Topic modeling of investment style news.

Auteur : Boemer, Dominik
Promoteur(s) : Ittoo, Ashwin
Faculté : HEC-Ecole de gestion de l'Université de Liège
Diplôme : Master en sciences de gestion, à finalité spécialisée en management général (Horaire décalé)
Année académique : 2019-2020
URI/URL : http://hdl.handle.net/2268.2/10346

Topic Modeling of Investment Style News

Promoter:
Ashwin Ittoo

Reviewers:
Cédric Gillain
John Pietquin

Master's thesis submitted by
Dominik Boemer
in partial fulfillment of the requirements for the
Master's degree in management sciences,
specializing in general management
Academic year 2019/2020

Topic Modeling of Investment Style News

Acknowledgments

First, I would like to thank *Ashwin Ittoo*, *Cédric Gillain* and *John Pietquin* for reading and reviewing this document. In particular, I want to thank Cédric Gillain for suggesting the topic, for our numerous discussions and for providing the data, which is analyzed in this thesis.

Moreover, I want to thank *Marie Lambert* and *Cédric Gillain* for their kind invitation to the *Asset & Risk Management Workshop* on July 8th, 2020.

Since the research topic was completely new to me, I relied on numerous *online courses*, which I would like to mention here since they are not cited in the following pages:

- *Finance Theory I* by Lo (2008)
- *Advanced Data Science* by Boyd-Graber (2018)
- *Natural Language Processing* by Jurafsky and Manning (2012)
- *Topic models* by Blei (2009)
- *Text Mining and Analytics* by Zhai (2016)
- *A Code-First Introduction to Natural Language Processing* by Thomas (2019)
- *Latent Dirichlet Allocation* and *Training Latent Dirichlet Allocation: Gibbs Sampling* by Serrano (2020)

Finally, I want to thank my *friends*, my *parents* and my *brother* for their support during this thesis and the shortly preceding PhD thesis.

Abstract

Smart beta exchange-traded funds (ETFs) are increasingly popular investment products among institutional investors. These ETFs can be categorized into different styles depending on the systematic risk factors to which they provide exposure. Hence, the question arises whether certain topics within the news coverage of specific styles influence the investment decision and thereby fund flows towards respective smart beta ETFs. This thesis focuses on partially answering this question by identifying the major topics in investment style news and their importance measured by their frequency of occurrence.

Based on a review of topic models, which are machine learning methods to discover topics in large collections of documents, latent Dirichlet allocation (LDA) is selected to identify the topics in investment style news. Moreover, the *most extensive literature survey of LDA in finance* (to the best of our knowledge) is compiled in order to optimally apply this method.

Subsequently, the major topics in a *unique corpus, which has never before been investigated by topic models* (to the best of our knowledge), are identified by LDA. This corpus consists of 1720 articles related to small-cap investing from 9 magazines targeting institutional investors.

The 5 major topics are “equity market (economy)”, “analyst research, trading and banking”, “retirement planing”, “indexes, ETFs and performance” and “fund management and fund launches”. These topics either persist, disappear or specialize when the number of topics to identify is increased. Dominant topics of individual magazines correspond to those proposed by the corpus specialist and the short descriptions of the magazines. The dominant topic over time is “fund management and fund launches”, which follows a seasonal trend characterized by lower coverage at the end of the year and higher coverage in January, thus suggesting that changes of fund management and fund launches preferentially occur at the beginning of the year.

Since the topic proportions of each article are identified, the correlation between the importance of topics over time and corresponding fund flows can be studied in future research.

Keywords: style investing, news coverage, topic modeling, latent Dirichlet allocation

Contents

Nomenclature

Acronyms

Acronym	Meaning
APT	Arbitrage pricing theory
CAPM	Capital asset pricing model
DJIA	Dow Jones Industrial Average
ETF	Exchange-traded fund
LDA	Latent Dirichlet allocation
LSA	Latent semantic analysis
LSI	Latent semantic indexing
NLP	Natural language processing
NMF	Non-negative matrix factorization
pLSA	Probabilistic latent semantic analysis
pLSI	Probabilistic latent semantic indexing
SVD	Singular value decomposition

Symbols

Symbol	Meaning
α	Parameter of the symmetric Dirichlet prior over θ_d
β_i	CAPM beta of asset i
β_{ik}	Sensitivity of asset i to factor k
β_{kw}	Probability of term w in topic k

$\boldsymbol{\beta} = (\beta_{kw})$	Per-topic word distributions
$\boldsymbol{\beta}_k$	Distribution of words in topic k
$\mathrm{Cov}(\cdot)$	Covariance
d	Label of a document
$\mathcal{D}$	Corpus of documents
e_i	Firm-specific return
$E(\cdot)$	Expected value
$E(R_k)$	Factor risk premium of factor k
$E(R_i)$	Risk premium of asset i
η	Parameter of the symmetric Dirichlet prior over β_k
F_k	Deviation of the common factor k from its expected value
$\mathbf{H}$	Matrix in NMF
$\mathbf{I}$	Identity matrix
K	Number of factors in a factor model
K	Number of topics in a topic model
λ	Parameter of the relevance measure in Eq. (4.1)
M	Number of documents in a corpus
μ_i	Expected return of asset i
$\hat{\mu}_i$	Estimated expected return of asset i
n	Number of assets in a portfolio
N_d	Number of words in document d
N_i	Number of assets (e.g. shares) i
$p(\cdot)$	Probability density or mass function
P_i	Price of asset i
$P_{i,t}$	Price of asset i at time t
r_f	Risk-free return (rate)
r_i	Return of asset i
$r_{i,t}$	Return of asset i at time t
r_M	Return of the market portfolio
r_p	Return of a portfolio
R_i	Excess return of asset i (with respect to r_f)

R_M	Excess return of the market portfolio (with respect to r_f)
σ_i	Standard deviation of the return of asset i
σ_i^2	Variance of the return of asset i
σ_M^2	Variance of the return of the market portfolio
σ_p^2	Variance of the return of a portfolio
$\hat{\sigma}_i^2$	Estimated variance of the return of asset i
$\mathbf{\Sigma}$	Diagonal matrix in the SVD
T	Number of periods in the estimation
$(\cdot)^T$	Transpose of matrix $(\cdot)$
θ_{dk}	Probability of topic k in document d
$\boldsymbol{\theta} = (\theta_{dk})$	Per-document topic distributions
$\boldsymbol{\theta}_d$	Distribution of topics in document d
$\mathbf{U}$	Matrix with orthonormal columns in the SVD
V	Number of words in a vocabulary
$\mathbf{V}$	Matrix with orthonormal columns in the SVD
w	Label of a word
$w_{d,n}$	Word at position n in document d
w_i	Portfolio weight of asset i
$\mathbf{W}$	Matrix in NMF
$\mathcal{W}$	Set of words in a vocabulary
X_{dw}	Frequency of term w in document d
$\mathbf{X} = (X_{dw})$	Document-term matrix
$\hat{\mathbf{X}}$	Approximation of the matrix $\mathbf{X}$
$\xi_{k,m}^a$	Absolute importance of topic k in magazine m
$\xi_{k,m}^r$	Relative importance of topic k in magazine m
$\xi_{k,t}^a$	Absolute importance of topic k in period t
$\xi_{k,t}^r$	Relative importance of topic k in period t
z	Label of a topic
$z_{d,n}$	Topic assignment of word n in document d
$\mathcal{Z}$	Set of topics

Chapter 1

Introduction

In this introduction, the *context* of this research, its *objectives*, the *outline* of this document and our *original contributions* are described.

1.1 Context

Investments in smart beta ETFs are clearly on the rise.[1] In fact, historical data of *ETFGI*, the leading independent research provider about the ETF/ETP industry, illustrate the global increase in assets under management (AUM) of smart beta ETFs/ETPs and their number in Fig. 1.1 (ETFGI, 2017).[2] According to *Morningstar's Global Guide to Strategic-Beta Exchange-Traded Products*, 1,493 smart beta ETPs exist as of December 31th, 2018, with about $797 billion AUM worldwide (Morningstar, 2019).[3] At the end of the same year, the Financial Times headlines "Smart beta moves into mainstream for large investors" (Riding, 2018). In addition, the most recent *EDHEC European ETF, Smart Beta and Factor Investing Survey* of 2019 further supports this statement: 51% of European professional asset managers participating in the survey already use smart beta and factor investing solutions, while 28% are considering to do so in the near future (Le Sourd and Martellini, 2019). Finally, in the *FTSE Russell 2019 Global Survey Findings from Asset Owners*, which surveyed 178 respondents with an estimated total AUM of over $5 trillion,

[1] *Exchange-traded funds* (ETFs) are "variants of mutual funds that allow investors to trade portfolios of securities just as they do shares of stocks" (Bodie et al., 2018). Smart beta ETFs are a subcategory of ETFs, which will be defined more precisely hereafter.

[2] ETFs are the most significant subcategory of *exchange-traded products* (ETPs) (Abner, 2016), with over 97% of the $5 trillion global ETP market consisting of ETFs (Small, 2018). This explains why both terms are often used interchangeably, although one should bear in mind that they have different meanings.

[3] "Strategic-beta" is Morningstar's terminology for smart beta (Ghayur et al., 2019).

57% of existing smart beta owners are evaluating additional allocations, while over 50% of those without smart beta, but currently evaluating it, plan to implement such a strategy in the near future (Russell FTSE, 2019).

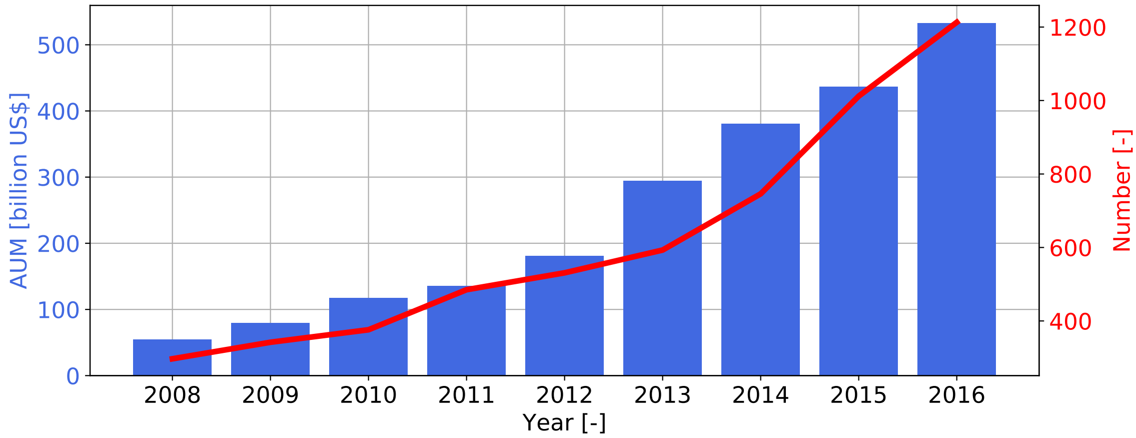


Figure 1.1: Assets under management (AUM, bars) and number of smart beta ETFs/ETPs worldwide (ETFGI, 2017).

A fundamental component of smart beta ETFs is the possibility to obtain *exposure to sources of systematic risk*, which are rewarded by corresponding risk premiums. For instance, a smart beta ETF of stocks with a small market capitalization (small cap) is expected to have higher average returns than a smart beta ETF of large capitalization stocks. Besides this *size factor*, which was described by Banz (1981), various other risk factors were identified in the literature (Harvey et al., 2016). Thus, different *investment styles* can be defined depending on which factor is captured, e.g. the small-cap investment style.

While investments in smart beta ETFs increase, one may wonder *what influence the media coverage of certain investment styles has on fund flows, i.e. net cash flows, towards/from corresponding smart beta ETFs*. This is the research questions of Gillain's PhD thesis (Gillain, 2020), which serves as the framework for this Master's thesis.

To answer the previous question, Gillain et al. created a *data set* of more than 100,000 articles from magazines, whose mission statement includes the production of *information for financial decision makers* (Gillain et al., 2019, 2020a,b). Since the number of articles about a certain investment style could serve as a proxy for investor attention, they further developed several *methods to identify the articles about a given style*. In particular, articles about the small-cap style were identified by a lexicon-approach, which is based on classifying articles depending on the occurrence of theme-specific words in their content. In simple terms, if an article contains the words "small cap" or a similar variation, this article is assumed to contain information about this style. In this way, it is possible to determine the number of articles related to the small-cap style for a given period t.

The previous research question can then be answered by *multiple linear regression models* similar to those in Sirri and Tufano (1998), Jenkinson et al. (2016) or Cao et al. (2017). In these models, *fund flows* of smart beta ETFs belonging to the small-cap style are dependent variables and the *media coverage* of this style is introduced as an independent variable, e.g. as in Fang et al. (2014). More precisely, it is assumed that fund flows of smart beta ETFs belonging to the small-cap style can be written as a linear combination of various factors including the media coverage of this style. Fund flows are usually normalized by total net assets of the fund, while the media coverage is computed as a function, e.g. a logarithm in Fang et al. (2014), of the number of articles related to the investment style. In addition, a temporal lag is introduced between fund flows and the media coverage since this coverage is assumed to cause fund flows. Basically, the normalized fund flow at time t can be denoted by "flow_t" and the coverage at $t-1$ by "coverage_{t-1}". Hence, the parameter β in the following equation should give some indication about the influence of the media coverage on the fund flow after its value is determined by the least-squares method, i.e. linear regression:

$$\text{flow}_t = \beta \, \text{coverage}_{t-1} + \dots \tag{1.1}$$

So far, the number of articles related to the small-cap style is used in the model to compute the coverage but one might anticipate that *not all articles of this style have the same influence on respective fund flows*. Possibly, only articles about certain topics impact these flows. This brings us to the research objective of this document.

1.2 Objectives

The main objective is to *identify the major topics in the investment style news about the small-cap style* by a text mining method based on machine learning. Any other approach would require significant prior expert knowledge in this field or reading several hundred articles of the previous data set, which would obviously be excessively time-consuming.

The text mining method should further allow to *cluster news articles according to their topics* in order to determine the frequency of occurrence of a certain topic during a given period or in a given magazine. This frequency could then be interpreted as the *importance* of this topic during this period or in this magazine. For instance, it should be possible to determine the importance of the topic "fund launches" for each month, if "fund launches" is one of the major topics within the data set. In this way, the informational granularity is increased from the investment style to the main topics within the articles related to this style. Ultimately, in future research, this should allow to determine whether the coverage of certain topics influences fund flows of small-cap smart beta ETFs.

To prepare this research and to better understand the data set, further objectives consist in analyzing the *topic coverage in each magazine* and the *importance of topics over time*.

1.3 Outline

This document has the following structure:

- Chapter 2 focuses on *defining the context in more detail* and on *reviewing the literature*. First, some fundamental concepts of investments, like modern portfolio theory and arbitrage pricing theory, are introduced to better understand factor investing and thus, smart beta ETFs. This section is useful to grasp the general context of this thesis but it is *not a requirement* for the following chapters. Afterwards, machine learning methods to identify topics in collections of documents, also known as *topic models*, are examined. And finally, the application of the most promising topic model, i.e. latent Dirichlet allocation (LDA), in finance is reviewed to identify how this method is optimally applied to our data set.
- Chapter 3 describes the *data* consisting of investment style news that are analyzed in this document.
- Chapter 4 explains the *methodology* that is applied to analyze the previous data. First, the programming environment, which is required for advanced text mining, is introduced as well as the 20 newsgroups corpus whose main topics are known.[4] Then, the pre-processing steps of the data, the topic model as such, and the post-processing of the results are detailed and illustrated by the previous corpus to validate the methodology.
- Chapter 5 finally discusses the *results*, i.e. the identified topics, the major topics in each news magazine and the importance of topics over time.

[4]A corpus is a collection of documents.

For the sake of completeness, various resources are included in the appendices. Appendix A contains an *example of the textual data in its initial format*. Appendix B includes the *Python code* of our topic modeling module with most pre- and post-processing functions, and a standard analysis script of the news that uses this module. Appendix C illustrates the *HTML result files* which are created by the previous code based on the previous example of textual data. And finally, appendix D details the results of the topic model that are discussed in Chap. 5.

1.4 Original contributions

Essentially *two original contributions*, which cannot be found in the existing scientific literature, are included in this thesis:

- First, a relatively extensive literature review analyzes the application of *latent Dirichlet allocation in finance*. No such review exists so far in the literature to the best of our knowledge, except for Loughran and McDonald (2016), who only cite a single reference in this field.
- The second originality is related to the data that are analyzed in this thesis. To the best of the author's knowledge, *topic modeling has not yet been applied to investment style news*.

Chapter 2

Contextualization and literature review

This chapter first introduces *basic concepts of investment theory* to better understand the topic of this research in its entirety, thereby complementing the explanations in the introduction. These concepts are, however, no mandatory requirement to understand the following sections and Sec. 2.1 can therefore be skipped. Subsequently, machine learning methods to identify topics in large collections of documents are introduced. These so-called *topic models* are also reviewed in order to determine the most promising method to extract topics of the investment style news. Finally, the literature about *latent Dirichlet allocation in finance* is surveyed to determine how this selected topic model is optimally applied to the investment style news in the following chapters.

2.1 Towards smart beta ETFs

In this section, some *fundamental concepts of investment theory* are summarized to understand the origin and definition of smart beta ETFs. These fundamental concepts are explained at great length in numerous textbooks, e.g. Amenc and Le Sourd (2003); Brealey et al. (2011); Elton et al. (2014); Reilly and Brown (2011); Vernimmen et al. (2018). Therefore, we restrict the following explanations to the most essential information, which is, however, no requirement for the following sections. Hence, this section can be skipped. Notice that the book *Investments* by Bodie et al. (2018) was selected as the main reference of the following subsections about fundamental concepts, i.e. modern portfolio theory, the capital asset pricing model and arbitrage pricing theory.

2.1.1 Modern portfolio theory

Modern portfolio theory is a methodology introduced by Markowitz (1952) to construct a portfolio of assets that maximizes the expected return for a given level of risk.

More precisely, the *return* $r_{i,t}$ of a risky asset i over a given time period t can be defined as its relative price change $P_{i,t} - P_{i,t-1}$ with respect to its price $P_{i,t-1}$ at the beginning of the period:[1]

$$r_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}} \tag{2.1}$$

On the one hand, the *expected return* is then defined by $\mu_i = E(r_i)$, which is the expected value of r_i.[2] On the other hand, *risk* is measured by the deviation of returns from the expected return, i.e. by the variance of the returns $\sigma_i^2 = E\left[(r_i - \mu_i)^2\right]$, where σ_i is the standard deviation. These values can be estimated by time series of past returns over T periods of the same length:

$$\hat{\mu}_i = \frac{1}{T}\sum_{t=1}^{T} r_{i,t} \qquad\qquad \hat{\sigma}_i^2 = \frac{1}{T-1}\sum_{t=1}^{T}(r_{i,t} - \hat{\mu}_i)^2 \tag{2.2}$$

A *portfolio* of n risky assets is defined as a collection of these assets. Hence, its return, its expected return and its risk are provided by the following expressions:[3]

$$r_p = \sum_{i=1}^{n} w_i r_i \qquad E(r_p) = \sum_{i=1}^{n} w_i E(r_i) \qquad \sigma_p^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \mathrm{Cov}(r_i, r_j) \tag{2.3}$$

where $w_i = N_i P_i / \left(\sum_{k=1}^{n} N_k P_k\right)$ is the weight of an asset in the portfolio, with N_i being its number and P_i its price such that $\sum_{i=1}^{n} w_i = 1$, and where $\mathrm{Cov}(r_i, r_j) = E\left[(r_i - \mu_i)(r_j - \mu_j)\right]$ is the covariance of returns r_i and r_j, which can also be estimated from past returns by an expression similar to that of the variance in Eq. (2.2).

The pairs of expected return $E(r_p)$ and risk σ_p^2 of all possible portfolios for given values of $E(r_i)$ and $\mathrm{Cov}(r_i, r_j)$ can be represented graphically as illustrated in Fig. 2.1. The individual assets and all possible portfolios are contained within a *minimum-variance frontier* that is defined by the minimum variance portfolio for a given expected return. The upper branch of the frontier is

[1]Depending on the type of asset, additional returns, like dividends for equity securities have to be considered. Similarly, transaction costs and taxes are neglected in this explanation.

[2]The subscript t was removed for readability.

[3]The first expression is a definition, while the following identities can be proven by the linearity of expectation and the definitions of variance and covariance.

the *efficient frontier of risky assets* since it corresponds to the portfolios with the highest returns for a given level of risk. Hence, Markowitz (1952) answered the question about which portfolio is the most efficient one for a given level of risk within this context.

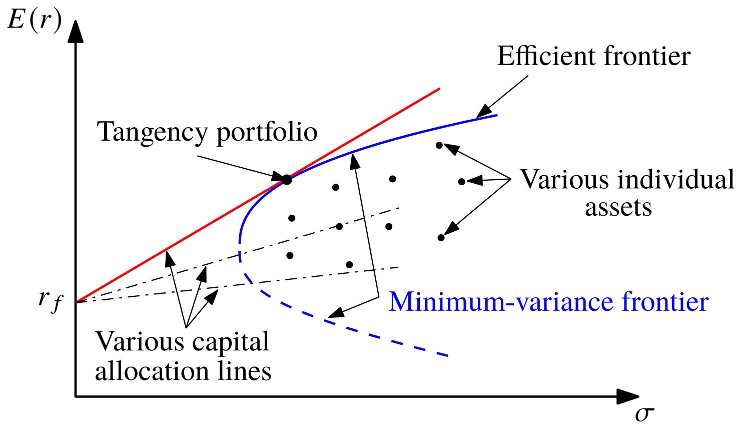


Figure 2.1: Markowitz portfolio optimization model.

In the presence of a risk-free asset, like Treasury bills classically, which have a risk-free return r_f, a number of additional portfolios become accessible by combining this asset with a risky portfolio. These new portfolios are defined by *capital allocation lines*, which join the risk-free asset with any risky portfolio. In particular, the return can be maximized for a given level of risk by combining the risk-free asset with the *tangency portfolio*, i.e. the portfolios defined by the red line in Fig. 2.1. Thus, the construction of an optimal portfolio can be separated into the independent tasks of identifying the optimal risky portfolio and the capital allocation between this portfolio and the risk-free asset, which is known as the *separation property* (Tobin, 1958).

Based on the previous framework, it is possible to illustrate two types of risk: systematic risk and nonsystematic risk. *Systematic risk* is risk that can be attributed to common risk sources among assets, which is why it is also called market risk. *Nonsystematic* risk, also known as firm-specific or idiosyncratic risk, is risk that can only be attributed to specific characteristics of an asset. In fact, the portfolio risk σ_p^2 in Eq. (2.3) can algebraically be re-written as follows by assuming that it is equally weighted, i.e. $w_i = 1/n$:[4]

$$\sigma_p^2 = \frac{1}{n}\,\overline{\sigma}^2 + \frac{n-1}{n}\,\overline{\text{Cov}} \quad \text{with} \quad \overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \quad \text{and} \quad \overline{\text{Cov}} = \frac{1}{n(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{i=1}^{n}\text{Cov}(r_i, r_j) \tag{2.4}$$

[4]The following conclusion is obviously also verified when assets are not equally weighted, provided that their individual risk contributions decrease when assets are added to the portfolio.

where $\overline{\sigma}^2$ and $\overline{\text{Cov}}$ are the average variance and the average covariance of the portfolio, respectively. The nonsystematic risk can be identified as the variance term, while the systematic risk is the covariance term.

When the number of assets n in the portfolio increases, the nonsystematic risk converges to zero, while $(n-1)/n \to 1$, so that systematic risk persists. This is an illustration of *diversification*, which consists in spreading a portfolio over many assets to reduce the exposure to specific sources of risk. For this reason, nonsystematic risk is also called *diversifiable risk*.

2.1.2 Capital asset pricing model

Based on Markowitz's portfolio theory, the *capital asset pricing model* (CAPM) by Treynor (1962), Sharpe (1964), Lintner (1965) and Mossin (1966) allows to characterize the tangency portfolio and the expected return of an asset, again under fairly strong assumptions.

Mainly, investors are assumed to have the same information about all assets, i.e. their expected returns $E(r_i)$ and risks $\text{Cov}(r_i, r_j)$. If investors can further borrow and lend at the same risk-free rate r_f, they would all find the same tangency portfolio. Finally, if they are rational mean-variance optimizers, i.e. investing according to Markowitz's theory, they would all invest in portfolios along the capital allocation line defined by the risk-free rate and the tangency portfolio. In consequence, since borrowing and lending is assumed to equal out, the *market portfolio*, which is the aggregation of all risky portfolios, has the same weights w_i as the tangency portfolio. And since the tangency portfolio is efficient, i.e. mean-variance optimal, this implies that the market portfolio is efficient, too.

In consequence, investing in a *capitalization-weighted market index*, like the S&P 500, is also efficient.[5,6] The index should indeed be cap-weighted since the previous market portfolio is cap-weighted as it contains all risky assets in proportion to their market capitalization. The previous idea was implemented by the launch of the first mutual index fund by Vanguard in 1976, which is the *Vanguard 500 Index Fund* tied to the S&P 500 (Kula et al., 2017).

Moreover, the contribution to the portfolio risk of asset i can be computed as follows by Eq. (2.3),

[5]A *capitalization-weighted or market-value-weighted index* is an index of components weighted by their market capitalization. For instance, if only a total of 10 assets A at \$2 and 30 assets B at \$4 exist on the market, their respective market capitalization is $10 \times \$2 = \$20$ and $30 \times \$4 = \$120$. Hence, a market-value-weighted index would contain $20/(20+120) = 14.3\%$ of A and $120/(20+120) = 85.7\%$ of B, like 1 A and 3 B. In fact, $1 \times \$2/(1 \times \$2 + 3 \times \$4) = 14.3\%$ and $3 \times \$4/(1 \times \$2 + 3 \times \$4) = 85.7\%$.

[6]Strictly speaking, the example is not fully appropriate since the S&P 500 contains not all available assets, and since only shares available for public trading (free float) are included in the capitalization weight.

by the definition of the covariance and by the previous assumptions:

$$w_i \sum_{j=1} w_j \operatorname{Cov}(R_i, R_j) = w_i \operatorname{Cov}\left(R_i, \sum_{j=1} w_j R_j\right) = w_i \operatorname{Cov}(R_i, R_M) \tag{2.5}$$

where $R_i = r_i - r_f$ is the excess return, which is introduced for the sake of brevity, and where R_M is the excess return of the market portfolio.

In addition, the *risk premium* $E(R_i)$ can be defined as the difference between the expected return $E(r_i)$ and the risk-free rate r_f, or, in other words, the expected value of excess returns. Hence, the contribution to the portfolio *risk premium* of asset i is $w_i E(R_i)$. If the market portfolio is in equilibrium, i.e. all investments have the same reward-risk ratio, the ratios corresponding to asset i and the market portfolio should be equal:

$$\frac{w_i E(R_i)}{w_i \operatorname{Cov}(R_i, R_M)} = \frac{E(R_M)}{\sigma_M^2} \tag{2.6}$$

Rearranging this equation finally results in the classical CAPM *expected return-beta relationship*, in which β_i can be interpreted as the sensitivity of the asset return to the market return:

$$E(r_i) = r_f + \beta_i \left[E(r_M) - r_f\right] \quad \text{with} \quad \beta_i = \frac{\operatorname{Cov}(R_i, R_M)}{\sigma_M^2} \tag{2.7}$$

Its main conclusion is that *only systematic risk is rewarded by a risk premium* since it cannot be diversified away unlike nonsystematic risk (see Eq. 2.4). So far, the risk premium of an asset is therefore a function of its market beta and the market risk premium.

2.1.3 Arbitrage pricing theory

Later, Ross (1976) introduced the *arbitrage pricing theory* (APT) to derive the risk premium due to the exposure to *multiple* common risk sources, instead of the single risk source in the CAPM. The APT is predicated on three main components: first, the excess returns of an asset can be described by the sum of its expected value, fluctuations due to common factors to all assets and asset-specific variations, i.e. by a *factor model*:

$$R_i = E(R_i) + \sum_{k=1}^{K} \beta_{ik} F_k + e_i \tag{2.8}$$

with β_{ik} the sensitivity of asset i to factor k, F_k the deviation of the common factor k from its expected value, such that $E(F_k) = 0$, and e_i the firm specific return, also such that $E(e_i) = 0$ (Amenc and Le Sourd, 2003). The deviation F_k could, for instance, be the difference between actual GDP growth and its expected growth.

Moreover, if nonsystemtic risk can be eliminated by diversification and if markets do not allow arbitrage opportunities to persist, the risk premium of asset i can be written as the linear combination of the *factor risk premiums* $E(R_k)$:[7,8]

$$E(R_i) = \sum_{k=1}^{K} \beta_{ik} E(R_k) \tag{2.9}$$

One should notice that if market risk is the only common risk factor, the previous equation is identical to Eq. (2.7) of the CAPM.

2.1.4 Factor investing

Arbitrage pricing theory showed that the risk premium of an asset is a function of the exposure of this asset to different factors, also known as *risk factors*. Hence, it is possible to invest in assets specifically because of their exposure to certain factors. This is called *factor investing* (Bender et al., 2013; Ghayur et al., 2019).

Throughout the years, more than 300 factors have been published in the scientific literature to explain the cross-section of expected returns, i.e. why some assets have higher returns than others (Harvey et al., 2016). Probably, the most influential publication in this field is literaly "The cross-section of expected returns" by Fama and French (1992), which gave rise to the Fama-French three-factor model (Fama and French, 1993). This model is based, among other things, on two empirical observations.

The first observation, which was initially documented by Banz (1981), is the *small-firm effect*. It states that stocks of firms with smaller *market capitalization*, which is the share price times the number of outstanding shares, have higher average returns than those with a larger market capitalization. Fig. 2.2 illustrates this observation by representing the value-weighted annual returns of 10 size-based portfolios. These portfolios are constructed by sorting all traded U.S.

[7] An *arbitrage opportunity* consists in the possibility of achieving risk-free profits without making a net investment, e.g. instantaneously buying for a low price and selling for a higher price.

[8] A simplified proof can be found in Bodie et al. (2018, p. 313-318), while a more formal proof is available in Amenc and Le Sourd (2003, p. 190-192).

stocks according to their market capitalization and then grouping them into 10 portfolios based on the sort order. One should notice that the small-firm effect persists even after adjusting the returns by the CAPM, so that this effect cannot exclusively be explained by the market risk factor.

The second observation is the *book-to-market effect*, which consists in average returns of stocks increasing with their *book-to-market ratio*. This ratio is defined as the accounting value of a firm divided by its market capitalization. Fig. 2.3 illustrates this observation based on a methodology of portfolio construction similar to the one that was previously applied for the small-firm effect.

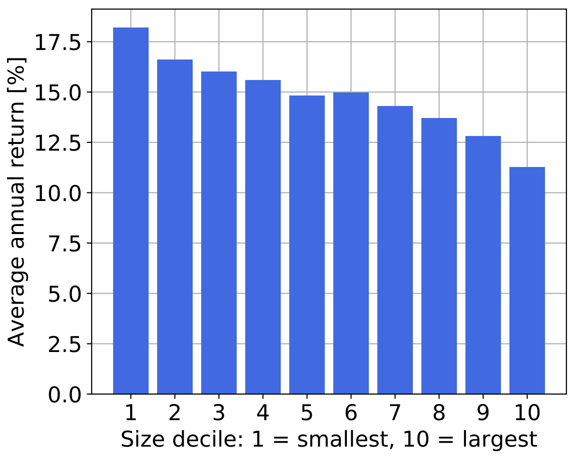


Figure 2.2: Average annual return from 1926 to 2019 for 10 portfolios constructed as a function of *market capitalization* based on data by French (2020).

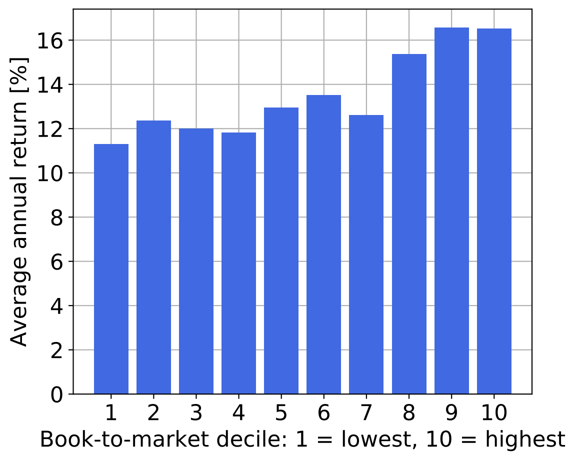


Figure 2.3: Average annual return from 1926 to 2019 for 10 portfolios constructed as a function of *book-to-market ratio* based on data by French (2020).

Hence, the Fama-French three factor model is defined by stating that the expected excess return of an asset $E(R_i)$, or more practically speaking, the average stock return $E(r_i) = E(R_i) + r_f$, is computed as the linear combination of three risk premiums, similar to Eq. (2.9) with $K = 3$:

$$E(R_i) = \beta_{iM} E(R_M) + \beta_{i\mathrm{SMB}} E(\mathrm{SMB}) + \beta_{i\mathrm{HML}} E(\mathrm{HML}) \tag{2.10}$$

While the first term is simply the market risk premium as in the CAPM, the two additional factors are SMB (Small Minus Big), i.e. the difference between the average returns of small-cap stocks and large-cap stocks, and HML (High Minus Low), i.e. the difference between the average returns of high book-to-market stocks (value stocks) and low book-to-market stocks (growth stocks).

Due to the strong predictive power of expected returns over different periods and markets by this model, *size* (via the market capitalization) and *value* (via the book-to-market ratio) became the most well-known factors. In particular, these are the categories of the *Morningstar style box* (Morningstar, 2002), which is used to characterize the investment positioning of mutual funds based on research by (Sharpe, 1992). This led to the notion of *style box investing* or more generally *style investing* that is defined as investing in groups of securities sharing a common attribute, e.g. small-cap or value stocks, instead of individual securities (Barberis and Shleifer, 2003).

In agreement with the findings by Fama and French (1992), a mutual fund could mainly select small capitalization stocks with high book-to-market ratios to increase their expected returns. This idea, which is also described by the phrase "tilting a portfolio towards factors" can explain the positive difference between some fund returns and capitalization-weighted benchmarks, like the S&P 500 (Bender et al., 2013). Various risk-based, behavioral or structural *reasons* are advanced to explain the higher expected returns associated with the previous factors (Ghayur et al., 2019). For instance, the risk premium of small capitalization stocks can be interpreted as the investor compensation for higher risk due to less investment information about small firms being available (*neglected-firm effect*, Merton, 1987).

2.1.5 Smart beta ETFs

Almost 20 year after the launch of Vanguard's first mutual index fund, America's first exchange-traded fund (ETF) was launched in 1993. In contrast to the Vanguard 500 Index Fund, the new Standard & Poor's Depository Receipt, also known as SPDR, allowed to buy and sell an S&P 500 index portfolio like a regular share of stock. More generally, *exchange-traded funds* are "variants of mutual funds that allows investors to trade portfolios of securities just as they do shares of stocks" (Bodie et al., 2018). Most ETFs are designed to passively track indexes as closely as possible (Goltz and Le Sourd, 2015). Their main advantages with respect to mutual funds are continuous trading, lower costs and transparency. The most well known ETF providers are BlackRock with the product line iShares, and Vanguard (Kula et al., 2017).

Classical ETFs, i.e. those that are tracking capitalization-weighted indexes, can, however, be *criticized* mostly for two reasons. On the one hand, these indexes might suffer from the *excessive overweighting of some companies* in the index, thus exposing investors to non-rewarded unsystematic risk. For instance, in 2017, the three largest companies on Euronext Brussels had the same market capitalization as the 130 remaining companies (Ghayur et al., 2019). On the other hand, these classical ETFs lack the possibility to *control the intentional exposure to risk factors*.

A solution to these criticisms are *smart beta ETFs*. Their definition is not unique but the following statements include the main characteristics. Amenc et al. (2015) define smart beta as a "new indexation approach that deviates from broad cap-weighted market indices", while Russell Investments (2014) describes them as "transparent, rules-based indexes designed to provide exposure to specific factors, market segments or systematic strategies". Hence, the fundamental elements of smart beta ETFs are *alternative weighting strategies* and *factor investing*. While factor investing, i.e. the selection of stocks within the fund based on their exposure to risk factors, has already been explained in Sec. 2.1.4, alternative weighting strategies consist in computing the weight of a stock in the fund differently than by its market capitalization. Numerous weighting schemes, like price weighting (one shear of each stock, e.g. DJIA) or equal weighting (same amount invested in each stock), with various characteristics, like optimal diversification, are described in the literature (Amenc et al., 2014; Kula et al., 2017).

In the context of this document, smart beta ETFs are of importance due to their recent *popularity*, which was illustrated in Chap. 1, and the *intentional exposure to systematic risk factors* that they offer. In particular, this document focuses on characterizing the news coverage related to the *small-cap investment style* to determine how this coverage influences fund flows towards smart beta ETFs, which provide exposure to the *small-size factor*, in future research.

2.2 Topic models

In this section, topic models are reviewed to determine *which model should at best be used* to identify the topics in investment style news. Before addressing specific models, topic modeling is first situated in the *context of machine learning* as an introduction, and necessary *fundamental concepts of natural language processing* are explained.

2.2.1 Topic models in machine learning

Due to the increasing quantity of digitized data, machine learning becomes increasingly necessary to find the information that we are looking for. *Machine learning* can be defined as "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" (Murphy, 2012).

Machine learning methods are classically divided into supervised learning and unsupervised learning. *Supervised learning* requires a labeled data set, also known as the training set, so that a computer can learn a mapping from an input set to an output set. For instance, Gillain et al. (2019) use the Naive Bayes algorithm (Jurafsky and Martin, 2019, p. 56) to classify investment style news (input set) into articles about small- and large-cap investment styles (output set). This requires manually specifying whether each article of the training set belongs to one, both or neither of these classes. It becomes quickly apparent that supervised learning is usually synonymous with significant human implication upfront. In contrast to supervised learning, *unsupervised learning* tries to find patterns in the data on itself and therefore does not require explicitly labeled data.

Unsupervised learning can be divided into two types of methods: dimensionality reduction and clustering. One the one hand, *dimensionality reduction* consists in projecting data from a high-dimensional space to a lower dimensional subspace, which is supposed to capture the essence of the original data. On the other hand, *clustering* focuses on sorting data elements into groups such that elements in the same group are similar.

Topic models can be defined as "algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents" (Blei, 2012).[9] They can be viewed as dimensionality reduction and clustering methods. First, they involve finding the subspace of topics in the high-dimensional space of words in documents and they can therefore be considered to be dimensionality reduction methods. Furthermore, documents can be clustered by topic models into topic groups. The difference between classical clustering methods, like k-means clustering (Coelho and Richert, 2015), and topic modeling is that a document can belong to multiple clusters (topics) instead of a single one (e.g. Grafe, 2010).

2.2.2 Fundamental concepts of natural language processing

In this section, some *fundamental concepts* of natural language processing, which are used in the following sections, are introduced.

Natural language processing (NLP) can be defined as the application of computational methods to the analysis of human language (Manning et al., 2008; Jurafsky and Martin, 2019). A topic model can therefore be considered to be an NLP method.

In NLP, a collection of text documents is called a *corpus*. To enable the analysis of a corpus by a computer, the text of the documents has to be *pre-processed*, i.e. converted to a more convenient representation to apply computational methods. The most common pre-processing operation is *tokenization*, which separates the running text of a document into tokens. These tokens are usually words, also known as terms. Hence, it is possible to count the frequency of occurrence of

[9]This definition has a rather inclusive view of topic modeling, which includes matrix factorization approaches, like LSA (Sec. 2.2.3), in the definition. Other authors, like Steyvers and Griffiths (2007), refer to probabilistic topic models (Secs. 2.2.5 and 2.2.6) by "topic models".

each term in an document. Other pre-processing techniques, like the removal of stop words, case folding, stemming, lemmatization, ... are also common. They are explained more thoroughly in Sec. 2.3.2 to prevent redundancy.

The term frequencies of all words in a corpus can be encoded in a *document-term matrix* $\mathbf{X}$.[10] The rows of this matrix correspond to documents, while its columns correspond to terms. For instance, the element X_{dw} of $\mathbf{X}$ is the frequency of term w in document d.

The following topic models are based on this data representation, which is called the *bag-of-words model* (Manning et al., 2008). In this model, the ordering of terms in documents is neglected, and only the term frequencies and their assignments to documents are of importance. Hence, the sentences "James is taller than Jennifer" and "Jennifer is taller than James" are equivalent according to this model. It seems, however, reasonable that documents with similar bag-of-words models have a similar content, so that this simplification is not overly detrimental to detecting topics.

Finally, one should notice that simple word counts in the document-term matrix might place too much importance on frequent words. The most popular solution to this problem is the *tf-idf weighting scheme*. Numerous variants of this scheme exist (Manning et al., 2008) but the underlying principle of all these variants is to discount the term frequencies (tf) in the document-term matrix by the document frequency of the term (df). This document frequency is the number of documents in the corpus, which contain the term. The most direct way of implementing this scheme is to divide all term frequencies in $\mathbf{X}$ by the corresponding document frequencies, which explains the name *tf-idf* (term frequency-inverse document frequency). In consequence, frequent terms in all documents have reduced scores in the document-term matrix.

In the following sections, the *most significant topic models* are reviewed.

2.2.3 Latent semantic analysis

Latent semantic analysis (LSA), which is also known as latent semantic indexing (LSI) in the field of information retrieval, is based on the *singular value decomposition* (SVD) of the document-term matrix $\mathbf{X}$ (Deerwester et al., 1990). Although the SVD is a classical operation in linear algebra (Strang, 2016), the full derivation of this method is not straightforward. For this reason, only the general idea is described hereafter with a practical example.

[10] Other authors use the transpose of the document-term matrix, which is the term-document matrix, e.g. Manning et al. (2008). In this document, the document-term matrix is chosen since this matrix is created by default in the Python machine learning library *scikit-learn* (Pedregosa et al., 2011), which will be used in Chap. 4.

If the document-term matrix contains M documents and V terms, the SVD consists in writing $\mathbf{X}^{M \times V}$ as the product of two matrices $\mathbf{U}^{M \times M}$ and $\mathbf{V}^{V \times V}$ with orthonormal columns, and a rectangular diagonal matrix $\mathbf{\Sigma}^{M \times V}$. The superscripts indicate the dimensions of the matrices with the exception of $(\cdot)^T$, which is the transpose of $(\cdot)$, in the decomposition:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2.11}$$

For instance, Tab. 2.1 represents a hypothetical document-term matrix $\mathbf{X}$ of documents about factor investing and topic modeling. A singular value decomposition of $\mathbf{X}$ can be written as follows by using the PYTHON library *NumPy* (Oliphant, 2006) and by rounding to the nearest tenth:

$$\begin{pmatrix} 1 & 1 & 2 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0.7 & -0.7 & 0 \\ 0 & 0.7 & 0.7 & 0 \\ 0.6 & 0 & 0 & 0.8 \\ 0.8 & 0 & 0 & -0.6 \end{pmatrix} \begin{pmatrix} 3.9 & 0 & 0 & 0 & 0 \\ 0 & 3.1 & 0 & 0 & 0 \\ 0 & 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0.9 & 0.4 \\ 0.7 & 0.2 & 0.7 & 0 & 0 \\ 0.7 & -0.5 & -0.5 & 0 & 0 \\ 0 & 0 & 0 & -0.4 & 0.9 \\ 0.3 & 0.8 & -0.5 & 0 & 0 \end{pmatrix} \tag{2.12}$$

	risk	premium	factor	data	model
Document 1	1	1	2	0	0
Document 2	2	0	1	0	0
Document 3	0	0	0	2	1
Document 4	0	0	0	3	1

Table 2.1: Example of document-term matrix of documents about factor investing and topic modeling.

The importance of this factorization in topic modeling can be explained by the following idea: one would expect words in one topic to occur less frequently in another topics. Hence, if topics are represented as vectors whose elements give some indication about the likelihood of words in these topics, topics can be expected to be orthogonal. It just so happens that the matrix $\mathbf{V}^T$ has as many columns as there are words in the example corpus (Tab. 2.1). Moreover, its rows are orthonormal due to the orthonormality of the columns of $\mathbf{V}$, i.e. $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Hence, the rows in $\mathbf{V}^T$ can be interpreted as *topic vectors*. For instance, in Eq. (2.12), the first topic (first row) is about “data” (0.9 in 4th column) and “model” (0.4 in last column), while the second topic (second row) is about “risk”, “factor”, and a bit less about “premium”. Hence, LSA is able to detect the initial topics. The following rows in $\mathbf{V}^T$ contain, however, negative weights, which makes it difficult to interpret these topics.

The singular values, which are the diagonal terms of $\mathbf{\Sigma}$, can be viewed as the *importance of a topic* in the corpus. In fact, they act as scaling factors in the operation $\mathbf{\Sigma V}^T$ since the rows of $\mathbf{V}^T$ have a unitary norm because of their orthonormality.

Furthermore, an approximation $\hat{\mathbf{X}}$ of $\mathbf{X}$ can be constructed by keeping only the K largest singular values and the remaining columns and rows of $\mathbf{U}$ and $\mathbf{V}^T$, respectively. This approximation is the closest matrix of rank K to $\mathbf{X}$ according to the least squares norm (Frobenius norm), i.e. such that $\sum_{d=1}^{M}\sum_{w=1}^{V}(X_{dw}-\hat{X}_{dw})^2$ is minimal (Deerwester et al., 1990). The previous approximation can be computed in our example by keeping only the largest singular values, e.g. 3.9 and 3.1 in Eq. (2.12):

$$\begin{pmatrix} 0 & 0.7 \\ 0 & 0.7 \\ 0.6 & 0 \\ 0.8 & 0 \end{pmatrix}\begin{pmatrix} 3.9 & 0 \\ 0 & 3.1 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 & 0.9 & 0.4 \\ 0.7 & 0.2 & 0.7 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1.5 & 0.4 & 1.5 & 0 & 0 \\ 1.5 & 0.4 & 1.5 & 0 & 0 \\ 0 & 0 & 0 & 2.1 & 0.9 \\ 0 & 0 & 0 & 2.8 & 1.2 \end{pmatrix} \tag{2.13}$$

By comparing $\hat{\mathbf{X}}$ in the previous equation with the initial document-term matrix $\mathbf{X}$ in Eq. (2.12), it can be seen that $\hat{\mathbf{X}}$ approximates $\mathbf{X}$ quite well, although some content has been lost. The previous operation is an illustration of topic modeling being a method of *dimensionality reduction*.

Finally, the rows of the matrix $\mathbf{U}$, which has not yet been analyzed, provide some indication about the *occurrence of a topic (column) in a document (row)*. For instance, the first matrix in Eq. (2.13) indicates that the second topic ("risk", "factor" and a bit less "premium") occurs in the first and second documents. A quick look at Tab. 2.1 shows that this is actually true.

In conclusion, LSA allows to discover latent semantic topics but the *interpretation of the matrix elements is not evident* due to the possibility of negative terms (Lee and Seung, 1999) and no statistical foundation (Hofmann, 1999), e.g. topics are no probability distributions over words but unit vectors.

2.2.4 Non-negative matrix factorization

Non-negative matrix factorization (NMF) consists like LSA in approximating the document-term matrix $\mathbf{X}$ by a product of lower-rank matrices (Lee and Seung, 1999): $\hat{\mathbf{X}}^{M\times V} = \mathbf{W}^{M\times K}\mathbf{H}^{K\times V}$, where the superscripts indicate again the dimensions of the matrices with the number of documents M, the size of the vocabulary V and the number of topics $K \ll \min(M, V)$.

The NMF approximation is constructed by an iterative process that optimizes a cost function, like the squared Frobenius norm $\sum_{d=1}^{M} \sum_{w=1}^{V} (X_{dw} - \hat{X}_{dw})^2$ (Lee and Seung, 2001), subject to the *non-negativity constraint* of $\mathbf{W}$ and $\mathbf{H}$. Hence, the advantage of NMF over LSA is the non-negativity of term weights in a topic, and the non-negativity of topic weights in a document.

The NMF decomposition of the document-term matrix in our previous *example* can be computed by the NMF implementation in the Python machine learning library *scikit-learn* (Pedregosa et al., 2011) for $K = 2$:

$$\begin{pmatrix} 0 & 1.3 \\ 0 & 1.2 \\ 0.9 & 0 \\ 1.3 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 2.3 & 0.9 \\ 1.2 & 0.4 & 1.2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1.6 & 0.6 & 1.6 & 0 & 0 \\ 1.4 & 0.5 & 1.4 & 0 & 0 \\ 0 & 0 & 0 & 2.1 & 0.8 \\ 0 & 0 & 0 & 2.9 & 1.1 \end{pmatrix} \tag{2.14}$$

One should notice that this last matrix $\hat{\mathbf{X}}$ approximates again the initial document-term matrix $\mathbf{X}$, and that the matrices $\mathbf{W}$ and $\mathbf{H}$ resemble the matrices $\mathbf{U}$ and $\mathbf{V}^T$ of the SVD decomposition in Eq. (2.13). Hence, the first document (first row in $\mathbf{W}$) is again about the second topic (non-zero second column), which is consists of "risk", "factor" and a bit less "premium" (non-zero elements in second row of $\mathbf{H}$).

While we were previously lucky to have found positive term weights and document weights in the example after keeping only the K singular values of the SVD, NMF ensures by construction that they are positive.[11] The interpretation of the weights is, however, still not intuitive since they are the result of a constrained optimization procedure to factorize a matrix. For this reason, *probabilistic topic models* are introduced in the following sections (Steyvers and Griffiths, 2007).

2.2.5 Probabilistic latent semantic analysis

In the previous topic models, documents are essentially linear combinations of topics, which are themselves linear combinations of words. While these topic models are based on linear algebra, the models in this section and the following one are built on a *probabilistic* foundation. This foundation allows to simplify the interpretation of topics (Steyvers and Griffiths, 2007).

Before delving more deeply into the definition of a specific probabilistic topic model, the *general idea* can be summarized as follows: probabilistic topic models first postulate an imaginary *generative model* that describes how documents are created, by selecting latent topics and choosing

[11] We were actually not lucky previously because an SVD decomposition with non-negative elements of the truncated matrices was *intentionally* chosen to simplify the explanations. In fact, the SVD is not unique so that multiplying the same columns of $\mathbf{U}$ and $\mathbf{V}$ by -1 is still a valid SVD (Strang, 2016).

words from these topics. The word "latent" refers to the fact that these topics are not observed. The generative model is then mathematically formalized so that the probability of generating a certain corpus can be computed. Finally, the resulting probabilistic model is inverted by statistical inference based on the observed data of the document-term matrix to uncover the latent topics. More specifically, the topics are determined in such a way that the likelihood of creating the observed data by the generative model is maximized.

The first probabilistic topic model was introduced by Hofmann (1999) under the name of *probabilistic latent semantic analysis* (pLSA), which is also known as probabilistic latent semantic indexing (pLSI) in the context of information retrieval. To explain pLSA, we have to define $d \in \mathcal{D} = \{d_1, \ldots, d_M\}$ as a document in the corpus $\mathcal{D}$, and $w \in \mathcal{W} = \{w_1, \ldots, w_V\}$ as a word in a vocabulary $\mathcal{W}$. These documents and words define the rows and columns of the document-term matrix $\mathbf{X}$. It can further be assumed that the observed data in the document-term matrix are related to each other by latent variables $z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$ such that $\mathcal{Z}$ is the set of topics.

By means of these definitions, the *generative probabilistic model* that describes the creation of a word w in a document d is the following one:

1. Choose a document d with the probability $p(d)$,
2. Choose a topic z with the probability $p(z|d)$,
3. Choose a word w with the probability $p(w|z)$.

The previous notation $p(z|d)$ indicates that the probability measure is a conditional probability, i.e. the probability of selecting the topic z, if the document d was picked. Likewise, $p(w|z)$ is the probability of selecting the word w, if the topic z was picked.

Fig. 2.4 *illustrates* the generative process based on an article belonging to the corpus of investment style news that will be introduced in Chap. 3. Each document can exhibit multiple topics that are common to all documents. For this reason, these topics are written next to the documents, i.e. on the left in Fig. 2.4. In this example, the topics could be "fund management", "geographical regions" and "performance". These topics as well as the annotations in this figure were obviously imagined for didactic purposes. Considering the content of the article, they seem, however, to be reasonable.

In probabilistic topic modeling, *topics* are probability *distributions over terms* of a fixed vocabulary. For instance, the green topic in Fig. 2.4 contains the word "fund", which has a probability of 0.05 in this topic, or more formally $p(w = \text{fund}|z = \text{fund management}) = 0.05$. One should

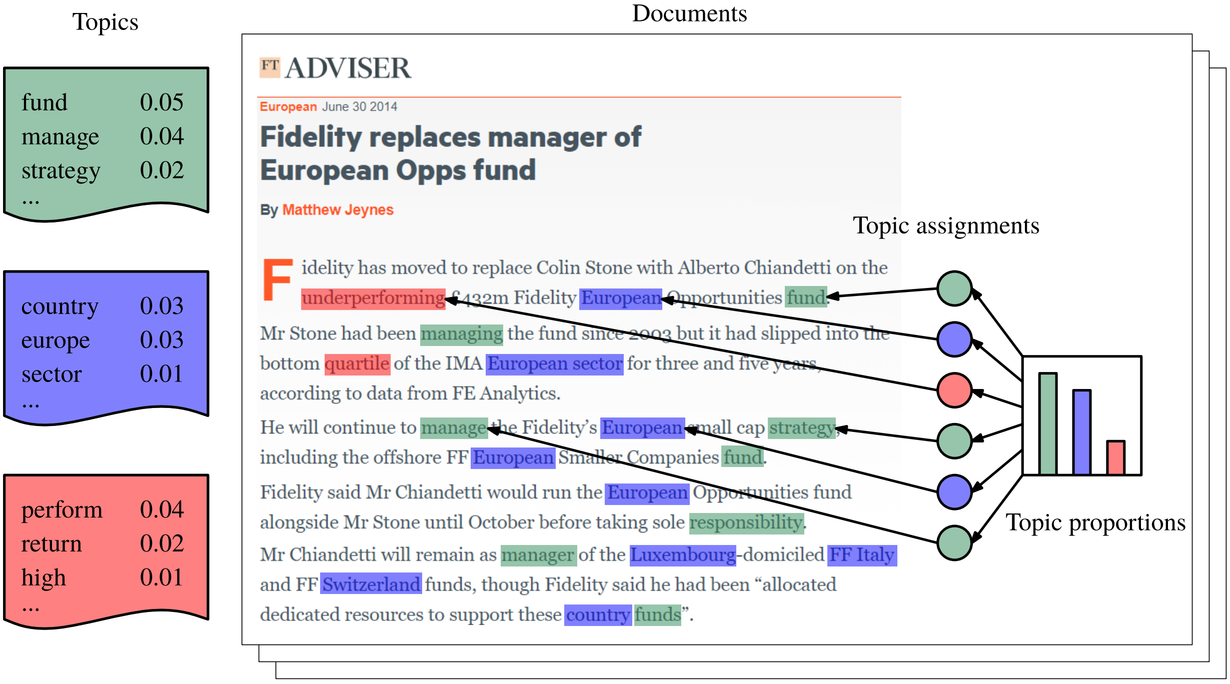


Figure 2.4: Generative probabilistic process to create a document illustrated based on an article of Jeynes (2014), which belongs to the corpus of investement style news (Chap. 3). The representation is similar to Fig. 1 in Blei (2012).

notice that topics are not mutually exclusive for a given word. This means that all words are included in all topics. They might, however, have different probabilities in different topics.

Documents in probabilistic topic models are mixtures of latent topics and therefore they may contain multiple topics. In our example, the article seems mainly to be about fund management according to the *topic proportions*, which are represented by the bar chart on the right in Fig. 2.4. These proportions, which are also known as the *distribution over topics*, indicate the probability of a topic occurring in the document. Mathematically, the green topic proportion is written $p(z = \text{fund management}|d)$, where d is the document in the example. Hence, to generate a word in a document, a topic is first randomly selected from the distribution over topics, as indicated by the topic assignments in Fig. 2.4. Then, a word is randomly chosen from this topic.

This generative process can be mathematically formalized by introducing the previous probabilities $p(d)$, $p(z|d)$ and $p(w|z)$ into the computation of the *joint probability* $p(w, d)$, i.e. the probability of observing a document d that contains the word w. First, by conditioning over all topics, one obtains:

$$p(d, w) = \sum_{z \in \mathcal{Z}} p(d, w|z) p(z) \tag{2.15}$$

Furthermore, a document d and a word w are assumed to be conditionally independent given a topic z. In simple terms, this means that words are selected independently of the specific document, if a topic has been chosen:

$$p(d, w) = \sum_{z \in \mathcal{Z}} p(d|z)p(w|z)p(z) \tag{2.16}$$

Then, the joint probability can finally be written as a function of the probabilities in the generative model by noticing that $p(d|z) = p(z|d)p(d)/p(z)$ according to Bayes' theorm:

$$p(d, w) = p(d) \sum_{z \in \mathcal{Z}} p(w|z)p(z|d) \tag{2.17}$$

This joint probability is only the probability of observing document d with the word w. However, a corpus may contain thousands of documents, which themselves contain thousands of words. If the corpus includes, for instance, only one document d_1 with three words (w_1, w_2, w_2), of which two are the same, the probability of observing this corpus would be $p(d_1, w_1)p(d_1, w_2)^2$ according to the bag-of-words assumption. In fact, this assumption allows writing these products of joint probabilities. Otherwise, one would have to condition with respect to other words in the document. Thus, the *probability of observing the entire corpus* can be quantified by the following *likelihood function*, which includes the counts of the document-term matrix $\mathbf{X} = (X_{dw})$:

$$L = \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} p(d, w)^{X_{dw}} \tag{2.18}$$

The underlying idea behind pLSA is then to assume that the best topics, i.e. $p(w|z)$, and topic proportions for each document, i.e. $p(z|d)$, are those that maximize the previous likelihood function given the observed data of the document-term matrix.[12] In statistics, this concept is known as *maximum likelihood estimation* (Murphy, 2012). In other words, a generative probabilistic model was defined to create a corpus. This model depends on various parameters, like the topics. The question is then to know which values take those parameters so that our specific corpus is the most likely to be created by the model.

[12]It is unclear why the probabilities $p(d)$ are not considered to be (unknown) parameters of the model in the literature, e.g. Blei et al. (2003) explain that pLSA has $KV + KM$ parameters, which correspond to $p(w|z)$ and $p(z|d)$, so that the probabilities $p(d)$ are not included. Possibly, these probabilities are set equal to the inverse of the number of documents in the corpus since all documents might be chosen with the same probability. Thus, $p(d)$ would be known and be no parameters anymore.

Hence, the topic model takes the form of an *optimization problem* to determine the (unknown) parameters $p(w|z)$ ($V \times K$ values) and $p(z|d)$ ($K \times M$ values) subject to the following constraints: the distribution over words should sum to one for each topic, i.e. $\sum_{w \in \mathcal{W}} p(w|z) = 1, \forall z \in \mathcal{Z}$, and the topic distributions should sum to one for each document, i.e. $\sum_{z \in \mathcal{Z}} p(z|d) = 1, \forall d \in \mathcal{D}$. This problem is finally solved by an *expectation-maximization algorithm*, which is an iterative method in statistics to find a local maximum of the likelihood function (Hofmann, 1999).

According to Blei et al. (2003), pLSA is, however, lacking a generative probabilistic model for the topic proportions of each document, which results in two *shortcomings*: on the one hand, the number of parameters of the model increases linearly with the number M of documents in the corpus. In fact, the previous paragraph shows that the number of parameters $p(z|d)$ increases with M. In a similar way to a polynomial regression with too many parameters, which leads to the interpolation of data points, the increasing number of parameters in pLSA suggests that it is prone to overfitting. On the other hand, it is unclear how to assign topic proportions to a new document, which was not previously included in the training set. Although the notions of training and testing sets are more commonly used in supervised learning, where algorithms first have to be trained by labeled data, they also exist in topic modeling. In fact, a topic model might learn some topics by itself and afterwards assign topics to new documents, i.e. the testing set. For these reasons, a topic model, which eliminates the previous shortcomings is introduced in the following section.

One might wonder why pLSA was presented in such detail before, if it is abandoned in favor of the following method. The answer is that the following method is partially built on pLSA and that this method is even more complex than pLSA in our opinion. Hence, we wanted to explain the underlying principle of probabilistic topic models in a simpler context than the following method.

2.2.6 Latent Dirichlet allocation

With more than 30,000 citations of the seminal publication (Blei et al., 2003) and numerous applications (Boyd-Graber et al., 2017), there is no doubt that *latent Dirichlet allocation* (LDA) is the most popular topic model. In contrast to pLSA, LDA introduces an assumption about how the topic proportions are generated for each document. In this way, LDA defines a more complete generative model, which includes the generation of topic proportions, so that the previous shortcomings of pLSA can be eliminated.

Similar to pLSA, LDA is based on the idea that the observed data in a corpus originate from a *generative probabilistic model* that includes hidden (latent) variables, which can be interpreted as the thematic structure of the corpus. Referring back to Fig. 2.4, this structure includes the topics, the topic proportions for each document and the topic assignments for each word.

Before explaining the generative model of LDA, one should notice that different versions of LDA exist. In this document, *LDA with symmetric Dirichlet priors and user-specified hyperparameters* is used for three reasons (Blei, 2012; Steyvers and Griffiths, 2007):[13] first, the Dirichlet prior on the per-topic word distributions favors topics defined by few words with high probability as will be explained later on. Secondly, this version of LDA is the most used version in the literature about LDA in finance (see Sec. 2.3). And thirdly, it is available in the PYTHON library *scikit-learn*, which is used in Chap. 4 to uncover the topics in investment style news.

The generative model of LDA can be described by its *plate notation* in Fig. 2.5. This notation allows to visually represent probabilistic models (Steyvers and Griffiths, 2007). The conditional dependences between random variables are indicated by arrows from independent to dependent variables. Shaded and unshaded circles represent observed and latent variables, respectively, while rectangles (plates) indicate repeated sampling of variables. The number in the right corner of rectangles specifies the number of these repetitions.

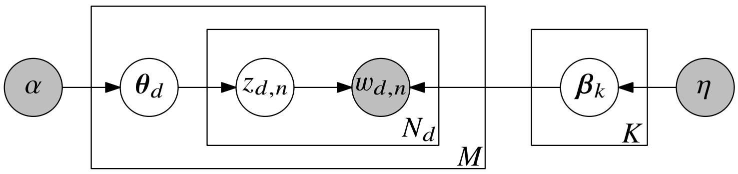


Figure 2.5: Plate notation of LDA with symmetric Dirichlet priors on per-document topic distributions and per-topic word distributions with user-specified hyperparameters α and η.

The generative process starts by creating K *topics* $\boldsymbol{\beta}_k$ with $k = 1, \ldots, K$. This explains the repetition indicated by the rectangle with subscript K in Fig. 2.5. As mentioned previously, a topic is a distribution over a vocabulary. If V represents the size of the vocabulary, i.e. the number of its words, each $\boldsymbol{\beta}_k$ can be interpreted as a V-dimensional vector whose elements are the probabilities of each word occurring in the topic k, as illustrated in Fig. 2.4 on the left. In

[13]The word "prior" is a concept of *Baysian statistics*. A full introduction to this theory was not added to this document but the following explanation could clarify the previous term by defining some key concepts of this theory (see Gelman et al., 2020 for more details). Some observed data x, which is called the *evidence*, might depend on an unobserved parameter θ. In this case, the prior probability distribution $p(\theta)$, also known as the *prior*, is the probability of θ before having seen x. *Hyperparameters* are parameters of prior distributions, here, the Dirichlet priors. A prior can be interpreted as our prior belief about θ. This belief could, however, change after having seen the previous evidence x, so that the posterior probability distribution, or simply *posterior*, is $p(\theta|x)$. According to Bayes' theorem, these distributions are related by the *likelihood* $p(x|\theta)$ and the marginal likelihood $p(x)$ such that $p(\theta|x) = p(x|\theta)p(\theta)/p(x)$. Finally, the *likelihood function* is the function $\theta \mapsto p(x|\theta)$.

LDA with a Dirichlet prior on these word distributions, the probabilities of each word in a topic are assumed to be drawn from a Dirichlet distribution. As mentioned previously, this distribution is chosen to be symmetric, so that it has a single hyperparameter η. The probability density function of such a distribution for $\eta < 1$ is illustrated in Fig. 2.6 to better understand why LDA works well in practice. If the vocabulary contains only three words, this density function can be represented in a barycentric coordinate system. This system has the advantage that coordinates of any point sum up to 1. If we assume, for instance, that topic k is drawn from the Dirichlet distribution as indicated in Fig. 2.6, the respective word probabilities are (approximately) given by $\boldsymbol{\beta}_k = (0.78, 0.08, 0.14)$. In this system of coordinates, the probability density function of the Dirichlet distribution is represented by different grayscale levels to indicate that its value increases near the corners and the edges if $\eta < 1$. Thus, the probability of a topic is higher near these corners and edges. This implies that topics contain only a few words with high probability, which is consistent with our intuition. For instance, the topic k in Fig. 2.6 is mainly determined by word 1 (78%). Since the per-topic word distribution is drawn from a symmetric Dirichlet distribution, the probability $p(\boldsymbol{\beta}_k|\eta)$ of a topic $\boldsymbol{\beta}_k$ is computed by the probability density function of this Dirichlet distribution with the hyperparameter η.

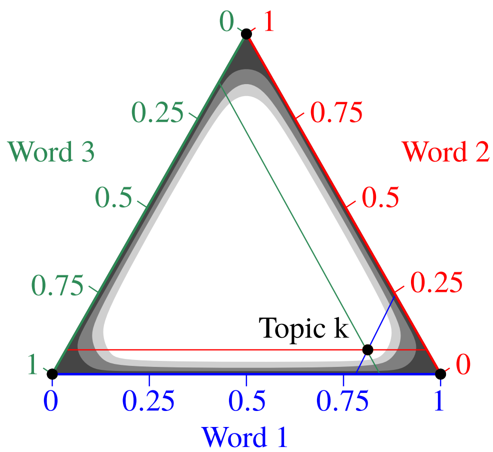


Figure 2.6: Probability distribution of a Dirichlet distribution in gray (white: low probability; black: high probability), which generates the probabilities of 3 words for a topic k.

So far, the topics have been generated. The next step is to create the *topic proportions* $\boldsymbol{\theta}_d$ for each document $d = 1, \ldots, M$, which explains why they are represented in a rectangle with subscript M in Fig. 2.5. These proportions were illustrated by the bar diagram in Fig. 2.4 for one document. Just as the distribution of words in a topic, the per-document topic proportions are sampled from a symmetric Dirichlet distribution with the hyperparameter $\alpha < 1$. The rationale behind this assumption is that documents are usually written about a few topics and not about all of them at once. Hence, the probability $p(\boldsymbol{\theta}_d|\alpha)$ can be computed by the probability density function of the symmetric Dirichlet distribution with the hyperparameter α.

Then, each word n in each document d containing N_d words is generated in two steps: first, a *topic* $z_{d,n}$ *is assigned* to the word $w_{d,n}$ based on the topic proportions $\boldsymbol{\theta}_d$ of the document.[14] This was previously represented by the color-filled circles in Fig. 2.4. While the topic proportions are drawn from a Dirichlet distribution, the topic assignments $z_{d,n}$ are drawn from a multinomial distribution based on the topic proportions $\boldsymbol{\theta}_d$. Hence, the probability of a topic assignment is written $p(z_{d,n}|\boldsymbol{\theta}_d)$, which can be computed by the probability mass function of the corresponding multinomial distribution. In simple terms, this function represents the probability of a biased K-sided die to fall on a specific side, i.e. a topic; the bias is introduced by the topic proportions, which favor the outcome of certain topics. Secondly, once the topic of a word is known, this *word is generated* by choosing a word within the selected topic $\boldsymbol{\beta}_{z_{d,n}}$. The corresponding probability $p(w_{d,n}|\boldsymbol{\beta}_{z_{d,n}})$ is also computed by the multinomial probability mass function since selecting the word within the topic is again like throwing a biased die, which is, however, V-sided this time. In fact, some words of the vocabulary have a higher probability within the topic and they are therefore more likely to occur.

As indicated in Fig. 2.5, the words $w_{d,n}$ and the hyperparameters α and η are the observed variables, while the distribution over words in a topic $\boldsymbol{\beta}_k$, the per-document topic proportions $\boldsymbol{\theta}_d$ and the topic assignments $z_{d,n}$ are the hidden variables.[15] To infer their values, the joint probability distribution of the entire model takes the form of the following equation (Blei, 2012), in which the notations $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\mathbf{z}$ and $\mathbf{w}$ were introduced for brevity to describe a full configuration of per-topic word distributions, per-document topic proportions, topic assignments and word selections in all documents. One should notice that this equation is an equivalent description of the model to the graphical representation in Fig. 2.5:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\alpha, \eta) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\eta) \prod_{d=1}^{M} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N_d} p(z_{d,n}|\boldsymbol{\theta}_d) p(w_{d,n}|\boldsymbol{\beta}_{z_{d,n}}) \tag{2.19}$$

The *likelihood* function of the corpus, like Eq. (2.18) for pLSA, can be derived from this distribution by marginalization (see Blei et al., 2003). This function is then again *maximized* to infer the values of the latent parameters. In consequence, these values are those for which

[14]Using the term "word" in this sentence might be confusing since the word has not yet been created. Alternatively, it could be said that a topic is assigned to the "position" n (of the future word) in document d.

[15]It is unclear why the hyperparameters were not shaded by Steyvers and Griffiths (2007), although these hyperparameters are user-specified, like the words in the corpus, and thus, observed.

the generation of the observed data is the most likely according to the previous probabilistic model. Various methods are suggested in the literature to perform this statistical inference, like variational expectation-maximization in Blei et al. (2003) or Gibbs sampling in Griffiths and Steyvers (2004); Resnik and Hardisty (2010), which is explained in most didactic introductions to LDA, e.g. Boyd-Graber (2018). One of the fastest algorithms is the online variational algorithm by Hoffman et al. (2010), which is implemented in the Python library *scikit-learn*.

Of the previous topic models, LDA seems to be best suited to discover topics in the data set of investment style news, which will be introduced in the following chapter. In fact, the model is built on a solid statistical foundation, unlike the matrix factorization approaches. Hence, the interpretation of topics and topic proportions in documents is more straightforward. In addition, it is based on a full generative probabilistic model, which includes the creation of topic proportions in contrast to pLSA. According to Blei et al. (2003), this reduces the susceptibility to overfitting of the model in comparison to pLSA. Moreover, it enables the assignment of probabilities to previously unseen documents, which was not possible in a natural way in pLSA. Finally, the introduction of Dirichlet priors encodes the intuition that documents are usually only about a few topics, and that topics can be described by a few words (Boyd-Graber et al., 2017). One should notice that pLSA is equivalent to LDA with maximum a posteriori estimation (instead of the maximum likelihood estimation) and uniform Dirichlet priors (Girolami and Kabán, 2003).[16] This uniform distribution suggests that the previous intuitions about documents and topics are not captured by pLSA. Due the previous reasons, *LDA is selected* to uncover the topics in the data of Chap. 3.

2.3 Latent Dirichlet allocation in finance

In this section, the literature about *latent Dirichlet allocation in finance* is analyzed to determine how to apply LDA at best to the investment style news of Chap. 3. This analysis is subdivided into the following sections: the context of LDA in finance and the corresponding data, its pre-processing, the implementation of the topic model and its parameters, and the post-processing of the results.

[16]Maximum a posteriori estimation is a concept of Bayesian statistics, which is defined in Murphy (2012). If $\theta \mapsto p(x|\theta)$ is a likelihood function with the observed data x and the unobserved parameter θ, the *maximum likelihood estimate* (ML) is $\hat{\theta}_{\mathrm{ML}} = \arg\max_\theta p(x|\theta)$. LDA and pLSA are based on this estimate. Alternatively, one can define the estimate $\hat{\theta}_{\mathrm{MAP}} = \arg\max_\theta p(\theta|x)$, which can be re-written by Bayes' theorem, if $p(\theta)$ is a prior distribution. Hence, the *maximum a posteriori estimation* (MAP) is $\hat{\theta}_{\mathrm{MAP}} = \arg\max_\theta p(x|\theta)p(\theta)$.

2.3.1 Context of LDA in finance and the corresponding data

Topic modeling by LDA in finance seems to be a fairly new field of research since a thorough review article about textual analysis in accounting and finance by Loughran and McDonald (2016) mentions only a single working paper of 2015 as one of the first applications of LDA. This paper, which was later published by Huang et al. (2018), examines the topical value added by prompt analyst reports with respect to transcripts of preceding quarterly earnings *conference calls* by comparing their topic proportions. Prior to the very exemplary introduction of LDA in the finance literature by the previous publication, Grafe (2010) was the first in our literature review to discover common topics among quarterly earnings conference calls by LDA.

The application of LDA is, however, not restricted to transcripts of conference calls in finance. Hence, a second subcategory is the analysis of the *scientific literature* itself. Moro et al. (2015) applied LDA to a subset of scientific articles selected by a keyword search in chosen journals to identify trends in applications of business intelligence in banking. In a similar way, Aziz et al. (2019) identified topics and their coverage over time in abstracts of academic articles about finance and machine learning.

Probably, the most significant number of articles in the finance literature including LDA focuses on topic modeling of the *various sections in 10-K forms*, which are annual reports about the financial performance of firms required by the SEC. In particular, *section 1A about risk factors* and *section 7 about management discussion and analysis (MD&A)* are examined. Concerning the former, Bao and Datta (2014) studied which topics in this risk disclosure might impact the stock return volatility to identify topics that investors actually perceive as risky. A similar analysis was also carried out by Israelsen (2014), who further regressed factor loadings (previously called sensitivities in Sec. 2.1.3) of the Fama-French 4-factor model with respect to disclosure frequencies of risk topics. Lopez-Lira (2019) built on the previous work to design a test that allows to classify identified topics into systematic and idiosyncratic risks, and he regressed excess returns with respect to risk weights to determine associated risk premiums of risk topics. Moreover, Hanley and Hoberg (2019) focussed only on risk disclosures of banks to identify arising risks in the financial sector. Besides the section 1A, Ball et al. (2015) and Hoberg and Lewis (2017) studied 10-K MD&A disclosures to explain the valuation of firms and to detect fraud, respectively. Finally, Dyer et al. (2017) determined topical trends in 10-Ks, which can mainly be explained by regulatory changes over time.

A fourth category can be defined by topic modeling of *news articles or equivalent disclosures* to take advantage of market inefficiencies when new information becomes available. Jin et al. (2013)

forecasted the evolution of foreign currencies by extracting selected topics from Bloomberg news and by interpreting their tone via sentiment analysis based on dictionnaries (see e.g. Loughran and McDonald, 2011). A similar approach was implemented by Larsen and Thorsrud (2017) but for Norwegian stocks based on the largest Norwegian business newspaper. Atkins et al. (2018) replaced the sentiment analysis by a Naïve-Bayes classifier trained on the basis of topic occurrences in Reuters US news and associated market movement. Instead of classical media, Feuerriegel et al. (2016) and Feuerriegel and Pröllochs (2018) used legally obligatory publications, like German ad hoc announcements and 8-K regulatory disclosures to study the effect of detected topics on German and US stock returns, respectively.

The previous references are summarized in Tab. 2.2 with some indication about the data that were analyzed. As mentioned previously, the most frequent data are 10-K forms, which were web scrapped from the SEC EDGAR database. Since *web scrapping*, i.e. the extraction of data from websites, became only possible with the availability of digital data, *sampling periods* usually start around the year 2000. The *size* of the data sets amounts on average to several thousand articles, which justifies the use of a machine learning method. Some sample sizes might seem inconsistent with the sampling period in Tab. 2.2, e.g. Israelsen (2014) and Hanley and Hoberg (2019). This inconsistency can be explained by various filtering methods. For instance, Hanley and Hoberg (2019) considered only 10-K filings of banks.

2.3.2 Data pre-processing

In Sec. 2.2.2, it was mentioned that textual data generally have to be *pre-processed* before they can be treated by LDA. On average, the literature is very vague about pre-processing, which can be different from one article to another. The following pre-processing steps can be found in the articles of Tab. 2.2:

- *Removal of irrelevant information*: numerous data sources contain irrelevant information for the topic identification, e.g. brokerage disclosures describing the stock-rating system or conflicts of interest in analyst reports (Huang et al., 2018), or web addresses (Atkins et al., 2018). This information can often be removed by text pattern searches via regular expressions. *Regular expressions* (regex) are sequences of characters to define text search strings (Jurafsky and Martin, 2019). For instance, the regex `/beg.n/` could be used to find the words within any character between “beg” and “n”, like “begin” or “begun”. Another example is `/\$[0-9]+\.[0-9][0-9]/`, which identifies dollar amounts like “$500.42”.
- *Combination of dependent words*: Huang et al. (2018) converted technical dependent words

Article	Data	Size	Period	Source
Grafe (2010)	Earnings call transcripts	3800	n/a	seekingalpha.com
Huang et al. (2018)	Earnings call transcripts	17,750	2003-2012	Thomson Reuters' StreetEvents
	Analyst reports	159,210	2003-2012	Thomson Reuters' Investext
Moro et al. (2015)	Scientific articles	219	2002-2013	Scientific journals
Aziz et al. (2019)	Scientific articles	5,123	1990-2018	Scientific journals
Bao and Datta (2014)	Secs. 1A in 10-K	14,799	2006-2010	SEC EDGAR database
Israelsen (2014)	Secs. 1A in 10-K	27,339	2006-2011	n/a (probably EDGAR)
Lopez-Lira (2019)	Secs. 1A in 10-K	79,304	2005-2019	SEC EDGAR database
Hanley and Hoberg (2019)	Secs. 1A in 10-K	10,558	1997-2015	SEC EDGAR database
Ball et al. (2015)	Secs. 7 in 10-K	52,835	1997-2011	SEC EDGAR database
Hoberg and Lewis (2017)	Secs. 7 in 10-K	55,666	1997-2011	SEC EDGAR database
Dyer et al. (2017)	10-Ks	75,991	1996-2013	SEC EDGAR database
Jin et al. (2013)	News articles	361,782	04/2010-03/2013	Bloomberg
Larsen and Thorsrud (2017)	News articles	459,745	05/1988-12/2014	Dagens Næringsliv newspaper
Atkins et al. (2018)	News articles	n/a	09/2011-09/2012	Reuters US news archive
Feuerriegel et al. (2016)	Ad hoc announcements	7645	01/2004-06/2011	DGAP information service
Feuerriegel and Pröllochs (2018)	8-K of NYSE firms	73,986	2004-2013	SEC EDGAR database

Table 2.2: Data in articles concerning LDA in finance with their definition, the sample size (after filtering), the sampling period and the data source. The studies are sorted according to the earlier order of presentation in the context section.

into one word, which is not separated by the tokenizer, to keep the initial meaning. For instance, "balance sheet" becomes "balance-sheet" or "earnings per share" is transformed to "EPS". An automatic method to create word combinations of co-occurring words, which is called "phrase modeling" by Lopez-Lira (2019), was suggested by this author but it depends on seemingly hard-to-estimate parameters.

- *Tokenization*: as mentioned in Sec. 2.2.2, tokenization is the separation of text into its individual terms (Jurafsky and Martin, 2019). The most basic tokenization algorithm consists in separating words based on whitespace characters between them. This operation is, however, problematic for word groups like "doesn't" or "they're". Hence, more sophisticated methods were developed. In addition, tokenization should be a fast operation because of the significant amount of data that usually has to be analyzed. For this reason, tokenizers are commonly built on deterministic regular expression rules. One of the most well-known tokenization standards is the Penn Treebank standard that is implemented in the Python library NLTK (Natural Language Toolkit; Bird et al., 2009), e.g. applied by Grafe (2010). A simpler tokenizer, i.e. the WordPunktTokenizer, which separates "doesn't" into [doesn, ', t] instead of [does, n't] according to the Penn Treebank tokenizer, was used by Atkins et al. (2018).

- *Case folding*: this step simply consists in replacing uppercase letters by their lowercase equivalent to prevent having two different entries in the document-term matrix for essentially the same word, e.g. one for “Market” and one for “market”.

- *Stemming*: morphologically different forms of a word might have almost the same meaning, like grammatically inflected words such as “market” and “markets”. To prevent having different terms of the same concept in the document-term matrix, inflected forms can be reduced to their roots. A crude heuristic method to achieve this objective is called stemming (Manning et al., 2008). It simply consists in chopping of word endings. The most common stemming algorithm is the Porter stemmer (Porter, 1980), which is based on a sequence of word modification rules, like “IES $\rightarrow$ I” that transforms “ponies” to “poni”. The Porter stemmer was used, amongst others, by Atkins et al. (2018), Aziz et al. (2019) and Feuerriegel and Pröllochs (2018). Stemming is fast due to its rules-based nature, but relatively aggressive, i.e. it easily transforms the meaning of words. For instance, the Porter stemmer transforms “university” to “univers” and “marketing” to “market”.

- *Lemmatization*: the previous shortcomings of stemming are alleviated by lemmatization. It consists in a proper “vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma” (Manning et al., 2008). In this way, it can transform “is” into “be”, or “saw” into “see” or “saw”, depending on additional information. Lemmatization was applied by Lopez-Lira (2019), while Huang et al. (2018) only transformed plural nouns to their singular form.

- *Removal of non-alphabetic characters and very short words*: numbers, punctuation and words with less than 3 letters (e.g. Atkins et al., 2018) have generally no real meaning in topic modeling and they are therefore removed.

- *Removal of stop words*: language includes high-frequency functional words, which contain no topical meaning, like “that” or “although”. These so-called *stop words* are removed from the corpus for the previous reason. Their removal is usually based on standard lists of stop words in the utilized text mining software. For instance, Feuerriegel and Pröllochs (2018) used the English stop word list in the text mining package of R (Feinerer et al., 2008).

- *Removal of names*: company names and tickers symbols were removed by Huang et al. (2018) with the intention of preventing companies from being detected as topics. Larsen and Thorsrud (2017) eliminated most common Norwegian surnames and given names, without, however, specifying their reasoning.

- *Removal of very rare and very frequent words*: Aziz et al. (2019) removed the top 10% of words according to a tf-idf ranking, and words that appear less than 5 times in the corpus. Similarly, Larsen and Thorsrud (2017) selected the 250,000 terms with the highest tf-idf score. Unfortunately, no justification of this pre-processing step was provided. Likewise, Grafe (2010) explained that he used arbitrary thresholds to remove words that occur in more than 50% or less than 2% of the documents. Atkins et al. (2018) applied a similar filter "to avoid difficulties caused by the dominance of frequently used words in learning and rare words only appearing in the training and test sets", which is rather vague (what difficulties? what is the problem with rare words?). Feuerriegel and Pröllochs (2018) eliminated infrequent words that appear in less than 5% of the documents with the objective of reducing the size of the document-term matrix. Hence, there seems to be no other reason to eliminate words based on their frequency than computational overhead, i.e. computation time, which is confirmed by Blei and Lafferty (2009).
- *Construction of the document-term matrix with term or tf-idf counts*: the final step consists in the construction of the document-term matrix, which is seemingly only based on tf-idf counts in Feuerriegel et al. (2016).

2.3.3 Implementation of the topic model and its parameters

On the basis of the pre-processed data, topics can be determined by an implementation of a *topic model* (software) and its *parameters*. Unfortunately, not all articles in Tab. 2.2 explain these elements, although they are a fundamental requirement for reaching their conclusions. Tab. 2.3 mentions the topic model software in the articles of Tab. 2.2 to guide our selection in Chap. 4.

Besides the software, its parameters are crucial in topic modeling. The most significant parameter is certainly the *number of topics* K, which is not chosen by the algorithm but by the user. Various methods are suggested in the literature to pick its value, i.a.:

- *Topic coherence*: the most straightforward way of selecting the number of topics K is to extract topics for various values of K and to assess their coherence by taking a look at the words with the highest probability for each topic. For instance, Feuerriegel et al. (2016) tested various numbers of topics from 5 to 150 after finally choosing 40. In this way, they and other authors observed that topics become very broad if K decreases, so that topics, which one would expect to be separated, are merged. If K increases, however, the topics

Article	Topic model software	Topics K	α	η
Grafe (2010)	n/a	5, 10	n/a	n/a
Huang et al. (2018)	Stanford Topic Modeling Toolbox	60^{*}	0.1	0.01
Moro et al. (2015)	R library topicmodels	19	n/a	n/a
Aziz et al. (2019)	R library topicmodels	20	$50/K$	n/a
Bao and Datta (2014)	Custom implementation (probably)	30	$50/K$	0.1
Israelsen (2014)	n/a	30	n/a	n/a
Lopez-Lira (2019)	PYTHON library gensim	25	n/a	n/a
Hanley and Hoberg (2019)	metaHeuristica (proprietary)	25	n/a	n/a
Ball et al. (2015)	metaHeuristica (proprietary)	100	n/a	n/a
Hoberg and Lewis (2017)	metaHeuristica (proprietary)	75^{**}	n/a	n/a
Dyer et al. (2017)	MALLET	150	n/a	n/a
Jin et al. (2013)	n/a	30	n/a	n/a
Larsen and Thorsrud (2017)	Custom implementation (probably)	80	$50/K$	$200/N$
Atkins et al. (2018)	PYTHON library gensim	100^{***}	n/a	n/a
Feuerriegel et al. (2016)	R library topicmodels	40	n/a	n/a
Feuerriegel and Pröllochs (2018)	R library topicmodels	20	n/a	n/a

Table 2.3: Topic model software, number of topics K and the hyperparameters of the Dirichlet priors for the articles in Tab. 2.2 (same order). Remarks: *: per industry of 5 industries; **: probably rather 100 since the article is based on Ball et al. (2015), who detected 100 topics and discarded 25; ***: the number of text body topics, since 20 topics were extracted from the titles; N is the total number of words in all documents.

become more and more specific until not being interpretable anymore. Topic coherence is further described in the following subsection on the analysis of results.

- *Perplexity*: previously, in the description of the LDA method, it was explained that this method is based on the maximization of a likelihood function. In fact, the latent parameters of the probabilistic generative model are chosen so that the probability of generating the analyzed corpus by this model is maximized. The likelihood function takes the form of the product $\prod_{d=1}^{M} p(\mathbf{w}_d)$, where $p(\mathbf{w}_d)$ is the probability of generating the document d in the corpus with $\mathbf{w}_d$ representing the words in this document. Then, the log-likelihood function, which is simply the logarithm of the likelihood function, can be written as the sum $\sum_{d=1}^{M} \log p(\mathbf{w}_d)$. Based on the theory of language model evaluation (Jurafsky and Martin, 2019), Blei et al. (2003) define the *perplexity* for a test corpus $\mathcal{D}_{\text{test}}$ in order to measure the performance of a trained LDA model as follows:

$$\text{perplexity}(\mathcal{D}_{\text{test}}) = \exp\left[-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right] \tag{2.20}$$

According to our explanation of the log-likelihood function, the perplexity decreases when the model becomes better, i.e. more likely, at generating the test corpus since the likelihood function increases in this case. In other words, the perplexity measures the ability of a trained LDA model to predict the words in the test corpus. This ability increases when the perplexity decreases. Although Blei et al. (2003) divides his data into a training and a test set, the perplexity in the literature about LDA in finance seems to be directly evaluated on the training set since the division into these sets is never mentioned in the articles of Tab. 2.2.

In practice, see e.g. Fig. B.1 in Huang et al. (2018), the perplexity decreases quickly with the number of topics K for the first few topics and the incremental improvement decreases. Hence, a heuristic rule is to select K such that the improvement for a larger number of topics than K becomes relatively insignificant.

As shown in Tab. 2.3, the number of topics ranges from 5 to 150 in this literature review. The minimum, median, mean and maximum numbers of documents per topic are respectively about 12, 500, 1900 and 12000, so that the range is quite wide.

Besides the number of topics, the *hyperparameters* of the Dirichlet priors have to be chosen. Tab. 2.3 shows that only very few publications mention these parameters. The corresponding values of the parameters can be traced back to Griffiths and Steyvers (2004) ($\alpha = 50/K$, $\eta = 0.1$), Steyvers and Griffiths (2007) ($\alpha = 50/K$, $\eta = 0.01$) and Kaplan and Vakili (2015) ($\alpha = 0.1$, $\eta = 0.01$), who chose these values because they return satisfying results in numerous cases. It is important to mention that $K \geq 50$ in Griffiths and Steyvers (2004), so that $\alpha \leq 1$. In Aziz et al. (2019) and Bao and Datta (2014), α is, however, greater than 1, which is very surprising. In fact, the probability density function of the symmetric Dirichlet distribution does not encourage documents with few topics under these circumstances, which is why one might question whether these authors were aware of the consequences of their choice.

2.3.4 Post-processing of the results

As mentioned in Sec. 2.2.6, the results of LDA are per-topic word distributions and per-document topic proportions. In practice, these results take the form of the matrices $\boldsymbol{\beta} = (\beta_{kw})$ and $\boldsymbol{\theta} = (\theta_{dk})$, such that β_{kw} is the probability of term w in topic k, and θ_{dk} is the probability of topic k in document d. Based on the articles in Tab. 2.2, the post-processing of these matrices can be divided into three steps: topic labeling, validation and the further analysis of the results. This last step is not explained in more detail hereafter since it simply consists in inferring additional knowledge of the per-topic word distributions and per-document topic proportions, e.g. by introducing the topic frequencies in regression models as in Feuerriegel et al. (2016).

Topic labeling

Topic labeling consists in finding a generic term, i.e. a label, for a topic. Topics created by LDA are commonly manually labeled by reading the words with the highest probability within the topic (Chang et al., 2009), e.g. the 20 top words in Huang et al. (2018). An example of a label is the word "fruit", if "apple, banana, kiwi, orange" are the highest-probability words. Topic labeling can be facilitated by various means:

- *Domain knowledge of experts*: Bao and Datta (2014) relied on topic labels that were created by experts who read hundreds of 10-K forms in the past.
- *Word clouds*: topics can be represented visually as word clouds, in which the size of a word increases with its probability in the topic, e.g. used by Bao and Datta (2014) and Lopez-Lira (2019).
- *Representative documents or paragraphs*: besides the per-topic word distributions, the per-document topic distributions are computed by LDA. Hence, a document or a paragraph with a very high proportion of a specific topic can be read to better understand the underlying meaning of a topic so that it can be labeled, as in Aziz et al. (2019), Hoberg and Lewis (2017) and Dyer et al. (2017).
- *LDAvis*: instead of the previous static methods to label topics, LDAvis is a web-based interactive visualization tool of topics, which further introduces an improved measure of word relevance to a topic (Sievert and Shirley, 2014). It was applied to LDA topics in finance by Feuerriegel and Pröllochs (2018) and it is explained in more detail in Sec. 4.4.

Validation

Validation focuses on checking whether LDA results are in line with the expectations for this method, i.e. whether the words in each topic actually form coherent topics that are included in the corpus. The following methods were applied in the previous literature about LDA in finance (Tab. 2.2):

- *Word intrusion test*: this test was introduced by Chang et al. (2009) and used by Bao and Datta (2014) to evaluate the semantic coherence of topics. In simple terms, it consists in presenting a few high-probability words of a topic and an intrusion word, which does not belong to the topic (like "cat" in our previous "fruit" topic), to a human evaluator. The more likely the evaluator is to detect the intrusion word, the more coherent is the topic.

- *Topic-dependence on external events*: for topics to be valid, it seems reasonable to expect their coverage frequency to be dependent on external events. For instance, (1) Bao and Datta (2014) detected an increase of the topic "macroeconomic risk" in 10-K filings around 2009, (2) Huang et al. (2018) observed a recent increase of the topics "smartphone business" and "wireless subscribers" in analyst reports, and (3) Aziz et al. (2019) noticed a shift from modeling-based topics towards data-based topics in the literature about machine learning in finance over time.

- *Manual topic assignment*: Huang et al. (2018) provided a human coder with topic labels that were created by an expert via LDA results. Based on these labels, the coder assigned topics to sentences in a sub-sample of the initial data set. Manual topic assignments were consistent in about 65% of the sentences with the LDA topics for each sentence, which is significantly greater than the 5% rate reached by random assignments in this context.

Chapter 3

Data

The data in this document were provided by Gillain since this Master's thesis is written within the context of his PhD thesis (Gillain et al., 2019, 2020a,b; Gillain, 2020). More precisely, 9 magazines of 5 media groups were selected because their mission statements include the compilation of information targeting *financial decision makers, and especially institutional investors*. The different media groups and the corresponding magazines are briefly described in Tab. 3.1 based on the few information available.

Magazines targeting institutional investors, i.e. entities, which invest money on behalf of others, like banks, mutual funds or pension funds, were essentially selected for two reasons (Brealey et al., 2011; Vernimmen et al., 2018): on the one hand, information concerning equity is described at an *aggregate level*. While real-time financial news from the Dow Jones Newswires, Bloomberg or Reuters are commonly about individual companies, it is expected that institutional media synthesize this information at a higher level of abstraction, like macroeconomics or asset allocation. Hence, it is anticipated that *information about style investing* (Sec. 2.1.4) is contained in these media. On the other hand, the transaction volume of institutional investors is far more important than the volume of retail investors (Davis Evans, 2009). Thus, to study the influence of news coverage on smart beta ETF flows in the long run, it seems appropriate to consider news, which most likely impact these flows significantly (Clifford et al., 2014).

The data were collected by *web scraping* the websites of the magazines via the Python modules *Beautiful Soup* (Richardson, 2020) and *Scrapy* (Scrapinghub, 2020). These programs first fetch web pages and then parse the underlying code to extract useful information. In total, 108,638 articles from January 1996 to July 2018 were collected with the following attributes: name of the

Media groups	Magazines
Euromoney Institutional Investor PLC "Euromoney is a global information services business providing essential B2B information to global and specialist markets. Euromoney provides price discovery, market intelligence and events across our segments. Euromoney is listed on the London Stock Exchange and is a member of the FTSE 250 share index." (Euromoney Institutional Investor, 2020)	**Euromoney** "Euromoney magazine was created in 1969 to cover the re-emergence of the international cross-border capital markets. The euromarket, after which the magazine is named, is the predecessor to today's mainstream *global capital markets*. Euromoney reported on, and championed, this market and its growth, in the process becoming the prime magazine of the *wholesale financial world*, its *institutions* and its users." (Euromoney, 2020) **Institutional Investor** "For 50 years, Institutional Investor has built its reputation on providing award-winning editorial for the world's most influential *decision makers in global asset management and banking*. This prestigious audience relies on Institutional Investor to provide in-depth coverage of the people and events impacting the world's economy and all facets of *institutional asset management*." (Institutional Investor, 2020)
FTAdvisor "FTAdviser.com is dedicated to the *financial intermediary market covering investments, mortgages, pensions, insurance, regulation and other key issues*." (FTAdviser, 2020)	**Financial Adviser** "The premier weekly newspaper for the UK´s *financial intermediary community*, Financial Adviser was launched in 1988 after the Financial Services Act 1986 defined for the first time the role of the independent financial adviser. Financial Adviser offers comprehensive and in–depth coverage of the retail finance landscape." (FTAdviser, 2020)
Global Fund Media "Founded in 2002, Global Fund Media publishes seven specialist newswires covering all asset classes within the *institutional investor* marketplace." (Global Fund Media, 2020)	**AlphaQ** "AlphaQ is also a subscription-based online analytical, product development and marketing resource for *institutional investors, wealth advisers and investment managers*, allowing them to share best-in-class ideas and strategies with their peers through the bimonthly journal." (AlphaQ, 2015) **Institutional Asset Manager** "Institutional Asset Manager provides news and features and reports for the *institutional investor* market place [...] covering *institutional pensions and managed funds*." (Institutional Asset Manager, 2020) **Wealth Adviser** "Wealth Adviser services private client wealth managers, family offices, trustees and their *investment advisers* with knowledge on assets across all asset classes." (Wealth Adviser, 2020)

Table 3.1: Description of selected media groups and corresponding magazines.

Media groups	Magazines
IPE International Publishers Ltd "IPE International Publishers Ltd, an independently-owned company founded in July 1996" (IPE International Publishers Ltd, 2020)	**Investment & Pension Europe (IPE)** "IPE is the leading European publication for *institutional investors* and those running *pension funds*." (IPE International Publishers Ltd, 2020)
Institutional Shareholder Services group of companies (ISS) "Founded in 1985, the Institutional Shareholder Services group of companies ("ISS") empowers investors and companies to build for long-term and sustainable growth by providing high-quality data, analytics, and insight. With nearly 2,000 employees spread across 30 U.S. and international locations, ISS is today the world's leading provider of corporate governance and responsible investment solutions, market intelligence and fund services, and events and editorial content for *institutional investors* and corporations, globally." (ISS, 2020a)	**PlanSponsor** "As the nation's leading authority on retirement and benefits programs, PLANSPONSOR is dedicated to helping employers navigate the complex world of *retirement plan* design and strategy. No other media source offers such a clear path to reach this influential group of *retirement plan decision makers* through an award-winning magazine, website, newsletters, events, multimedia and social connections." (ISS, 2020b) **PlanAdviser** "PLANADVISER with its reputation for editorial integrity, objectivity, and leadership, is the trusted information and solutions resource for America's *retirement benefits decision makers*. PLANADVISER is the only magazine to address the specific needs and concerns of *advisers* who specialize in the sale and servicing of *institutional retirement plans*, including 401(k), 403(b), 457 and defined benefit (DB) plans." (ISS, 2020b)

Table 3.1: Description of selected media groups and corresponding magazines (cont.).

magazine, date, title, author, section and textual content. The format of one of these articles is illustrated in appendix A.

Since this thesis focuses on topic modeling of investment style news, the news related to the *small-cap style* were chosen from the initial data set by Gillain via a lexicon-based classifier (Gillain et al., 2019, 2020b). More specifically, if an article contains the variation "smallcap", "small-cap" or "small cap" of "small cap" or its uppercase or plural form (e.g. "Small Caps"), this article is extracted from the initial data set and assumed to be related to this investment style. In addition to the variations of "small cap", articles containing similar variations of "micro cap" and "mid cap" were also extracted from the initial data set since they are usually grouped with small caps. Moreover, the selected news were *restricted to the period* of January 1st, 2010 to July 12, 2018 because the magazines with most articles (Institutional Investor, Wealth Adviser, PlanSponsor) only have a representative number of articles since 2010, and because the news were collected until about half of July 2018.

The resulting data set concerning small-cap investing over the previous period contains 1720 articles. Fig. 3.1 illustrates the cumulated number of articles per magazine for each year. One notices that the total number of articles per year is about 200, except for the last year due to the time restriction. Moreover, some magazines, like the Wealth Adviser, contribute numerous articles to the data set, while other, like AlphaQ, are rare. This could be explained by some magazines, i.e. the Financial Adviser, AlphaQ and the Institutional Asset Manager, only being published on the internet for more recent editions.

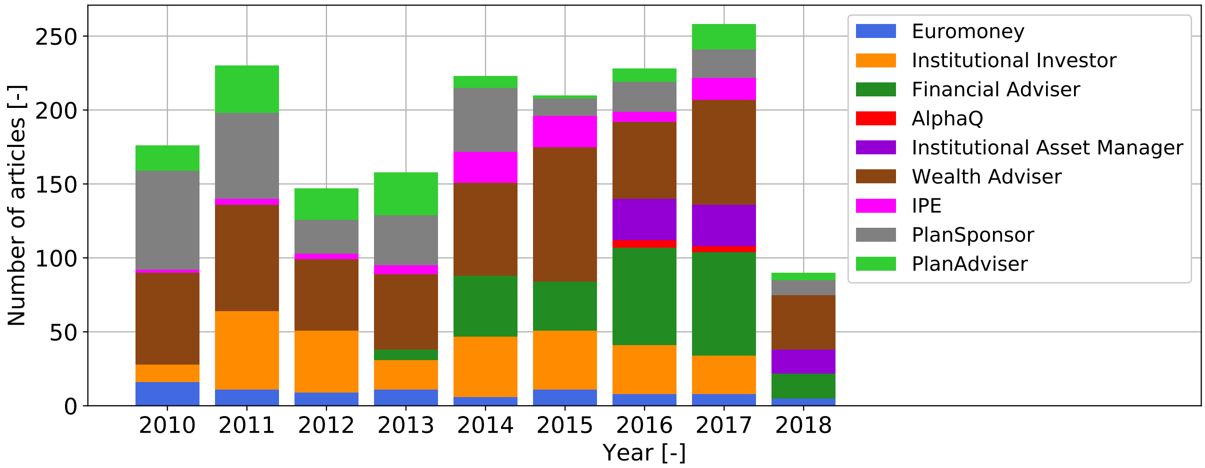


Figure 3.1: Cumulated number of articles per journal per year.

Finally, one should notice that the *titles* are added to the corresponding text bodies of the articles, which will be analyzed by the topic model. This seems to be a reasonable operation since titles usually summarize the content of the article.

Chapter 4

Methodology

This chapter focuses on the *methodology* to *identify topics* in the investment style news of the previous chapter. Furthermore, it highlights how these *news can be clustered according to their topics* to determine the coverage of a specific topic, e.g. via the number of articles about this topic, in a given magazine or during a particular period.

Because of the relatively thorough explanation of key concepts, like latent Dirichlet allocation (LDA), in Chap. 2 for better understanding, these explanations are *not repeated* in this chapter but simply referenced for the sake of brevity. Moreover, the methodology is illustrated on the basis of the popular *20 newsgroups corpus* in order to present this methodology in a more didactic way and to simultaneously validate it.

The methodology is explained by first introducing the *programming environment*, that is required for more advanced data processing like topic modeling, and the *20 newsgroups corpus*. Then, the text mining procedure is described by focusing on the *pre-processing* of the news, the specific *topic model*, which is applied, and the *post-processing* of the results.

4.1 Programming environment and 20 newsgroups corpus

Concerning the *programming environment*, the PYTHON programming language is selected because of its ease of use and the significant number of available libraries, also known as modules. The most popular PYTHON library for natural language processing (NLP) is the *Natural Language Toolkit* (NLTK; Bird et al., 2009), which is used for all classical NLP operations, e.g. tokenization and lemmatization. Since no implementation of LDA exists in NLTK and due to additional reasons that are explained in Sec. 4.3, the LDA implementation in the library *scikit-learn* (Pedregosa et al., 2011) is selected. This library is one of the most popular machine learning libraries in PYTHON. Besides NLTK and scikit-learn, the libraries *matplotlib* (plots), *re* (regular expressions), *NumPy* (numerical computing) and *pyLDAvis* (post-processing LDA results) were used to create this document.

A standard PYTHON *analysis script of the investment style news* and our *topic modeling module* can be found in appendix B. This module provides access to the data structures and functions required to pre-process the news and post-process the results. In particular, it allows to create various HTML documents to better understand these results. Examples of these documents are illustrated in appendix C based on the article that previously served as an illustration of the generative process in probabilistic topic models (Fig. 2.4).

The *20 newsgroups corpus* is a collection of about 19,000 messages of newsgroups, i.e. online forums, evenly split across 20 different newsgroups (Rennie, 2008). Each of these newsgroups has a topic like baseball, Christian religion, motorcycles or space. This information has the advantage that a topical ground truth, as Sievert and Shirley (2014) call it, is known for each document. For instance, the following paragraph is a message of the newsgroup about space:

```
With the continuin talk about the "End of the Space Age" and complaints
by government over the large cost, why not try something I read about
that might just work.
Announce that a reward of $1 billion would go to the first corporation
who successfully keeps at least 1 person alive on the moon for a year.
Then you'd see some of the inexpensive but not popular technologies begin
to be developed.  THere'd be a different kind of space race then!
```

The messages of the previous four topics are selected to illustrate the following explanations and to validate the method.

4.2 Pre-processing the corpus

As stated in Sec. 2.3.2, various pre-processing steps of the textual data are required to build the document-term matrix that is analyzed by the topic model. In the framework of this study, the following operations are applied to the data in the respective order. The corresponding code can be found in the `prepro_corpus()` function in appendix B. Only operations, which include special choices with respect to those in Sec. 2.3.2, are explained in more detail:

- *Removal of irrelevant information*: web and e-mail addresses, remnants of webscrapping, like "#paragraph#", as well as general information following (and including) the expression "For more information" are removed.

- *Tokenization*: NLTK's recommended tokenizer is used to separate the running text into tokens.[1] Since it does not separate quite frequent word combinations between sentences like "initiatives.Although" into "initiatives . Although", this separation is explicitly added. By analogy, words joined by a hyphen (e.g. five-year) or a slash (e.g. broker/dealer) are also separated.

- *Case folding*

- *Lemmatization with POS-tagging*: lemmatization is selected because of its superiority over stemming and because of the reasonable size of the corpus so that lemmatization is not too time-intensive.[2] NLTK's lemmatizer is based on WordNet (Fellbaum, 1998), which is a database of lexical relations, so that lemmas can be determined by a combination of morphology functions and look-ups. In addition, part-of-speech tagging (POS-tagging), which consists in labeling words according to their class (Jurafsky and Martin, 2019), is applied to improve lemmatization.[3] In this context, POS-tagging is limited to the most fundamental grammatical classes, i.e. nouns, verbs, adjectives and adverbs, to lemmatize more consistently words like "saw" in sentences such as "A saw is a tool." (saw) or "She saw the sun." (see).

- *Removal of non-alphabetic characters and very short words*: words with less than 3 letters are removed as well as characters different from those in the English alphabet.

- *Removal of stop words*: words in NLTK's English stop word list, which contains 318 words, are removed from the corpus.

- *Document-term matrix*: this matrix is created by simple term counts. Tf-idf counts were also tested but the topic model (LDA) then returns a few uninterpretable topics with high-probability words that are very rare in the corpus. The exact reason behind this observation is unclear to the author. Tf-idf weighting is possibly inconsistent with the probabilistic

[1]Precisely, an improved `TreebankWordTokenizer` along with the `PunktSentenceTokenizer` for the English language.

[2]The total computation time of the standard analysis script in appendix B, which includes pre-processing, topic modeling and post-processing, is about 2.5 minutes.

[3]In retrospect, our approach could further be improved by POS-tagging before case folding so that the algorithm could take into account the additional information of uppercase letters.

generative model that is adopted to determine the topics. In this model, a word is either selected or not (integer count), i.e. it cannot occur 0.4 times in a document, which is possible by tf-idf.

If the previous list is compared to the list about pre-processing operations in the literature in Sec. 2.3.2, one notices that no words are deleted because of their *low/high frequency of occurrence*. The reason behind this decision is too use as much topical information as available. For instance, a word might be very rare but it occurs only in a specific topic. Hence, deleting this word deletes information that could be used to detect topics. Moreover, the computation time is not excessively long, so that a reduction of the document-term matrix is not required. And finally, frequent words, which could render topic labeling difficult, are penalized according to a relevance measure in Sec. 4.4, so that their removal also seems to be no requirement.

If the previous operations are applied to the *example* of the space newsgroup, the words highlighted in yellow are kept:

```
With the continuin talk about the "End of the Space Age" and complaints
by government over the large cost, why not try something I read about
that might just work.
Announce that a reward of $1 billion would go to the first corporation
who successfully keeps at least 1 person alive on the moon for a year.
Then you'd see some of the inexpensive but not popular technologies begin
to be developed. THere'd be a different kind of space race then!
```

The corresponding *tokens* after pre-processing are the following:

```
continuin talk end space age complaint government large cost try read just
work announce reward billion corporation successfully person alive moon year
inexpensive popular technology begin develop different kind space race
```

Besides the significant removal of stop words, one should notice the lemmatization of words like "complaints" or "technologies" to their singular forms. A similar example is available in Sec. C.1 for an articles in the corpus of investment style news to illustrate the HTML file that is created by our Python function `write_prepro_html()` in appendix B.

The *corpus of the investment style news* contains 1720 documents, 612,522 tokens and 22,210 different tokens after the previous pre-processing.

4.3 Topic model

As explained at the end of Sec. 2.2.6, *latent Dirichlet allocation* (LDA) seems to be the best topic model within the set of models that are presented in Sec. 2.2. Moreover, the version with symmetric Dirichlet priors for the per-topic word distributions and per-document topic proportions is selected. This version seems to be the most popular one among the authors, who went a bit further in trying to understand LDA, i.e. those who specified the model parameters, according to Tab. 2.3.

The *input parameters* of this LDA algorithm are essentially the corpus, the number of topics K, the hyperparameter α of the prior regarding the per-topic word distributions and the hyperparameter η of the prior regarding the per-document topic proportions. The number of topics is determined by the perplexity (Sec. 2.3.3) and a manual evaluation of topic coherence. The values of the hyperparameters initially take the default values of the implementation and then some variations consistent with the literature are tested. In addition, several *numerical parameters*, like the maximum number of iterations (of the expectation-maximization loop) until convergence, have to be specified. The default values of these parameters in the selected implementation are used.[4] The *results* of the algorithm are the per-topic word distributions $\boldsymbol{\beta}$ and the per-document topic distributions $\boldsymbol{\theta}$.

Concerning the *implementation*, Tab. 2.3 contains only one Python library including the LDA algorithm, i.e. gensim (Řehůřek and Sojka, 2010). Gensim is a library specialized in topic models with an LDA model based on the implementation of the online LDA algorithm by Hoffman et al. (2010). The same implementation is the basis of the LDA method in the machine learning library scikit-learn (Pedregosa et al., 2011). Since the research in this document is a first feasibility study about topic modeling of investment style news, it was preferred to select the less specialized library, i.e. scikit-learn; any later transfer to gensim is obviously possible due to similar APIs (application programming interfaces).[5]

Finally, our *example* about the newsgroup data can be used to check whether the *perplexity* is a good indicator to choose the number of topics. Fig. 4.1 shows the perplexity as a function of this number for the default parameters in scikit-learn, i.e. $\alpha = 1/K$, $\eta = 1/K$. The default maximum number of iteration had, however, to be increased from 10 to 20 to obtain this figure so that the

[4]It is, however, verified, that the perplexity does not decrease significantly anymore from one iteration to the next close to the final iteration, which can be interpreted as the convergence of the algorithm.

[5]The basic features of both libraries are the same but gensim offers more choice, like asymmetric priors. Notice also that the batch version of the LDA algorithm in Hoffman et al. (2010) is selected by default in the LDA method of scikit-learn.

perplexity becomes stationary, i.e. "converges". It can be seen that the perplexity is minimal for $K = 4$, which is precisely the number of different newsgroups that were selected, i.e. baseball, Christian religion, motorcycles or space. Hence, the perplexity seems to be a good indicator to choose the number of topics.[6]

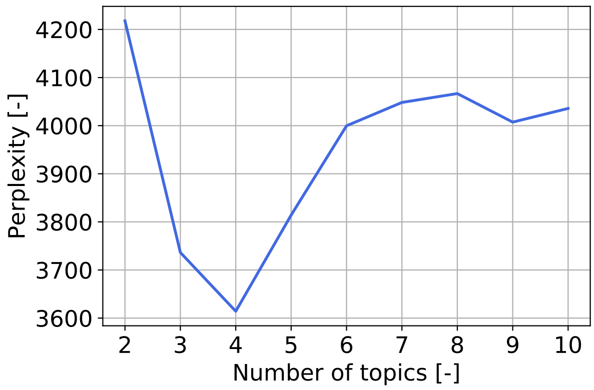


Figure 4.1: Perplexity as a function of the number of topics for the newsgroup data set about 4 topics (default parameters, i.e. $\alpha = 1/K$, $\eta = 1/K$, but 20 iterations instead of 10).

4.4 Post-processing the results

Various *post-processing methods* were explained in the literature review (Sec. 2.3.4) to label topics, to validate the method and to further analyze the results. In this section, some of these methods and others are explained in more detail to *label the topics in the investment style news* and to determine the *topic coverage in each magazine*, as well as the *importance of topics over time*.

4.4.1 Topic labeling

As mentioned in Sec. 2.3.4, topics are usually labeled on the basis of the *first few high-probability words of each topic*, which are known by the matrix $\boldsymbol{\beta}$ of per-topic word distributions. In the

[6]The variation of the perplexity with the number of topics has a similar trend when the hyperparameters take the constant values $\alpha = 0.1$ and $\eta = 0.01$, which are used by Huang et al. (2018). The rebound after $K = 4$ is, however, smaller with these latter values.

newsgroup example, the following word lists can be generated after applying LDA with $K = 4$:[7]

```
Topic 1:  bike like just know make think dod look time motorcycle
Topic 2:  space launch nasa use satellite orbit year data mission earth
Topic 3:  year game good win think team run player hit like
Topic 4:  god say people know christian jesus think believe church make
```

A quick look at these words, immediately allows us to label the topics as motorcycles, space, baseball and Christian religion in this order by the knowledge of the ground truth, i.e. the labels of the newsgroups.[8] Without this knowledge (and without more high-probability topic words), topic 3 could possibly also be football. Moreover, the previous topics contain a lot of stop words that are not included in NLTK's list of stop words, e.g. "like", "just", "know". So, if the ground truth was not known in advance, as for the investment style news of the previous chapter, topic labeling would be much harder. For this reason, various methods are introduced to *facilitate topic labeling*, as mentioned in Sec. 2.3.4.

Domain knowledge of experts

One of these methods is *domain knowledge of experts*. Considering that topics have not yet been suggested for the previous data set in the literature (to the best of our knowledge) and considering that the objective of this thesis is precisely to not read hundreds of articles manually, only few topical information is known.

The best expert about this data set is certainly *Gillain*, who manually labeled a significant number of articles to train machine learning classifiers (Gillain et al., 2019). Gillain noticed that some magazines seem to have a dominant topic (Gillain and Lambert, 2020): "strategy" in the Institutional Investor, "past performance" in PlanSponsor and "new funds" in the Wealth Adviser. In addition, the *category labels of articles on the websites* of the magazines are read to facilitate topic labeling.

LDAvis

Apart from expert knowledge, *data visualization tools* like word clouds or LDAvis simplify data labeling. In this document, we focus on LDAvis that offers much more ways to analyze

[7]The remaining parameters are the same as those required to create Fig. 4.1. The colors of these topics have no meaning so far but they will become useful at the end of this section.

[8]The newsgroup messages were shuffled before using LDA to ensure that finding such good results is not simply due to luck.

topic-relevant word lists than word clouds. Before addressing its visualization component, the relevance measure in LDAvis is described. As mentioned previously, the interpretation of word lists suffers from frequent terms without topical information. To penalize these terms, Sievert and Shirley (2014) introduced the following *relevance measure*, which depends on the user-specified parameter λ:

$$r(w, k|\lambda) = \lambda \log(\beta_{kw}) + (1-\lambda) \log\left(\frac{\beta_{kw}}{p_w}\right) \quad \text{with} \quad p_w = \frac{\sum_{d=1}^{M} X_{dw}}{\sum_{d=1}^{M} \sum_{v=1}^{V} X_{dv}} \tag{4.1}$$

If $\lambda = 1$, one recovers the classical high-probability ranking since β_{kw} is the probability of term w in topic k. However, if λ decreases, the importance of $\log(\beta_{kw}/p_w)$ increases. This term penalizes high-frequency words in the topic. More precisely, p_w is the marginal probability of term w in the corpus, i.e. the corpus-wide frequency of this term divided by the total count of terms since the component X_{dw} of the document-term matrix is the count of term w in document d. In consequence, the quantity β_{kw}/p_w, which is called *lift*, decreases, if the frequency of term w in the corpus increases. The optimal value of λ was found to be 0.6 by a user study. In this study, participants were asked to label topics based on words lists that were generated for random values of λ. Since the ground truth was known, λ could be determined by selecting the value for which labels corresponded most frequently to this truth.

If the previous relevance measure is applied to the *newsgroup example* with $\lambda = 0.6$, the top words of each topic are the following. In comparison to the previous lists, irrelevant words either disappear or they have a lower ranking, thus, occurring later in these lists:

```
Topic 1:  bike dod motorcycle ride like just dog rid helmet buy
Topic 2:  space launch nasa satellite orbit data mission program shuttle earth
Topic 3:  game year win team player run good hit play baseball
Topic 4:  god say christian jesus people church believe know christ think
```

In addition to the improved word lists, LDAvis allows to *visually interact* with them in the browser based on javascript in an HTML document created by the PYTHON library *pyLDAvis* (Sievert and Shirley, 2014). An example is illustrated in the appendix (Sec. C.2) for the investment style news. For instance, the overall term frequencies in the corpus (blue bars) as well as the estimated term frequencies in a specified topic (red bars) are shown. By selecting a word, its importance in all topics is represented by the size of circles corresponding to the topics.

Representative titles and documents

Besides the results derived from the per-topic word distributions $\boldsymbol{\beta}$, those derived from the per-document topic proportions $\boldsymbol{\theta}$ can also be used to facilitate topic labeling. Hence, the *titles of articles with the highest proportions for a given topic* can be read to infer a label of this topic. An example is shown in Sec. C.3 of the appendix for investment style news, which was created by our PYTHON function `write_title_hmtl()` in appendix B.

In addition to reading documents with high topic proportions for a single topic, their *words can be color-coded* depending on the topics of these words. Two different ways can be considered to determine the color, i.e. the topic, of a word. On the one hand, the topic k with the highest probability for a given term w can be selected independently of the topic proportions of the document, in which it is located:

$$k^* = \arg\max_k \beta_{kw} \tag{4.2}$$

Based on this criterion, our previous newsgroup message is colored as follows, thus, mainly containing words of the topic "space":

```
With the continuin talk about the "End of the Space Age" and complaints
by government over the large cost, why not try something I read about
that might just work.
Announce that a reward of $1 billion would go to the first corporation
who successfully keeps at least 1 person alive on the moon for a year.
Then you'd see some of the inexpensive but not popular technologies begin
to be developed.  THere'd be a different kind of space race then!
```

This criterion disregards, however, the context of the words, which is provided by the remaining words in the document. For instance, a "bank" in a document about loans is likely a financial institution, while a "bank" in a document about rivers is the land along this river. Similarly, the word "play" might stand for a theater play or a verb representing the engagement in a recreational activity. This coexistence of multiple meanings for a word is known as *polysemy* (Steyvers and Griffiths, 2007). Hence, the topic of a word can be chosen as the topic with the *highest probability for a given word in the document* to include the context of this document:[9]

$$k^{**} = \arg\max_k \beta_{kw}\theta_{dk} \tag{4.3}$$

[9] The general derivation of this formula could not be found in the literature. It seems, however, that it can be derived by Eqs. [5], [6] and [7] in Griffiths and Steyvers (2004) within the framework of Gibbs sampling.

Based on this criterion, the words in the newsgroup message have the following topics. Their frequencies of occurrence are mostly consistent with the (rounded) topic proportions of this message, which are extracted of $\boldsymbol{\theta}$ and also included hereafter:[10]

```
Topic 1: 1%    Topic 2: 72%    Topic 3: 1%    Topic 4: 27%

With the continuin talk about the "End of the Space Age" and complaints
by government over the large cost, why not try something I read about
that might just work.
Announce that a reward of $1 billion would go to the first corporation
who successfully keeps at least 1 person alive on the moon for a year.
Then you'd see some of the inexpensive but not popular technologies begin
to be developed.  THere'd be a different kind of space race then!
```

Additional examples based on the news about investment styles are included in Secs. C.4.1 and C.4.2 of the appendix. These examples were created by writing HTML documents via our Python function `write_article_html()` in appendix B.

4.4.2 Topic coverage in each magazine

It was mentioned in the previous section that some magazines seem to have specific dominant topics. To test this statement, the *topic coverage in each magazine* could be estimated by counting the number of articles about each topic in a given magazine. LDA offers, however, the possibility to detect multiple topics in each document. So, instead of assuming that the topic of an article is the dominant topic in this article or the topic with a proportion greater than a certain threshold, e.g. 40%, topic proportions θ_{dk} are summed. In this way, no arbitrary threshold has to be introduced and all available information is fully taken into account. Thus, the *absolute importance* $\xi^a_{k,m}$ of topic k in *magazine* m is computed as follows, where $\mathcal{D}_m$ is the corpus of articles published in magazine m:

$$\xi^a_{k,m} = \sum_{d \in \mathcal{D}_m} \theta_{dk} \tag{4.4}$$

To compare the topic coverage in a magazine to the coverage in another magazine, it is necessary to normalize the previous values of absolute importance. Otherwise, the coverage of one topic might simply be more important in a certain magazine because it contains in general more articles

[10] Although no word seems to have been drawn from topics 1 and 3, their percentages are not zero. This can be explained by Dirichlet smoothing (Blei et al., 2003), see e.g. the influence of the hyperparameters in Eqs. [6] and [7] in Griffiths and Steyvers (2004).

than the other magazine. Hence, the *topic proportions of a magazine* are defined by the *relative importance* $\xi^r_{k,m}$ of topic k in *magazine* m:

$$\xi^r_{k,m} = \frac{\xi^a_{k,m}}{\sum_{k^*=1}^{K} \xi^a_{k^*,m}} \tag{4.5}$$

4.4.3 Importance of topics over time

Some topics might preferentially occur at some moments in time, as previously illustrated in Sec. 2.3.4. To detect these trends, the *absolute importance* $\xi^a_{k,t}$ of topic k during the period t can be determined in a similar way than in the previous section, where $\mathcal{D}_t$ is the corpus of documents published during *period* t:

$$\xi^a_{k,t} = \sum_{d \in \mathcal{D}_t} \theta_{dk} \tag{4.6}$$

Furthermore, it seems reasonable that an investor is not only influenced by the absolute importance of a topic but also by its relative importance with respect to all topics. For instance, if a magazine writes exclusively about silver and gold in equal proportions, both metals can be assumed to grab equal amounts of attention. This seems still reasonable, if the number of articles about silver increases by the same number as those about gold. If, however, suddenly, the journal writes 90% of its articles about gold, it is likely that attention shifts towards gold. Hence, it is useful to compute the *relative importance* $\xi^r_{k,t}$ of topic k during the *period* t:

$$\xi^r_{k,t} = \frac{\xi^a_{k,t}}{\sum_{k^*=1}^{K} \xi^a_{k^*,t}} \tag{4.7}$$

Chapter 5

Results and discussion

This chapter discusses the *results of the topic model obtained by processing the investment style news* (Chap. 3) by latent Dirichlet allocation as described in Chap. 4. First, *topics are identified* for various parameters of the model. Secondly, the *topic coverage by each magazine* is examined. And finally, the *importance of these topics over time* is analyzed.

5.1 Topic identification

To identify the topics, the number of topics is first estimated by the *perplexity*. Then, the *influence of user-specified parameters* is analyzed. And, finally, the *topics are labeled*.

5.1.1 Determination of the number of topics by the perplexity

Labeling topics requires specifying the *number of topics* K to determine the topics. To estimate this number, Fig. 5.1 represents the *perplexity* as a function of K for different values of the hyperparameters. The first set of values is the default set in scikit-learn, while the second one is the most used consistent combination in the literature according to Sec. 2.3.3.[1] Moreover, the number of iterations of the expectation-maximization algorithm in LDA was increased from 10 (default) to 20 in both cases to reach a more converged state. More precisely, Fig. 5.2 illustrates the perplexity as a function of this number of iterations for a constant number of topics. It can

[1]If $\alpha = 50/K$ was used instead, α would be greater than 1 in our case, which would favor having a lot of topics in each document. It seems, however, more reasonable that a document is only about a few topics. Moreover, Kaplan and Vakili (2015) and Huang et al. (2018) justify using $\eta = 0.01$ by the desire to find topics with few high-probability words.

be seen that this perplexity still decreases relatively significantly at 10 iterations, especially for the red curve ($\alpha = 0.1$, $\eta = 0.01$ and $K = 10$), while the variation becomes quite small after 20 iterations.

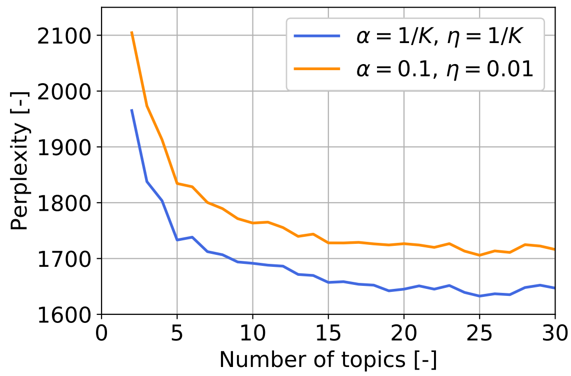


Figure 5.1: Perplexity as a function of the number of topics K (20 iterations).

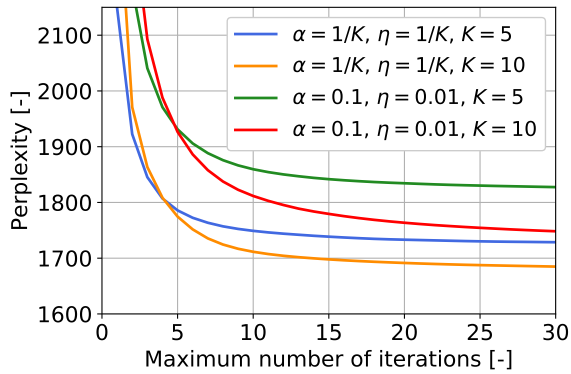


Figure 5.2: Perplexity as a function of the maximum number of iterations (of the expectation-maximization algorithm).

Based on Fig. 5.1, choosing *up to 15 topics* seems to be a reasonable decision since the remaining decrease of the perplexity after $K = 15$ is relatively small. Topics could obviously be extracted for $K > 15$, but their interpretation becomes increasingly difficult. In fact, it is shown in Sec. 5.1.3 that spurious topics start to appear even for $K = 15$.

5.1.2 Influence of parameters

To understand the influence of the different parameters on the results, the configurations in Tab. 5.1 are tested. The corresponding word lists with the titles of the most relevant articles can be found in appendix D.[2] The word lists are constructed based on the relevance measure in Eq. (4.1) with $\lambda = 0.6$, as suggested in Sec. 4.4.1.

Configurations 1 and 2 are tested to illustrate that the method is by default *non-deterministic*, i.e. that the same input parameters (excluding the random seed) do not necessarily lead to the exact same results. In fact, the latent variables are initialized by drawing pseudorandom samples. A pseudorandom number generator depends on a random seed for its initialization. In configuration 2, this seed is changed to check whether the initialization has a significant impact on the results. Figs. D.1 and D.2 in the appendix show that the topics are on average still the

[2] The maximum number of iterations of the expectation-maximization algorithm is kept at 20 according to the explanations in the previous section.

Configuration	K	α	η	Seed
1	5	$1/K$	$1/K$	0
2	5	$1/K$	$1/K$	1
3	5	0.1	0.01	0
4	10	$1/K$	$1/K$	0
5	10	0.1	0.01	0
6	15	$1/K$	$1/K$	0

Table 5.1: Configurations of parameters.

same. They are shuffled and partially modified, though. For instance, topic 3 in configuration 1 is essentially topic 1 in configuration 2:

```
Topic 3 (1): plan pension retirement fund percent investment fee participant
Topic 1 (2): plan retirement pension fund participant sponsor fee investment
```

Topic 1 in configuration 1 is, however, not the direct counterpart of topic 2 in configuration 2, which should be its counterpart after assigning the remaining topics in configuration 1 to their most closely related topics in configuration 2:

```
Topic 1 (1): market year equity sector say company investor growth return stock
Topic 2 (2): team management join esg investment manager equity analyst appoint
```

Non-determinism is inherent to LDA since it is based on the optimization of a complex function so that the convergence to the global optimum cannot be guaranteed. In consequence, different local optima are reached depending on the initial configuration. To ensure the reproducibility of our results, the value of the random seed is kept at 0 hereafter.

Besides the non-determinism, the *influence of the hyperparameters* can be analyzed by comparing the word lists of configuration 1 (Fig. D.1) to those of configuration 3 (Fig. D.3) for $K = 5$, as well as the word lists of configuration 4 (Fig. D.4) to those of configuration 5 (Fig. D.6) for $K = 10$. These word lists are essentially the same except for some small differences in word order, e.g.

```
Topic 1 (1): market year equity sector say company investor growth return stock
Topic 1 (3): market year sector equity say growth investor company european rate
```

According the previous results, choosing the default values of the hyperparameters $\alpha = 1/K$ and $\eta = 1/K$ seems reasonable. Hence, the topics of the configurations 1, 4 and 6 are labeled in the following section to study the influence of the number of topics, too.

5.1.3 Topic labeling

In this section, the topics in investment style news are labeled for either 5, 10 or 15 topics. It should be noted beforehand that labeling topics is *not simple*, especially not in such a specific context as investment style news. Thus, we do not claim that the following labels are the most appropriate ones, but the best that we could find. In the following lines, our approach and findings are described from the smallest to the largest number of topics. The results are summarized in Tabs. 5.2, 5.3 and 5.4, which contain the topic labels and the top 20 words of each topic according to the relevance measure in Eq. (4.1) with $\lambda = 0.6$. More detailed words lists including the titles of the top articles can be found in Secs. D.1, D.4 and D.6 in the appendix.

Five topics

Tab. 5.2 contains the results for *5 topics*. The choices of the topic labels are motivated in the following paragraphs by trying to use most of the top words in these paragraphs and by referring to the top articles; sentences might appear simplistic because of these forced word choices.

Topic label	Top 20 words
Equity market (economy)	market year equity sector say company investor growth return stock high european rate economy manager rise china price yield economic
Analyst research, trading and banking	bank firm market business trading say trade client research capital deal company banking year ipo new private finance million analyst
Retirement planning	plan pension retirement fund percent investment fee participant say sponsor asset make share newsdash hedge fiduciary endowment use active money
Indexes, ETFs and performance	index etf cap market vanguard russell fund billion msci equity large stock inflow return factor small performance emerge quarter month
Fund management and fund launches	fund investment management manager portfolio strategy cap asset equity manage small team company value investor launch global invest growth client

Table 5.2: Inferred topic labels and top 20 words for 5 topics.

Topic 1 mainly focuses on the *equity market*, which is closely tied to the *economy*.[3] In fact, the equity market enables investors to buy stock of companies in different sectors at certain prices to obtain a return. Growth implies high stock returns, which are usually associated with certain years and geographic regions (Europe, China). Top articles of this topic are about the impact

[3]Since the separation between "equity market" and "economy" is not distinct according to the word list, "economy" is written in parentheses next to "equity market" in Tab. 5.2 and hereafter.

of European political uncertainty on the equity market, Japanese equities or indicators of equity returns.

Topic 2 is not easily summarized by a single label. It seems to focus on *analyst research* (firm, client, research, company, finance, analyst, ...), *trading* (market, trading, trade, deal, ipo, ...) and *banking* (bank, banking, capital, ...). The 1st, 2nd, 4th and 5th top articles are about rankings of equity sell-side analyst teams (e.g. All-America Research Team), who provide research to clients about companies and who are usually employed by banks. The 3rd, 6th and 7th top articles concern high-frequency trading and the influence of tick size, i.e. the minimum movement of a security price, on trading. Finally, the 8th top article focuses on banking services required by SMEs.

Topic 3 is clearly about *retirement planning* since a pension plan is a special form of retirement plan, since these plans usually consist in investments of money from plan participants in funds that come with a fee, since these plans are set up by sponsors and since fiduciaries manage them. Top titles of topic 3 mention lifetime income plans, DC (defined contribution) plans, Larry Fink's opinion on retirement planning, plan fees, multiemployer plans, ...

Topic 4 is about *indexes, ETFs and performance*. It appears reasonable to find these concepts in the same topic since performance is usually compared to or measured by indexes, and since indexes can be traded via ETFs. Hence, it is no surprise to find index providers, like Russell and MSCI, in the top words as well as Vanguard as one of the largest ETF issuers. To include some additional top words in this paragraph, one can say that indexes usually aggregate price information about stock equity traded on a market, and their performance is measured by monthly or quarterly returns. Top articles are, for instance, about the largest stock and bond funds, which are index funds of Vanguard available as ETFs, and the difference in performance of actively and passively managed funds. Some titles of top articles, however, also mention DC plans, like 401(k)s, which one would rather expect in topic 3. Finding them in topic 4 can, however, be explained by the strong focus on indexes and performance in these articles. For instance, in the top article "DC Participants Less Active Traders in March", the words "plan", "pension", "retirement" are not mentioned at all, but "index" and "performance" numerous times.

Topic 5 is mainly about *fund management and fund launches*. Hence, by using the top words, one can say that the focus lies on the investment strategy of the fund via its portfolio of assets including equity, and its team that manages and grows capital for clients. Concerning the top 8 articles, 5 are about the modification of fund management/strategy, while 3 are about new fund launches. It should be noted that this topic seems to mainly focus on actively managed funds since passively managed funds, like index funds and ETFs, as well as articles about their launches can be found in topic 4 when the top 20 articles are examined.

As a side note, one should take a look at the topic proportions of the article in Fig. C.5 in the appendix, which was previously presented to explain probabilistic topic models (Fig. 2.4). This article is about the replacement of a fund manager, which one would most likely classify into the previous topic “fund management and fund launches”. Without surprise, the corresponding topic proportion is actually 60%. The remaining 40% are mainly from the “equity market (economy)” topic, probably because the word “European” appears 5 times in this short article and because this word is strongly associated in other articles with the topic “equity market (economy)”.

In conclusion, despite the complexity of the data (in comparison to the newsgroup messages), LDA is *able to detect related concepts*, which can, however, not always be summarized by a single overarching label. In fact, some topics seem to be combinations of topics, which one would separate, like “analyst research, trading and banking” in the second topic. Hence, one may wonder whether these topics are actually separated, if LDA is applied with 10 or 15 topics.

More than five topics

The 10 and 15 topics in Tabs. 5.3 and 5.4, respectively, were also labeled by analyzing the top words lists, top titles and corresponding articles. Instead of explaining again the reasoning behind each label, which might be repetitive, we will rather focus on more general findings.

First, *labeling becomes much more complex for some topics*, while it becomes *much easier for others* after increasing the number of topics: on the one hand, no real label could be found for the last topic about “Asia, private equity and the search for asset managers via IPE Quest” in Tab. 5.3. On the other hand, the topic “analyst research” in Tab. 5.4 groups articles about the “All-America Research Team”, a ranking of American research analysts in the Institutional Investor magazine, so that labeling is relatively easy. Likewise, the top 8 articles of the last topic in Tab. 5.4 are all about fund flows, while the top words include “inflow, outflow, flow”, thus leaving no room for ambiguity.

Secondly, *some topics are kept almost unchanged* when the number of topics is increased. This can certainly at least partially be traced back to initializing the latent variables in the same way for different numbers of topics since the random seed is unchanged. Hence, the algorithm always reaches certain local optima. For instance, the first topic is the “equity market (economy)” for all values of K. Similarly, the topic about “fund management and fund launches” is present for all

Topic label	Top 20 words
Equity market (economy)	market year sector growth equity company high stock european rate investor rise return economy say cap yield price manager small
Trading	trading market ipo exchange trade trader russia russian company volume say bat order firm listing broker stock moscow new commission
Pension	percent pension newsdash retirement employee employer activist plan say new worker endowment state board school health university cio benefit public
Indexes and ETFs	index etf cap msci russell market weight billion ishares inflow emerge small large stock spdr factor exposure etfs volatility global
Fund management and fund launches	fund investment management portfolio team cap manager small manage strategy equity asset company launch growth value global capital join invest
Retirement plan products	plan vanguard fund fee share expense sponsor retirement participant class option hancock investment fiduciary cost plaintiff john offer complaint ratio
Banking, analyst research	bank banking loan finance business lending bnp paribas year credit smes corporates lender capital client euromoney analyst billion runner trade
Performance	quarter return fund target plan asset equity date bond year fixed allocation income average participant tdfs increase maturity performance flow
Investment strategy	manager investor think active use risk say investment portfolio make research time way firm lot strategy different want like need
Asia, private equity, searches for asset managers via IPE Quest	china private fund billion equity hedge hong kong chinese asset market million say percent capital emerge firm year management asia

Table 5.3: Inferred topic labels and top 20 words for 10 topics.

numbers of topics.

Thirdly, *topics are actually separated into more specific topics* when the number of topics is increased, as previously anticipated. For example, the topic "analyst research, trading and banking" for $K = 5$ is divided into "trading" and "banking, analyst research" for $K = 10$. Later, for $K = 15$, "trading" disappears and "banking, analyst research" becomes "European banking", "corporate banking" and "analyst research". Likewise, a separation between "indexes and ETFs" occurs from $K = 10$ to 15. In fact, the top 8 articles of the topic "ETF launches" for $K = 15$ are exclusively about ETF launches, while those of the topic "Indexes and ethical investing" are mainly about index launches.

Fourthly, at $K = 15$, *special kinds of topics* emerge. The first special topic is the "Vanguard and John Hancock" topic, which mainly focuses on these names. For instance, the top 8 articles of

Topic label	Top 20 words
Equity market (economy)	market sector year growth equity company high european investor rise stock rate economy say price cap yield dividend earnings manager
ETF launches	etf market trading index exchange cap trade weight factor stock small volatility emerge beta exposure bat nasdaq nyse launch volume
Pension	percent pension newsdash retirement activist worker employer employee new board state say health kemna cio endowment school wisconsin public walker
Indexes and ethical investing	index esg russell msci sri wilshire dow jones environmental aon cap hewitt sustainable sustainability measure social acwi market company frontier
Fund management and fund launches	fund investment management manager portfolio cap small strategy team equity manage asset company growth value launch global invest capital investor
Vanguard and John Hancock	vanguard hancock john expense etf tax ratio index firearm fund explorer international ast timessquare dividend basis ing mcnabb transamerica municipal
European banking	bank loan finance italy germany lender banking european german lending spain italian helaba hungary trade eurobank debt greek billion smes
Performance	return quarter equity asset allocation year bond fixed fund target average plan income performance gain real maturity rate end median
Investment strategy	investor manager portfolio think research active strategy firm say stock hedge use make investment return time like factor risk beta
Emerging markets, IPO, private equity	china private market deal ipo billion hong kong say capital russian chinese million raise firm company percent russia fund gso year
Corporate banking	bank client business need euromoney say bnp paribas want corporates customer liquidity technology make cash people banking work lot just way
Analyst research	analyst firm research evercore morgan team america year runner merrill join liquidnet university merrin client lynch work independent goldman senior
Retirement plan products	plan fund fee participant sponsor retirement investment active share option class target date fiduciary passive asset fidelity adviser use nextpage
Articles with an exclusive frequent word	alphadex franklin bullishness bissett goalmaker templeton ifunds rollins ave maria ifas dashboard schneider family ibillionaire redwood albion fma polley factsheet
Fund flows	billion inflow etf outflow month million flow ishares net category cap gold saw respectively market spdr asset commodity bond large

Table 5.4: Inferred topic labels and top 20 words for 15 topics.

this topic all contain the word “Vanguard”. The second special topic is “Articles with an exclusive frequent word”. More precisely, these articles are characterized by a word that is only frequent in them and in very few other articles. For instance, the top word “AlphaDEX” occurs 44 times in the corpus but only in 3 articles. The top articles of this topic have merely a corresponding topic proportion of about 25%, while the respective topic proportions of top articles for other topics are generally between 90 and 100% (see Sec. D.6 in the appendix). Hence, determining at most around *15 topics as suggested by the perplexity* in Sec. 5.1.1 is a *good choice*.

To conclude, the topic identification by LDA is very *powerful* in our opinion since topics were determined without having to read a significant portion of the corpus. The approach is, however, *limited by the expert knowledge that is required to label topics*. The next step is to use the previous topics to extract further information of the corpus.

5.2 Topic coverage in each magazine

The *topic coverage in each magazine* is analyzed to characterize the magazines and to validate the previous findings.[4] Tab. 5.5 contains the topic proportions of each magazine for 5 topics, i.e. the values of relative importance according to Eq. (4.5). The topic proportions for 10 and 15 topics are available in Tabs. D.1 and D.2 in the appendix. Based on all these data, the magazines likely have the following *dominant topics*:[5]

- *Euromoney* focuses on “corporate” and “European banking”, “trading” and the “equity market (economy)”;
- *Institutional Investor* is the magazine with the *most homogeneous* coverage of all topics. When the number of topics is increased, “investment strategy” becomes the major topic;
- *Financial Adviser* covers topics *most heterogeneously*, so that it focuses almost exclusively on the “equity market (economy)”, and secondarily, on “fund management and fund launches”;
- *AlphaQ* has similar but less concentrated topic proportions than the Financial Adviser;

[4]The characterization of magazines should not be over-interpreted due to the preliminary filtering of articles as explained in Chap. 3. In other words, this characterization is based on the articles related to the small-cap investment style but not on all articles of the magazines.

[5]The approach to determine the dominant topics is partially subjective. We started by considering the most significant topic proportions of each magazine in Tab. 5.5, e.g. “analyst research, trading and banking” for Euromoney. And then, we tried to refine these topics based on the most significant proportions in Tabs. D.1 and D.2, e.g. “trading”, “European banking” and “corporate banking” again for Euromoney.

- *Institutional Asset Manager* reports mainly on “fund management and fund launches”, and secondarily, on the “equity market (economy)”, “trading” as well as “indexes and ETFs”;
- *Wealth Adviser* covers mostly “fund management and fund launches” as well as “ETF launches” and the “equity market (economy)”;
- *Investment & Pension Europe* provides information about “retirement planning”, “investment strategy”, and “searches for asset managers”;
- *PlanSponsor* and *PlanAdvisor* focus on “indexes, ETFs, performance” including “fund flows”, and “retirement planning”.

Magazine	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Euromoney	32%	59%	4%	2%	3%
Institutional Investor	25%	26%	23%	12%	13%
Financial Adviser	65%	4%	3%	4%	24%
AlphaQ	34%	4%	14%	9%	39%
Institutional Asset Manager	14%	25%	2%	17%	41%
Wealth Adviser	16%	6%	2%	18%	59%
Investment & Pension Europe	21%	9%	27%	7%	36%
PlanSponsor	5%	3%	21%	49%	21%
PlanAdviser	10%	3%	17%	48%	23%

(a) Topic coverage in each magazine.

Topic 1	Equity market (economy)
Topic 2	Analyst research, trading and banking
Topic 3	Retirement planning
Topic 4	Indexes, ETFs and performance
Topic 5	Fund management and fund launches

(b) Topic labels.

Table 5.5: Topic proportions of each magazine.

Concerning the validation of the results, the previous *dominant topics actually correspond to those suggested by the expert of the data set* in Sec. 4.4.1. In fact, the Institutional Investor preferentially reports on “strategy” by the topic “investment strategy”, which emerges for $K = 10$ and $K = 15$. Moreover, the Wealth Advisor mainly focuses on “new funds” via the topics “fund management and fund launches” and “ETF launches”. And finally, PlanSponsor covers the “past performance” by the topic “indexes, ETFs and performance”, which can be refined to “performance” and “fund flows” by increasing the number of topics. These findings validate the method at least partially.

Furthermore, the results of the topic model seem consistent for the following reasons. First, PlanSponsor and PlanAdviser frequently contain the same articles. Hence, their topic proportions should be similar, which is actually true. Secondly, the topic coverage of each magazine should agree with its short description in Tab. 3.1. This is also true since Euromoney covers “European banking”, and Investment & Pension Europe, PlanSponsor and PlanAdviser significantly report on “retirement planning”, for instance.

5.3 Importance of topics over time

Before presenting the results about the importance of topics over time, let us remind that the ultimate objective of this research is to determine whether the coverage of style investing in news affects fund flows of corresponding smart beta ETFs. A regression model of fund flows versus temporally lagged style coverage should shed light on this relation. In simple terms, the question is whether the publication of more articles about a certain investment style at time t (e.g. January) influences fund flows of smart beta ETFs of this style at time $t+1$ (e.g. February). Topic modeling is introduced in this research to increase the granularity of the available information in the regression model. Instead of studying the relation between fund flows and news related to small-cap investing as a whole, these news can be divided into topics to study their individual relations to fund flows. In this thesis, it was decided to focus exclusively on the topic modeling of investment style news and not on the regression model. Nevertheless, the independent variables of this regression model can be presented in this document. These variables are the absolute or relative importance of topics, as defined in Sec. 4.4.3, for given periods.

Any topic of those in the previous Sec. 5.1 could be selected to compute its temporal importance. Without any prior knowledge about which topics impact fund flows most significantly, a reasonable first step is to choose the topics determined by the minimum *number of topics*, i.e. $K = 5$. Moreover, two different *kinds of periods* are considered: on the one hand, the articles are grouped for each year to detect *historical trends*.[6] On the other hand, articles are grouped per month (independently of the year) to identify *trends that occur on average in each year*. Considering the complexity of the data and its filtering, it is obviously difficult to relate topical trends to external drivers.[7] Hence, the following analysis is mainly descriptive.

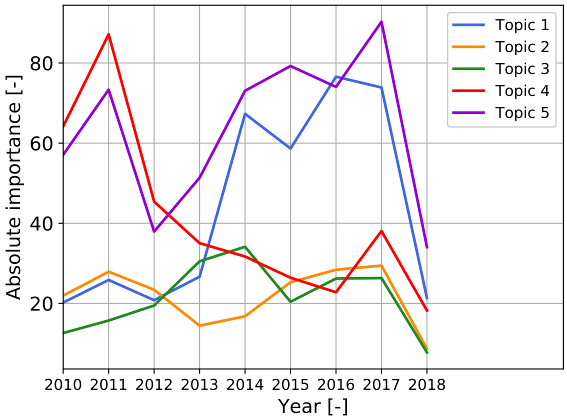
Figs. 5.3a and 5.3b illustrate the absolute and relative importance of the 5 topics from January 2010 until half of July 2018 for each year. The yearly *absolute importance* of each topic is mainly represented for two reasons. First, it provides an idea about the topic frequencies per year. So, at least around 10 and at most around 90 “articles” exist about a certain topic per year.[8] Secondly,

[6]Smaller intervals than years could obviously be chosen but they would be difficult to read in a plot. For instance, if the importance of topics was plotted for each single month from January 2010 until half of July 2018, there would be 103 single months with relatively strong coverage variations from month to month.

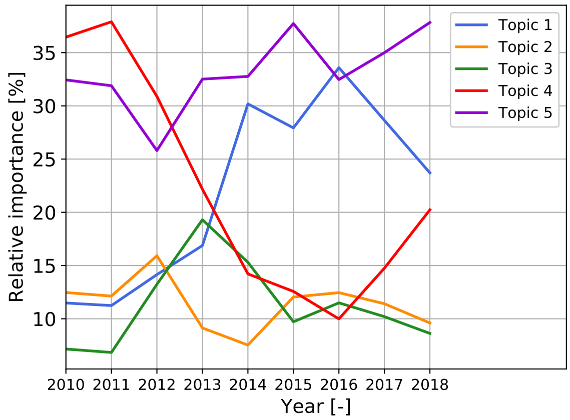
[7]The initial articles of news targeting institutional investors were filtered for news related to small-cap investing. Hence, significant expert knowledge would be required to understand why a topic like the equity market suddenly becomes more important at some moment in time. In other words, the problem is more complex than relating the increase of articles about macroeconomic risk in a business newspaper around 2008 to the corresponding financial crisis.

[8]The quotes around “articles” are added since topic proportions (instead of articles) are summed to compute the absolute importance of a topic according to Eq. (4.4).

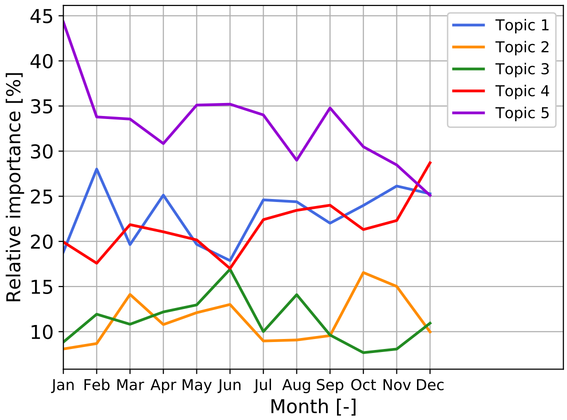
it shows that the relative importance might be a better metric to measure the importance of a topic than the absolute importance due to the fluctuations of the total annual number of articles. For instance, fewer articles are available in 2012, and especially, in 2018 since news were only collected up to about half of that year.



(a) Absolute importance per year.



(b) Relative importance per year.



(c) Relative (average) importance per month.

Topic 1	Equity market (economy)
Topic 2	Analyst research, trading and banking
Topic 3	Retirement planning
Topic 4	Indexes, ETFs and performance
Topic 5	Fund management and fund launches

(d) Topic labels.

Figure 5.3: Importance of 5 topics per year or per month.

According to the *relative importance* in Fig. 5.3b, the *most important topic* during the entire period is “fund management and fund launches”. At the beginning of the period, “indexes, ETFs and performance” has about the same importance but it *decreases subsequently while being compensated by* “equity market (economy)”. These major trends can also be observed for 10 and 15 topics in Figs. D.5b and D.8b in the appendix. This comparison is possible since the topics “equity market (economy)” and “fund management and fund launches” reappear for the different values of K. The topic “indexes, ETFs and performance” specializes, however, to “ETF launches” for $K = 15$, which has the same historical trend. Thus, the decreasing importance of this topic in Fig. D.8b could possibly be interpreted as a *decrease of small-cap smart beta ETF launches*. This decrease is surprising in consideration of the rising number smart beta ETFs/ETPs in Fig. 1.1 in the introduction. However, one could imagine that small-cap smart beta ETFs were mainly launched early on in the history of smart beta ETFs as a first go-to solution since the small-size factor is one of the major factors. Later on, small-cap smart beta ETFs are complemented by other kinds of smart beta ETFs, so that launches of the small-cap style become less frequent. This explanation is obviously hypothetical and it could be verified in future research by analyzing databases of smart beta ETF launches. Coming back to $K = 5$ in Fig. 5.3b, we observe that the topics “retirement planning” and “analyst research, trading and banking” have about the same importance since 2015. As mentioned previously, it is difficult to further analyze these findings in a meaningful way.

Concerning the average relative importance of a topic over a year, i.e. the *seasonality of topics*, the most significant trend in Fig. 5.3c is certainly the increase of the topic “fund management and fund launches” in January and its decrease at the end of the year. This trend can also be observed in Figs. D.5c and D.8c for $K = 10$ and 15. It suggests that changes of fund management and fund launches preferentially occur at the beginning of the calendar year, which seems reasonable.[9] Meanwhile, the importance of the topic “indexes, ETFs and performance” increases towards the end of the year. An intuitive and hypothetical interpretation might be that management changes and fund launches are the response to past performances. Due to synchronicity, the previous observations could be linked to the *turn-of-the-year effect*, also known as the January effect, but the relation is unclear since the origin of this effect is not yet precisely known. This effect consists in the concentration of the small-size effect, i.e. abnormally high returns of small-capitalization stocks, at the beginning of the year (Lynch et al., 2014; Sikes, 2014).

[9]One should notice that fund launches in this topic are rather launches of actively managed funds than ETFs and index funds, which mostly occur in the topic “indexes, ETFs and performance”.

Chapter 6

Conclusion

In this thesis, a machine learning method called latent Dirichlet allocation (Chap. 2) was adopted to *identify the major topics* in a unique corpus of magazine articles related to small-cap investing (Chap. 3). Moreover, the *topic proportions* in each article were determined so that the importance of topics measured by their frequency of occurrence could be quantified for each magazine and during time periods (Chap. 4). Thereby, the *magazines* and the *topical trends over time* were characterized and analyzed (Chap. 5). Ultimately, these results allow to check whether the coverage of specific topics in news related to the small-cap style could influence fund flows of smart beta ETFs focusing on this style.

In the following section, these general conclusions are described more explicitly by summarizing the *outcomes* of this research. Finally, *future research perspectives* are provided in the last section.

6.1 Summary and main contributions

The most significant outcomes including our main contributions are summarized by following the structure of this document.

Chapter 2 - Contextualization and literature review

Basic concepts of investment theory were introduced to better understand the research topic in its entirety. Based on Markowitz portfolio theory and the capital asset pricing model (CAPM), it was shown that a cap-weighted market index maximizes the expected return for a given level of risk. This return is a function of the exposure to systematic risk factors, while unsystematic risk is not rewarded. The arbitrage pricing theory allows to generalize the risk premium due to market exposure of the CAPM to additional risk factors. In particular, a small market capitalization and high book-to-market ratios of stocks can be considered to proxy for risk factors since risk premiums increase with these features according to the Fama-French three-factor model. Smart beta exchange-traded funds (ETFs) are a cost-effective and transparent way to gain exposure to factors and hence, achieve higher returns than classical cap-weighted index funds. In addition, smart beta ETFs offer alternative weighting schemes that could reduce unsystematic risk exposure of these index funds. Due to the resulting popularity of smart beta ETFs, the objective of this thesis is to identify topics and their frequency in magazine articles related to the small-cap investment style. Thus, it can be tested in future research whether the coverage of certain topics influences fund flows of smart beta ETFs associated with this style.

Topics are most efficiently extracted from large collections of documents by a machine learning approach that is called *topic modeling*. It is built on a bag-of-words representation of these documents via a document-term matrix that contains the counts of each term for each document. A review of topic models showed that results obtained by factorizing this matrix, i.e. by latent semantic analysis (LSA) or non-negative matrix factorization (NMF), are more difficult to interpret than those of probabilistic topic models. These latter models assume that topics are distributions over words and that documents are distributions over topics. The corresponding distributions are computed by statistical inference based on a probabilistic generative model of textual data and the observed corpus itself. Latent Dirichlet allocation (LDA), which is the most popular probabilistic topic model, was selected for the data analysis in this thesis instead of probabilistic latent semantic analysis (pLSA) since LDA includes Dirichlet priors on the previous distributions. Among other things, this encodes the intuition that documents are about a few topics and that topics are defined by a few high-probability words.

Finally, *LDA in finance* was extensively reviewed for the first time (to the best of our knowledge) in order to determine how to optimally apply this method in this context. In particular, data pre-processing steps, topic model implementations and their parameters, and post-processing methods were surveyed.

Chapter 3 - Data

The textual *data* in this thesis consist of 1720 articles from 2010 to July 2018, which include the bigram “small cap” or a similar variation, so that they are related to small-cap investing. These articles were selected from a unique and much larger corpus of articles that was created by Gillain et al. (2019). They collected these articles from 9 magazines of 5 media groups whose mission statement includes the production of information for financial decision makers. In consequence, this corpus is expected to contain information about style investing and to provide this information to institutional investors, who are most likely to influence fund flows.

Chapter 4 - Methodology

The previous data were *pre-processed* by the removal of irrelevant information (like web addresses), tokenization, case folding, lemmatization based on part-of-speech-tagging and the removal of non-alphabetic characters, very short words and stop words. Then, per-topic word distributions and per-document topic distributions were computed by *LDA with symmetric Dirichlet priors* for different choices of parameters. In particular, the user-specified number of topics was estimated by the *perplexity*, which measures the ability of a trained LDA model to predict the testing data (of the same corpus), and topic coherence. The resulting topics were *labeled* by *manually* examining the respective top words, titles and articles of these topics. In particular, the top words were chosen by a *relevance measure* that penalizes very frequent words. Finally, the topic coverage in each magazine and the importance of topics over time was computed based on the topic proportions in each article.

The previous *data processing* was mainly carried out by the Python modules NLTK and scikit-learn as well as our own topic modeling module, which offers the option of outputting HTML documents of the results to simplify their evaluation. Moreover, the previous methodology was satisfactorily *validated* by the 20 newsgroups corpus, for which the underlying topics are known.

Chapter 5 - Results and discussion

The *number of topics* was estimated at about 15 by the perplexity. In addition, 5 and 10 topics were also separately extracted to better understand the influence of this parameter. The previous estimate proved to be valid since spurious topics started to appear for 15 topics. Moreover, the *hyperparameters* of the Dirichlet priors had no significant influence on the results within the tested range, whereas the *random seed* that initializes LDA, however, slightly modified the topics.

When 5 topics were identified, these topics were *labeled* as “equity market (economy)”, “analyst research, trading and banking”, “retirement planning”, “indexes, ETFs and performance” and “fund management and fund launches”. For more than 5 topics, some of them persisted, some disappeared and others were specialized. For instance, “analyst research, trading and banking” was separated into “analyst research”, “European banking” and “corporate banking”, while "trading" disappeared for 15 topics.

The *topic coverage in each magazine* allowed to validate the LDA results since dominant topics in specific magazines suggested by the expert of the corpus were actually dominant topics of these magazines according to LDA. Moreover, the topic coverage was consistent with the short descriptions of the magazines.

Concerning the *evolution of topics over time*, "fund management and fund launches" was the most frequent topic over the analyzed period, whereas "indexes, ETFs and performance" decreased in contrast to "equity market (economy)". Considering the complexity of the data, we could only hypothesize that this decrease might be due to a decreasing number of small-cap smart beta ETF launches. Moreover, the analysis of the topic coverage over a year, i.e. the seasonality of topics, showed that the topic "fund management and fund launches" preferentially occurs in January and less frequently at the end of the year, thus suggesting that changes of fund management and fund launches especially take place at the beginning of the calendar year.

In conclusion, LDA turned out to be a *very powerful method* since major topics could be identified and since articles could be clustered without having to read a significant portion of the 1720 articles. The approach is, however, limited by the expert knowledge required to label topics and to interpret historical or seasonal trends.

6.2 Future research perspectives

This thesis was written within the context of Gillain's PhD thesis (Gillain, 2020) whose research question is to determine whether the media coverage of investment styles influences fund flows of corresponding smart beta ETFs via investors who read these media. Future research should therefore focus on *combining the previous results with quantitative data of smart beta ETFs*. Thus, the previous hypothesis about the decreasing number of small-cap smart beta ETF launches could be verified.

More importantly, however, the *regression model* in the introduction (Chap. 1) that relates fund flows of small-cap smart beta ETFs at time t to the media coverage associated with small-cap investing at $t-1$, i.e.

$$\text{flow}_t = \beta \text{ coverage}_{t-1} + \dots \tag{6.1}$$

could be *refined thanks to the increased informational granularity* provided by this thesis. More precisely, the coverage could be separated according to the different topics:

$$\text{flow}_t = \beta_1 \text{ (coverage topic 1)}_{t-1} + \beta_2 \text{ (coverage topic 2)}_{t-1} + \dots \tag{6.2}$$

In this way, the individual influences of these topics on fund flows could be estimated by the resulting values of $\beta_1, \beta_2, \ldots$ to better understand how investment style news potentially influence fund flows.

Appendix A

Example of the initial textual data

The initial textual data has the following structure for each article.

Listing A.1: Example of an article in the data set

```
Date :
2014-06-30 00:00:00
Title :
 Fidelity replaces manager of European Opps fund
Magazine :
FTAdviser
ID :
63617




#paragraph# Fidelity has moved to replace Colin Stone with Alberto Chiandetti on the underperforming £432m Fidelity European Opportunities fund.

#paragraph# Mr Stone had been managing the fund since 2003 but it had slipped into the bottom quartile of the IMA European sector for three and five years, according to data from FE Analytics.

#paragraph# He will continue to manage the Fidelity’s European small cap strategy, including the offshore FF European Smaller Companies fund.

#paragraph# Fidelity said Mr Chiandetti would run the European Opportunities fund alongside Mr Stone until October before taking sole responsibility.

#paragraph# Mr Chiandetti will remain as manager of the Luxembourg-domiciled FF Italy and FF Switzerland funds, though Fidelity said he had been "allocated dedicated resources to support these country funds".
```

Appendix B

PYTHON scripts

This appendix contains the standard PYTHON analysis script of the investment style news and our topic modeling module that includes most of the pre- and post-processing functions.

B.1 Standard analysis script of news

Listing B.1: Standard analysis script of news - style_news.py

```
# -*- coding: utf-8 -*-

import pyLDAvis
import pyLDAvis.sklearn

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

from topic_modeling import load_news, prepro_corpus, write_prepro_html, \
                           write_topics_terminal, write_article_html, \
                           words_df_max, write_title_html, \
                           write_magazine_html, plot_topic_month, \
                           plot_topic_year, write_articles_topics_xlsx

news = load_news("input/Small_news.txt", nbr_max=-1)

# Create corpus that includes titles
corpus = [article.title + "\n\n" + article.text for article in news]

# Pre-process corpus
corpus_prepro = prepro_corpus(corpus,
                              lowercase_on=True,
                              remove_stop_words_on=True,
                              stemming_on=False,
                              simple_lemmatization_on=False,
```

```
                              pos_lemmatization_on=True,
                              token_pattern_on=True)

# Ignore words that have a df strictly higher/lower than max_df/min_df
vectorizer = CountVectorizer(max_df=1.0,
                             min_df=1)

# Create document-term matrix
dtm = vectorizer.fit_transform(corpus_prepro)

# Create LDA model
lda = LatentDirichletAllocation(n_components=5, random_state=0,
                                verbose=1, evaluate_every=1,
                                max_iter=20)
lda.fit(dtm)

# Get model data
vocab = vectorizer.get_feature_names()
topic_word_distr = lda.components_
doc_topic_distr = lda.transform(dtm)

# Write results to terminal
nbr_topic_words = 30
write_topics_terminal(topic_word_distr, vocab, nbr_topic_words, dtm,
                      lambda_=0.6)

# Write results to file
path = "output/stylenews/"

write_prepro_html(716, corpus, corpus_prepro, path+"prepro.html", news)

vis_data = pyLDAvis.sklearn.prepare(lda, dtm, vectorizer, sort_topics=False)
pyLDAvis.save_html(vis_data, path+"pyLDAvis.html")

write_article_html(716, corpus, corpus_prepro, topic_word_distr,
                   doc_topic_distr, vocab, nbr_topic_words,
                   path+"article.html", dtm, 0.6, True, news=news)

write_title_html(topic_word_distr, doc_topic_distr, vocab, nbr_topic_words,
                 news, 8, dtm, 0.6, path+"title.html")

write_magazine_html(topic_word_distr, doc_topic_distr, vocab, nbr_topic_words,
                    news, dtm, 0.6, path+"magazines.html")

plot_topic_month(news, doc_topic_distr, path+"month.pdf")

plot_topic_year(news, doc_topic_distr, path+"year.pdf")

write_articles_topics_xlsx(news, doc_topic_distr, path+"articles_topics.xls")
```

B.2 Topic modeling module for pre- and post-processing

Listing B.2: Topic modeling module - topic_modeling.py

```
# -*- coding: utf-8 -*-

import datetime

import matplotlib
import matplotlib.pyplot as plt
MYCOLORS = ['royalblue', 'darkorange', 'forestgreen', 'red', 'darkviolet',
            'saddlebrown', 'fuchsia', 'gray', 'limegreen', 'cyan',
            'lightsteelblue', 'moccasin', 'lightgreen', 'lightcoral', 'plum',
            'peru', 'pink', 'lightgray', 'yellowgreen', 'paleturquoise']
MYCOLORS = [matplotlib.colors.CSS4_COLORS[color] for color in MYCOLORS]

from nltk import word_tokenize, pos_tag
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer, PorterStemmer

import numpy as np

import pandas as pd

import re

from sklearn.feature_extraction.text import ENGLISH_STOP_WORDS


class Article(object):

    def __init__(self, magazine="", date=datetime.date(1,1,1),
                 title="", text="", id_nbr=-1):
        self.magazine = magazine
        self.date = date
        self.title = title
        self.text = text
        self.id_nbr = id_nbr


def load_news(filepath, nbr_max=-1):

    article_list = []

    with open(filepath, 'r', encoding='utf8') as file:

        article_list = []

        line = file.readline()
        while line:

            # New article (starts with "Date :")
            if re.match(r"^Date :", line):

                # Stop if specified number of news read
                if len(article_list) >= nbr_max and nbr_max != -1:
                    break

                # Read date
                line = file.readline()
                date = datetime.datetime.strptime(line,"%Y-%m-%d %H:%M:%S ")

                # Read title
```

```
                file.readline()
                title = file.readline().strip()

                # Read magazine
                file.readline()
                magazine = file.readline().strip()

                # Read ID
                file.readline()
                id_nbr = int(file.readline())

                text = ""
                article_list.append(Article(magazine, date, title, text,
                                            id_nbr))

                # Skip empty lines
                file.readline()
                file.readline()
                file.readline()
                file.readline()
            else:
                article_list[-1].text += line

            line = file.readline()

    return article_list


def prepro_corpus(corpus,
                  remove_noise_on=True,
                  lowercase_on=False,
                  remove_stop_words_on=False,
                  stemming_on=False,
                  simple_lemmatization_on=False,
                  pos_lemmatization_on=False,
                  token_pattern_on=False,
                  remove_proper_nouns_on=False):

    c = list(corpus)

    for i, doc in enumerate(corpus):

        # Remove irrelevant information
        if remove_noise_on:
            c[i] = remove_noise(c[i])

        # Remove proper nouns
        if remove_proper_nouns_on:
            c[i] = " ".join([token for token in tokenize(c[i]) \
                             if pos_tag(token)[0][1].startswith('NNP')
                             is False])

        # Case folding
        if lowercase_on:
            c[i] = c[i].lower()

        # Lemmatization (before stop words, e.g. for "'s") without pos-tagging
        if simple_lemmatization_on:
            wnl = WordNetLemmatizer()
            c[i] = " ".join([wnl.lemmatize(w) for w in tokenize(c[i])])

        # Lemmatization (before stop words, e.g. for "'s") with pos-tagging
        if pos_lemmatization_on:
            tokens = tokenize(c[i])
            tokens_pos = pos_tag(tokens)
```

```python
            wnl = WordNetLemmatizer()
            c[i] = " ".join([wnl.lemmatize(token, get_wordnet_pos(pos))
                             for (token, pos) in tokens_pos])

        # Keep only words with 3 or more letters
        if token_pattern_on:
            c[i] = " ".join(re.findall(r"\b[a-z]{3,}\b", c[i]))

        # Stop words
        if remove_stop_words_on:
            stop_words = ENGLISH_STOP_WORDS.union(["per", "cent"])
            c[i] = " ".join([w for w in tokenize(c[i])
                             if w not in stop_words])

        # Stemming (after stop words, since stems might not be in stop words)
        if stemming_on:
            ps = PorterStemmer()
            c[i] = " ".join([ps.stem(w) for w in tokenize(c[i])])

        printProgressBar(i+1, len(corpus), prefix="Prepro:", length=50)

    return c


def remove_noise(doc):

    # Remove "For more information ..."
    d = re.sub(r"For more information.*", "", doc, flags=re.DOTALL)

    # Remove words between #
    d = re.sub(r"#.+#", "", d)

    # Remove URLs
    d = re.sub(r"(www|http:|https:)+[^\s]+[\w]", "", d)

    # Remove email addresses
    d = re.sub(r"(?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*|\"(?:[\
        x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*\"
        )@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])
        ?|\[(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.)
        {3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c
        \x0e-\x1f\x21-\x5a\x53-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])", "", d)

    # Remove words ending in ".com" or ".ru"
    d = re.sub(r"[\w]+(\.com|\.ru)", "", d)

    return d


def get_wordnet_pos(treebank_tag):

    # Transform treebank_tag of pos_tag to wordnet tag
    if treebank_tag.startswith('J'):
        return wordnet.ADJ
    elif treebank_tag.startswith('V'):
        return wordnet.VERB
    elif treebank_tag.startswith('N'):
        return wordnet.NOUN
    elif treebank_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN


def printProgressBar(iteration, total, prefix = '', suffix = '', decimals = 1,
```

```
                       length = 100, fill = '\u2588', printEnd = "\r"):

    percent = 100 * (iteration / float(total))
    percent = ("{0:." + str(decimals) + "f}").format(percent)
    filledLength = int(length * iteration // total)
    bar = fill * filledLength + '-' * (length - filledLength)
    print(f"\r{prefix} |{bar}| {percent}% {suffix}", end=printEnd)
    if iteration == total:
        print()


def write_prepro_html(index, corpus, corpus_prepro, filename, news=None):
    """
    Compatible with lemmatization including pos-tagging.

    """

    doc_orig = corpus[index]
    doc_proc = corpus_prepro[index]

    header = "<!DOCTYPE html>\n" \
             "<html>\n" \
             "<body>\n"

    # Titel (magazine, date, id)
    title = ''
    if news != None:
        title += "<h3>" + news[index].title
        title += " (" + news[index].magazine + ", "
        title += "{:%d/%m/%Y}".format(news[index].date) + ", "
        title += str(news[index].id_nbr) + ")</h3>\n"

    # Pre-processed document
    txt_proc ="<h4>Preprocessed document</h4>\n"
    txt_proc += doc_proc

    # Highlighted document, i.e. text that is kept after preprocessing
    tokens_orig = tokenize(doc_orig)
    tokens_proc = tokenize(doc_proc)

    tokens_cmp = [token.lower() for token in tokens_orig]
    wln = WordNetLemmatizer()

    simple_lemmatization_on = False
    if simple_lemmatization_on:
        tokens_cmp = [wln.lemmatize(token) for token in tokens_cmp]
    else:
        tokens_pos = pos_tag(tokens_cmp)
        tokens_cmp = [wln.lemmatize(token, get_wordnet_pos(pos))
                      for (token, pos) in tokens_pos]

    i = 0
    j = 0
    is_kept_after_prepro = [False]*len(tokens_cmp)
    while i < len(tokens_cmp) and j < len(tokens_proc):
        if tokens_cmp[i] == tokens_proc[j]:
            is_kept_after_prepro[i] = True
            j += 1
        i += 1

    txt_highlight = "\n\n\n<h4>Highlighted document</h4>\n"
    for i in range(len(tokens_orig)):
        if is_kept_after_prepro[i]:
            txt_highlight += " <span style=\"background-color:yellow\">"
            txt_highlight += tokens_orig[i]
```

```
                txt_highlight += "</span> "
            else:
                txt_highlight += " " + tokens_orig[i] + " "

        # Original document
        txt_orig ="\n\n\n<h4>Original document</h4>\n"
        txt_orig += doc_orig.replace("\n","<br />\n")

        footer = "</body>\n" \
                 "</html>"

        # Write to file
        f = open(filename, 'w', encoding='utf-8')
        txt = header + title + txt_proc + txt_highlight + txt_orig + footer
        f.write(txt)
        f.close()


    def tokenize(doc):
        """
        Required to tokenize e.g.:
            - holiday.This ==> holiday This
            - mid/small ==> mid small

        This function should only be called after noise removal so that web and
        mails adresses can still be identified by regex, to remove them.
        """

        tokens = word_tokenize(doc)
        new_tokens = []
        for token in tokens:
            t = re.split("[/'.-]", token)
            for item in t:
                if item != "":
                    new_tokens.append(item)
        return new_tokens


    def write_topics_terminal(topic_word_distr, vocab, n_words, dtm, lambda_=1.0):

        relevance = compute_relevance(topic_word_distr, dtm, lambda_)

        print(" ")
        for i, topic in enumerate(relevance):
            txt = "Topic "+str(i+1)+": "
            txt += " ".join([vocab[i] for i in topic.argsort()[:-n_words-1:-1]])
            txt += "\n-"
            print(txt)


    def compute_relevance(topic_word_distr, dtm, lambda_):
        topic_word_distr_norm = topic_word_distr\
                                /topic_word_distr.sum(axis=1)[:, np.newaxis]
        term_proportion = np.array(dtm.sum(axis=0))/dtm.sum()
        log_lift = np.log(topic_word_distr_norm / term_proportion)
        relevance = lambda_*np.log(topic_word_distr_norm) + (1-lambda_)*log_lift
        return relevance


    def write_article_html(index, corpus, corpus_prepro, topic_word_distr, \
                           doc_topic_distr, vocab, nbr_topic_words, filename, \
                           dtm, lambda_, word_topic_by_doc, news=None):

        doc_orig = corpus[index]
        doc_proc = corpus_prepro[index]
```

```

    header = "<!DOCTYPE html>\n" \
             "<html>\n" \
             "<body>\n" \
             "<head>\n" \
             "<style>\n" \
             "table, th, td {\n" \
             "  border: 1px solid black;\n" \
             "  border-collapse: collapse;\n" \
             "}" \
             "</style>\n" \
             "</head>\n"

    # Titel (magazine, date, id)
    title = ''
    if news != None:
        title += "<h3>" + news[index].title
        title += " (" + news[index].magazine + ", "
        title += "{:%d/%m/%Y}".format(news[index].date) + ", "
        title += str(news[index].id_nbr) + ")</h3>\n"

    # Topics
    topic_word_distr_norm = topic_word_distr\
                            /topic_word_distr.sum(axis=1)[:, np.newaxis]

    # Compute theta_d * beta_k
    if word_topic_by_doc:
        theta_d = doc_topic_distr[index]
        topic_word_distr_norm = np.multiply(topic_word_distr_norm,
                                            theta_d[:, np.newaxis])

    relevance = compute_relevance(topic_word_distr, dtm, lambda_)

    txt_topics ="<h4>Topics</h4>\n"
    txt_topics +="<table>\n"
    for i, topic in enumerate(relevance):
        txt_topics += "  <tr>\n"
        txt_topics += "    <td>"
        txt_topics += "{:>5.0%}".format(doc_topic_distr[index,i])
        txt_topics += "</td>\n"
        txt_topics += "    <td width=60 style=\"background-color:"
        txt_topics += MYCOLORS[i]
        txt_topics += "\">"
        txt_topics += "Topic "+str(i+1)
        txt_topics += "</span>"
        txt_topics += "</td>\n"
        txt_topics += "    <td>"
        txt_topics += " ".join([vocab[i] for i \
                            in topic.argsort()[:-nbr_topic_words-1:-1]])
        txt_topics += "</td>\n"
        txt_topics += "  </tr>\n"
    txt_topics+="</table>"

    # Highlighted text
    tokens_orig = tokenize(doc_orig)
    tokens_proc = tokenize(doc_proc)

    tokens_cmp = [token.lower() for token in tokens_orig]
    wln = WordNetLemmatizer()

    simple_lemmatization_on = False
    if simple_lemmatization_on:
        tokens_cmp = [wln.lemmatize(token) for token in tokens_cmp]
    else:
        tokens_pos = pos_tag(tokens_cmp)
```

```
        tokens_cmp = [wln.lemmatize(token, get_wordnet_pos(pos))
                      for (token, pos) in tokens_pos]

    i = 0
    j = 0
    topic = [-1]*len(tokens_cmp)
    while i < len(tokens_cmp) and j < len(tokens_proc):
        if tokens_cmp[i] == tokens_proc[j]:
            try:
                i_vocab = vocab.index(tokens_cmp[i])
                topic[i] = np.argmax(topic_word_distr_norm[:,i_vocab])
            except ValueError:
                # Possible due to frequency criterion in Vectorizer
                topic[i] = -1
            j += 1
        i += 1

    txt_highlight = "\n\n\n<h4>Highlighted document</h4>\n"
    for i in range(len(tokens_orig)):
        if topic[i]>-1:
            txt_highlight += " <span style=\"background-color:"
            txt_highlight += MYCOLORS[topic[i]]
            txt_highlight += "\">"
            txt_highlight += tokens_orig[i]
            txt_highlight += "</span> "
        else:
            txt_highlight += " " + tokens_orig[i] + " "

    # Original text
    txt_orig ="\n\n\n<h4>Original document</h4>\n"
    txt_orig += doc_orig.replace('\n','<br />\n')

    footer = "</body>\n" \
             "</html>"

    # Write to file
    f = open(filename, "w", encoding="utf-8")
    txt = header + title + txt_topics + txt_highlight + txt_orig + footer
    f.write(txt)
    f.close()


def write_title_html(topic_word_distr, doc_topic_distr, vocab, nbr_topic_words,
                     news, nbr_titles, dtm, lambda_, filename):

    header = "<!DOCTYPE html>\n" \
             "<html>\n" \
             "<body>\n" \
             "<head>\n" \
             "<style>\n" \
             "table, th, td {\n" \
             "  border: 1px solid black;\n" \
             "  border-collapse: collapse;\n" \
             "}" \
             "</style>\n" \
             "</head>\n"

    # Topics
    relevance = compute_relevance(topic_word_distr, dtm, lambda_)

    txt_topics = ""
    for i, topic in enumerate(relevance):

        # Topic
        txt_topics += "<table>\n"
```

```
        txt_topics += "  <tr>\n"
        txt_topics += "    <td width=60 style=\"background-color:"
        txt_topics += MYCOLORS[i]
        txt_topics += "\">"
        txt_topics += "Topic "+str(i+1)
        txt_topics += "</span>"
        txt_topics += "</td>\n"
        txt_topics += "    <td width=1000>"
        txt_topics += " ".join([vocab[i] for i \
                                in topic.argsort()[:-nbr_topic_words-1:-1]])
        txt_topics += "</td>\n"
        txt_topics += "  </tr>\n"

        # Print titles with highest topic probability and probability
        indices = doc_topic_distr[:,i].argsort()[:-nbr_titles-1:-1]
        titles  = [news[j].title for j in indices]
        proba   = [doc_topic_distr[j,i] for j in indices]
        for j, title in enumerate(titles):
            txt_topics += "  <tr>\n"
            txt_topics += "    <td>"
            txt_topics += "{:>5.0%}".format(proba[j])
            txt_topics += "</td>\n"
            txt_topics += "    <td>"
            txt_topics += title
            txt_topics += "</td>\n"
            txt_topics += "  </tr>\n"
        txt_topics += "</table>"
        txt_topics += "<br />\n"

    footer = "</body>\n" \
             "</html>"

    # Write to file
    f = open(filename, 'w', encoding='utf-8')
    txt = header + txt_topics + footer
    f.write(txt)
    f.close()


def write_magazine_html(topic_word_distr, doc_topic_distr, vocab,
                        nbr_topic_words, news, dtm, lambda_, filename):

    header = "<!DOCTYPE html>\n" \
             "<html>\n" \
             "<body>\n" \
             "<head>\n" \
             "<style>\n" \
             "table, th, td {\n" \
             "  border: 1px solid black;\n" \
             "  border-collapse: collapse;\n" \
             "}" \
             "</style>\n" \
             "</head>\n"

    # Topics
    relevance = compute_relevance(topic_word_distr, dtm, lambda_)

    txt_topics  = "<h4>Topics</h4>\n"
    txt_topics += "<table>\n"
    for i, topic in enumerate(relevance):
        txt_topics += "  <tr>\n"
        txt_topics += "    <td width=60 style=\"background-color:"
        txt_topics += MYCOLORS[i]
        txt_topics += "\">"
        txt_topics += "Topic "+str(i+1)
```

```
        txt_topics += "</span>"
        txt_topics += "</td>\n"
        txt_topics += "    <td>"
        txt_topics += " ".join([vocab[i] for i \
                        in topic.argsort()[:-nbr_topic_words-1:-1]])
        txt_topics += "</td>\n"
        txt_topics += "  </tr>\n"
    txt_topics += "</table>\n"

    # For each magazine, compute topic distribution
    magazines = list(dict.fromkeys([article.magazine for article in news]))
    if len(magazines) == 9:
        magazines = ['Euromoney', 'InstitutionalInvestor', 'FTAdviser',
                     'AlphaQ', 'InstitutionalAsset', 'WealthAdviser',
                     'IPE', 'PlanSponsor', 'PlanAdviser']
    nbr_articles_magazine = np.zeros(len(magazines))
    magazine_topic_distr = np.zeros((len(magazines), doc_topic_distr.shape[1]))

    for i, article in enumerate(news):
        magazine_index = magazines.index(article.magazine)
        nbr_articles_magazine[magazine_index] += 1
        magazine_topic_distr[magazine_index] += doc_topic_distr[i]

    for i, nbr in enumerate(nbr_articles_magazine):
        magazine_topic_distr[i] /= nbr

    txt_mag = "<h4>Topic distribution for each magazine </h4>\n"
    n_col = doc_topic_distr.shape[1]+1 # Nbr topics + 1
    n_row = len(magazines)+1

    txt_mag += "<table>\n"
    for i in range(n_row):
        for j in range(n_col):
            if i == 0:
                if j == 0:
                    txt_mag += "  <tr>\n"
                    txt_mag += "    <td>"
                    txt_mag += "</td>\n"
                else:
                    txt_mag += "    <td width=60 style=\"background-color:"
                    txt_mag += MYCOLORS[j-1]
                    txt_mag += "\">"
                    txt_mag += "Topic "+str(j)
                    txt_mag += "</span>"
                    txt_mag += "</td>\n"
            else:
                if j == 0:
                    txt_mag += "  <tr>\n"
                    txt_mag += "    <td>"
                    txt_mag += magazines[i-1]
                    txt_mag += "</td>\n"
                else:
                    txt_mag += "    <td width=60 style=\"background-color:"
                    cmap = matplotlib.cm.get_cmap('Greens')
                    rgba = cmap(magazine_topic_distr[i-1,j-1])
                    txt_mag += matplotlib.colors.to_hex(rgba)
                    txt_mag += "\">"
                    txt_mag += "{:>5.0%}".format(magazine_topic_distr[i-1,j-1])
                    txt_mag += "</span>"
                    txt_mag += "</td>\n"

    txt_mag += "</table>\n"

    footer = "</body>\n" \
             "</html>"
```

```

    # Write to file
    f = open(filename, 'w', encoding='utf-8')
    txt = header + txt_topics + txt_mag + footer
    f.write(txt)
    f.close()


def plot_topic_year(news, doc_topic_distr, filename=''):

    # Compute topic percentages for each year
    dates = [article.date for article in news]
    years = max(dates).year - min(dates).year + 1
    n_topics = doc_topic_distr.shape[1]

    topic_year_distr = np.zeros((n_topics, years))
    news_freq_month = np.zeros(years)

    for i, article in enumerate(news):
        year = article.date.year - min(dates).year
        topic_year_distr[:, year] += doc_topic_distr[i,:]
        news_freq_month[year] += 1

    rel_importance = np.zeros(topic_year_distr.shape)
    for year in range(years):
        rel_importance[:, year] = topic_year_distr[:, year]\
                                  /news_freq_month[year]

    # Plot absolute importance for each year
    fig, ax = plt.subplots(figsize=(6, 4.5))
    for i in range(n_topics):
        ax.plot(np.arange(1, years+1), topic_year_distr[i,:],
                color=MYCOLORS[i], linewidth=2)
    legend = ["Topic "+str(i+1) for i in range(n_topics)]
    ax.legend(legend, fontsize=11, loc='upper right', framealpha=1)
    ax.set_xlabel("Year [-]", fontsize=14)
    ax.set_ylabel("Absolute importance [-]", fontsize=14)
    ticks = [str(i) for i in np.arange(min(dates).year,max(dates).year+1)]
    plt.xticks(np.arange(1, years+1), ticks)
    ax.set_xlim(1,12)
    plt.xticks(fontsize=11)
    plt.yticks(fontsize=14)
    plt.grid(True)
    plt.tight_layout()

    if filename != '':
       filename1 = "{0}_{2}.{1}".format(*filename.rsplit('.', 1) + ["abs"])
       plt.savefig(filename1, format='pdf')

    # Plot relative importance for each year
    fig, ax = plt.subplots(figsize=(6, 4.5))
    for i in range(n_topics):
        ax.plot(np.arange(1, years+1), rel_importance[i,:]*100,
                color=MYCOLORS[i], linewidth=2)
    legend = ["Topic "+str(i+1) for i in range(n_topics)]
    ax.legend(legend, fontsize=11, loc='upper right', framealpha=1)
    ax.set_xlabel("Year [-]", fontsize=14)
    ax.set_ylabel("Relative importance [%]", fontsize=14)
    ticks = [str(i) for i in np.arange(min(dates).year,max(dates).year+1)]
    plt.xticks(np.arange(1, years+1), ticks)
    ax.set_xlim(1,12)
    plt.xticks(fontsize=11)
    plt.yticks(fontsize=14)
    plt.grid(True)
    plt.tight_layout()
```

```
    if filename != '':
       filename2 = "{0}_{2}.{1}".format(*filename.rsplit('.', 1) + ["rel"])
       plt.savefig(filename2, format='pdf')


def plot_topic_month(news, doc_topic_distr, filename=''):

    # Compute topic percentages for each month
    dates = [article.date for article in news]
    n_topics = doc_topic_distr.shape[1]

    topic_month_distr = np.zeros((n_topics, 12))
    news_freq_month = np.zeros(12)

    for i, article in enumerate(news):
        month = article.date.month-1
        topic_month_distr[:, month] += doc_topic_distr[i,:]
        news_freq_month[month] += 1

    for month in range(12):
        topic_month_distr[:, month] /= news_freq_month[month]

    # Plot topic percentages for each month
    fig, ax = plt.subplots(figsize=(6, 4.5))
    for i in range(n_topics):
        ax.plot(np.arange(1, 13), topic_month_distr[i,:]*100,
                color=MYCOLORS[i], linewidth=2)
    legend = ["Topic "+str(i+1) for i in range(n_topics)]
    ax.legend(legend, fontsize=11, loc='upper right', framealpha=1)
    ax.set_xlabel("Month [-]", fontsize=14)
    ax.set_ylabel("Relative importance [%]", fontsize=14)
    ticks = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep',
             'Oct', 'Nov', 'Dec']
    plt.xticks(np.arange(1, 13), ticks)
    ax.set_xlim(1,16)
    plt.xticks(fontsize=11)
    plt.yticks(fontsize=14)
    plt.grid(True)
    plt.tight_layout()

    if filename != '':
       plt.savefig(filename, format='pdf')


def write_articles_topics_xlsx(news, doc_topic_distr, filename):

    dates  = ["{:%d/%m/%Y}".format(article.date) for article in news]
    titles = [article.title for article in news]
    magazines = [article.magazine for article in news]
    ids = [article.id_nbr for article in news]
    data = list(zip(dates, titles, magazines, ids))
    df1 = pd.DataFrame(data, columns=['Date','Title','Magazine','ID'])

    columns = []
    nbr_topics = doc_topic_distr.shape[1]
    for i in range(nbr_topics):
        columns.append('Topic '+str(i+1))
    df2 = pd.DataFrame(doc_topic_distr, columns=columns)

    df = pd.concat([df1, df2], axis=1)

    df.to_excel(filename)
```

Appendix C

Examples of HTML result files

To control the pre-processing of the textual data and to interpret the results of the topic model, a number of HTML documents can be created by the PYTHON functions in the previous chapter of the appendix. These documents are illustrated in this appendix. They were created with the parameters of configuration 1 in Tab. 5.1.

C.1 Results of pre-processing

Fidelity replaces manager of European Opps fund (FTAdviser, 30/06/2014, 63617)

Preprocessed document

fidelity replaces manager european opps fund fidelity replace colin stone alberto chiandetti underperforming fidelity european opportunity fund stone manage fund slip quartile ima european sector year accord data analytics continue manage fidelity european small cap strategy include offshore european small company fund fidelity say chiandetti run european opportunity fund alongside stone october sole responsibility chiandetti remain manager luxembourg domicile italy switzerland fund fidelity say allocate dedicated resource support country fund

Highlighted document

Fidelity replaces manager of European Opps fund # paragraph # Fidelity has moved to replace Colin Stone with Alberto Chiandetti on the underperforming £432m Fidelity European Opportunities fund # paragraph # Mr Stone had been managing the fund since 2003 but it had slipped into the bottom quartile of the IMA European sector for three and five years , according to data from FE Analytics # paragraph # He will continue to manage the Fidelity ’ s European small cap strategy , including the offshore FF European Smaller Companies fund # paragraph # Fidelity said Mr Chiandetti would run the European Opportunities fund alongside Mr Stone until October before taking sole responsibility # paragraph # Mr Chiandetti will remain as manager of the Luxembourg domiciled FF Italy and FF Switzerland funds , though Fidelity said he had been “ allocated dedicated resources to support these country funds ”

Original document

Fidelity replaces manager of European Opps fund

#paragraph# Fidelity has moved to replace Colin Stone with Alberto Chiandetti on the underperforming £432m Fidelity European Opportunities fund.

#paragraph# Mr Stone had been managing the fund since 2003 but it had slipped into the bottom quartile of the IMA European sector for three and five years, according to data from FE Analytics.

#paragraph# He will continue to manage the Fidelity’s European small cap strategy, including the offshore FF European Smaller Companies fund.

#paragraph# Fidelity said Mr Chiandetti would run the European Opportunities fund alongside Mr Stone until October before taking sole responsibility.

#paragraph# Mr Chiandetti will remain as manager of the Luxembourg-domiciled FF Italy and FF Switzerland funds, though Fidelity said he had been “allocated dedicated resources to support these country funds”.

Figure C.1: Results of pre-processing.

C.2 LDAvis file

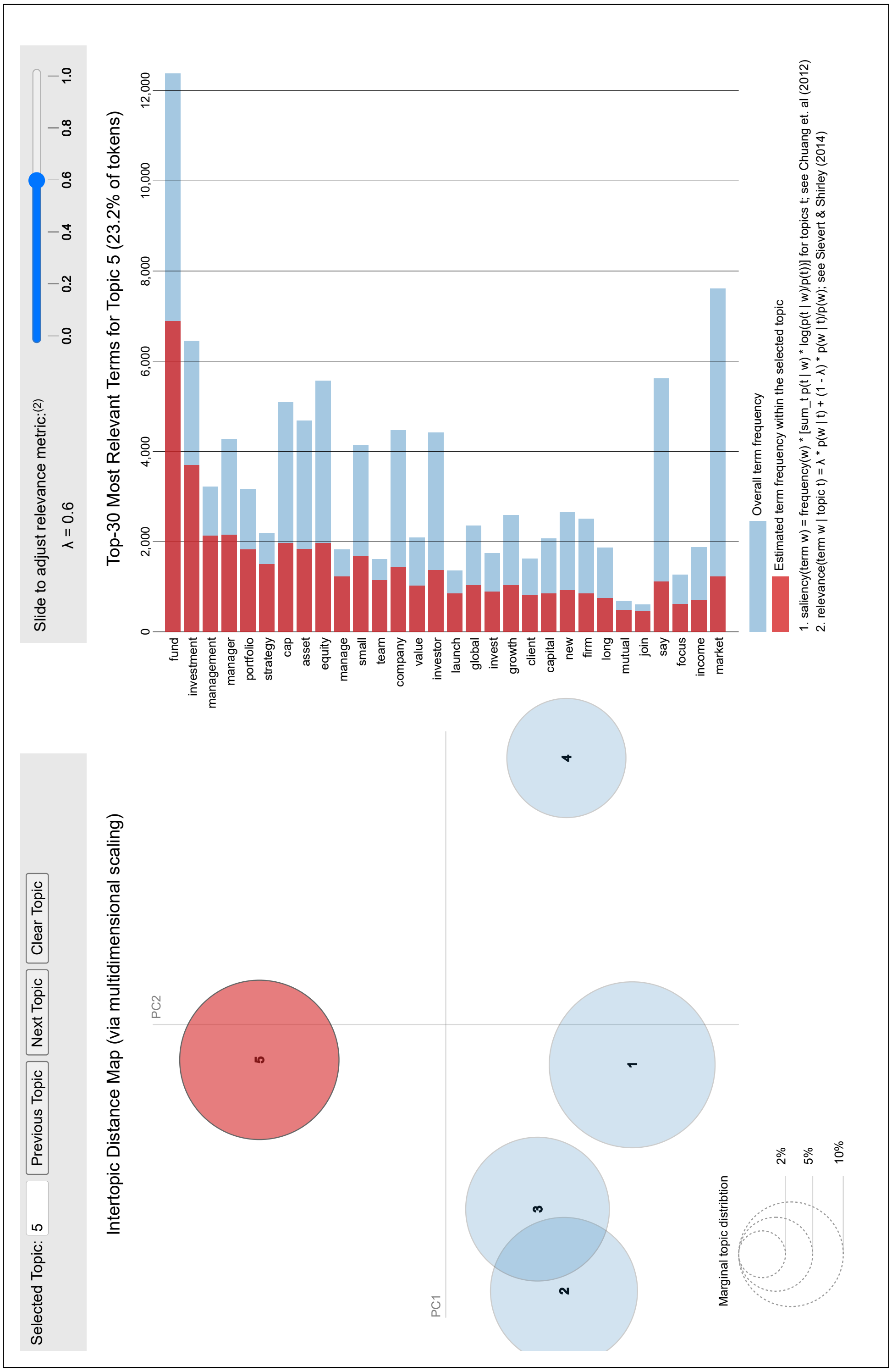


Figure C.2: LDAvis file.

C.3 Topic-title file

Topic 1	market year equity sector say company investor growth return stock high european rate economy manager rise china price yield economic fund bond term europe earnings asset small risk expect strong
100%	WisdomTree warns political risk is at European gates
100%	Multi
100%	WisdomTree is bullish on Japan
100%	Half year report shows some key indicators ‘flashing red’
100%	Markets in 2016 How to Separate Signals from Noise
100%	European equities the future drivers of returns
100%	Upbeat diagnosis for healthcare
100%	Dollar-equity correlation conjures up memories of dotcom boom

Topic 2	bank firm market business trading say trade client research capital deal company banking year ipo new private finance million analyst liquidity make work exchange loan need big want broker time
100%	2015 All-America Research Team Meet the Rising Stars
100%	Yearn to Learn Molding the Rising Stars of Wall Street
100%	Nasdaq and AX Trading Look at Block Trade Alternative To HFT
100%	J.P. Morgan’s Joseph Greff Joins All-America Hall of Fame
100%	2 Firms Share Title of America’s Top Corporate Access Provider
100%	Five Questions CA Cheuvreux’s Ian Peacock on HFT Anxiety
100%	Flight Path for the SEC’s Tick-Size Pilot
100%	SMEs shift gears as cross-border trade grows

Topic 3	plan pension retirement fund percent investment fee participant say sponsor asset make share newsdash hedge fiduciary endowment use active money think employee cost option manager target time return year portfolio
100%	United Technologies CIO Robin Diamonte Has Lifetime Income Plans
100%	Tune Up Your DC Plan in 2014
100%	BlackRock CEO Mulls Retirement in Twitter Era
100%	BlackRock CEO Mulls Retirement in Twitter Era
100%	Making Sure Plan Fees Are Reasonable
100%	Are Multiemployer Plans Understating Their Liabilities
100%	What Plan Sponsors Should Know About the Final Fiduciary Rule
100%	A Plan Sponsor Hires a 3(38) Investment Manager

Topic 4	index etf cap market vanguard russell fund billion msci equity large stock inflow return factor small performance emerge quarter month total weight asset beta ishares category exposure bond benchmark low
100%	World’s largest stock and bond funds report lower expense ratios
100%	Actively Managed Funds Fail to Beat Benchmarks
100%	Passively Managed Funds Trounce Actively Managed Funds
100%	October Brought Heavy Trading in 401(k)s
100%	DC Participants Less Active Traders in March
100%	Mercer Finds Equity Markets End 2009 Strong
100%	No Strong Participant Reaction to Market Swings
100%	Target Maturity Fund Performance Climbs Back Up in 2009

Topic 5	fund investment management manager portfolio strategy cap asset equity manage small team company value investor launch global invest growth client capital new firm long mutual join say focus income market
100%	Manulife launches 15 new funds
100%	Mairs & Power Mutual Funds announce co-portfolio manager and officer changes
100%	Wednesday people roundup
100%	Sentry Investments adds two senior portfolio managers
100%	Balter converts London and NYC-based hedge funds to liquid alts mutual funds
100%	Empire Life launches seven new global funds
100%	Neuberger Berman introduces Absolute Return Multi-Manager Fund
100%	Franklin Templeton proposes changes for two Bissett Balanced Fund mandates

Figure C.3: Topic-title file with the respective topic proportion in the corresponding article.

C.4 Topic-article files

C.4.1 Without taking the context of the document into consideration

Fidelity replaces manager of European Opps fund (FTAdviser, 30/06/2014, 63617)

Topics

39%	Topic 1	market year equity sector say company investor growth return stock high european rate economy manager rise china price yield economic fund bond term europe earnings asset small risk expect strong
0%	Topic 2	bank firm market business trading say trade client research capital deal company banking year ipo new private finance million analyst liquidity make work exchange loan need big want broker time
0%	Topic 3	plan pension retirement fund percent investment fee participant say sponsor asset make share newsdash hedge fiduciary endowment use active money think employee cost option manager target time return year portfolio
0%	Topic 4	index etf cap market vanguard russell fund billion msci equity large stock inflow return factor small performance emerge quarter month total weight asset beta ishares category exposure bond benchmark low
60%	Topic 5	fund investment management manager portfolio strategy cap asset equity manage small team company value investor launch global invest growth client capital new firm long mutual join say focus income market

Highlighted document

Fidelity replaces manager of European Opps fund # paragraph # Fidelity has moved to replace Colin Stone with Alberto Chiandetti on the underperforming £432m Fidelity European Opportunities fund # paragraph # Mr Stone had been managing the fund since 2003 but it had slipped into the bottom quartile of the IMA European sector for three and five years , according to data from FE Analytics # paragraph # He will continue to manage the Fidelity ’ s European small cap strategy , including the offshore FF European Smaller Companies fund # paragraph # Fidelity said Mr Chiandetti would run the European Opportunities fund alongside Mr Stone until October before taking sole responsibility # paragraph # Mr Chiandetti will remain as manager of the Luxembourg domiciled FF Italy and FF Switzerland funds , though Fidelity said he had been “ allocated dedicated resources to support these country funds ”

Original document

Fidelity replaces manager of European Opps fund

#paragraph# Fidelity has moved to replace Colin Stone with Alberto Chiandetti on the underperforming £432m Fidelity European Opportunities fund.

#paragraph# Mr Stone had been managing the fund since 2003 but it had slipped into the bottom quartile of the IMA European sector for three and five years, according to data from FE Analytics.

#paragraph# He will continue to manage the Fidelity’s European small cap strategy, including the offshore FF European Smaller Companies fund.

#paragraph# Fidelity said Mr Chiandetti would run the European Opportunities fund alongside Mr Stone until October before taking sole responsibility.

#paragraph# Mr Chiandetti will remain as manager of the Luxembourg-domiciled FF Italy and FF Switzerland funds, though Fidelity said he had been “allocated dedicated resources to support these country funds”.

Figure C.4: Topic-article file without taking the context of the document into consideration, i.e. Eq. (4.2).

C.4.2 Taking the context of the document into consideration

Fidelity replaces manager of European Opps fund (FTAdviser, 30/06/2014, 63617)

Topics

39%	Topic 1	market year equity sector say company investor growth return stock high european rate economy manager rise china price yield economic fund bond term europe earnings asset small risk expect strong
0%	Topic 2	bank firm market business trading say trade client research capital deal company banking year ipo new private finance million analyst liquidity make work exchange loan need big want broker time
0%	Topic 3	plan pension retirement fund percent investment fee participant say sponsor asset make share newsdash hedge fiduciary endowment use active money think employee cost option manager target time return year portfolio
0%	Topic 4	index etf cap market vanguard russell fund billion msci equity large stock inflow return factor small performance emerge quarter month total weight asset beta ishares category exposure bond benchmark low
60%	Topic 5	fund investment management manager portfolio strategy cap asset equity manage small team company value investor launch global invest growth client capital new firm long mutual join say focus income market

Highlighted document

Fidelity replaces manager of European Opps fund # paragraph # Fidelity has moved to replace Colin Stone with Alberto Chiandetti on the underperforming £432m Fidelity European Opportunities fund # paragraph # Mr Stone had been managing the fund since 2003 but it had slipped into the bottom quartile of the IMA European sector for three and five years , according to data from FE Analytics # paragraph # He will continue to manage the Fidelity ' s European small cap strategy , including the offshore FF European Smaller Companies fund # paragraph # Fidelity said Mr Chiandetti would run the European Opportunities fund alongside Mr Stone until October before taking sole responsibility # paragraph # Mr Chiandetti will remain as manager of the Luxembourg domiciled FF Italy and FF Switzerland funds , though Fidelity said he had been " allocated dedicated resources to support these country funds "

Original document

Fidelity replaces manager of European Opps fund

#paragraph# Fidelity has moved to replace Colin Stone with Alberto Chiandetti on the underperforming £432m Fidelity European Opportunities fund.

#paragraph# Mr Stone had been managing the fund since 2003 but it had slipped into the bottom quartile of the IMA European sector for three and five years, according to data from FE Analytics.

#paragraph# He will continue to manage the Fidelity's European small cap strategy, including the offshore FF European Smaller Companies fund.

#paragraph# Fidelity said Mr Chiandetti would run the European Opportunities fund alongside Mr Stone until October before taking sole responsibility.

#paragraph# Mr Chiandetti will remain as manager of the Luxembourg-domiciled FF Italy and FF Switzerland funds, though Fidelity said he had been "allocated dedicated resources to support these country funds".

Figure C.5: Topic-article file by taking the context of the document into consideration, i.e. Eq. (4.3).

C.5 Topic-magazine file

Topics

Topic 1	market year equity sector say company investor growth return stock high european rate economy manager rise china price yield economic fund bond term europe earnings asset small risk expect strong
Topic 2	bank firm market business trading say trade client research capital deal company banking year ipo new private finance million analyst liquidity make work exchange loan need big want broker time
Topic 3	plan pension retirement fund percent investment fee participant say sponsor asset make share newsdash hedge fiduciary endowment use active money think employee cost option manager target time return year portfolio
Topic 4	index etf cap market vanguard russell fund billion msci equity large stock inflow return factor small performance emerge quarter month total weight asset beta ishares category exposure bond benchmark low
Topic 5	fund investment management manager portfolio strategy cap asset equity manage small team company value investor launch global invest growth client capital new firm long mutual join say focus income market

Topic distribution for each magazine

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Euromoney	32%	59%	4%	2%	3%
InstitutionalInvestor	25%	26%	23%	12%	13%
FTAdviser	65%	4%	3%	4%	24%
AlphaQ	34%	4%	14%	9%	39%
InstitutionalAsset	14%	25%	2%	17%	41%
WealthAdviser	16%	6%	2%	18%	59%
IPE	21%	9%	27%	7%	36%
PlanSponsor	5%	3%	21%	49%	21%
PlanAdviser	10%	3%	17%	48%	23%

Figure C.6: Topic-magazine file.

Appendix D

Results of the topic model

In this chapter, the results of the topic model are reproduced for the different configurations of parameters in Tab. 5.1.

D.1 Configuration 1

Topic 1	market year equity sector say company investor growth return stock high european rate economy manager rise china price yield economic fund bond term europe earnings asset small risk expect strong
100%	WisdomTree warns political risk is at European gates
100%	Multi
100%	WisdomTree is bullish on Japan
100%	Half year report shows some key indicators 'flashing red'
100%	Markets in 2016 How to Separate Signals from Noise
100%	European equities the future drivers of returns
100%	Upbeat diagnosis for healthcare
100%	Dollar-equity correlation conjures up memories of dotcom boom

Topic 2	bank firm market business trading say trade client research capital deal company banking year ipo new private finance million analyst liquidity make work exchange loan need big want broker time
100%	2015 All-America Research Team Meet the Rising Stars
100%	Yearn to Learn Molding the Rising Stars of Wall Street
100%	Nasdaq and AX Trading Look at Block Trade Alternative To HFT
100%	J.P. Morgan's Joseph Greff Joins All-America Hall of Fame
100%	2 Firms Share Title of America's Top Corporate Access Provider
100%	Five Questions CA Cheuvreux's Ian Peacock on HFT Anxiety
100%	Flight Path for the SEC's Tick-Size Pilot
100%	SMEs shift gears as cross-border trade grows

Topic 3	plan pension retirement fund percent investment fee participant say sponsor asset make share newsdash hedge fiduciary endowment use active money think employee cost option manager target time return year portfolio
100%	United Technologies CIO Robin Diamonte Has Lifetime Income Plans
100%	Tune Up Your DC Plan in 2014
100%	BlackRock CEO Mulls Retirement in Twitter Era
100%	BlackRock CEO Mulls Retirement in Twitter Era
100%	Making Sure Plan Fees Are Reasonable
100%	Are Multiemployer Plans Understating Their Liabilities
100%	What Plan Sponsors Should Know About the Final Fiduciary Rule
100%	A Plan Sponsor Hires a 3(38) Investment Manager

Topic 4	index etf cap market vanguard russell fund billion msci equity large stock inflow return factor small performance emerge quarter month total weight asset beta ishares category exposure bond benchmark low
100%	World's largest stock and bond funds report lower expense ratios
100%	Actively Managed Funds Fail to Beat Benchmarks
100%	Passively Managed Funds Trounce Actively Managed Funds
100%	October Brought Heavy Trading in 401(k)s
100%	DC Participants Less Active Traders in March
100%	Mercer Finds Equity Markets End 2009 Strong
100%	No Strong Participant Reaction to Market Swings
100%	Target Maturity Fund Performance Climbs Back Up in 2009

Topic 5	fund investment management manager portfolio strategy cap asset equity manage small team company value investor launch global invest growth client capital new firm long mutual join say focus income market
100%	Manulife launches 15 new funds
100%	Mairs & Power Mutual Funds announce co-portfolio manager and officer changes
100%	Wednesday people roundup
100%	Sentry Investments adds two senior portfolio managers
100%	Balter converts London and NYC-based hedge funds to liquid alts mutual funds
100%	Empire Life launches seven new global funds
100%	Neuberger Berman introduces Absolute Return Multi-Manager Fund
100%	Franklin Templeton proposes changes for two Bissett Balanced Fund mandates

Figure D.1: Top 30 words and top 8 articles for the topics in configuration 1.

D.2 Configuration 2

Topic 1	plan retirement pension fund participant sponsor fee investment newsdash fiduciary contribution say share target endowment employee date asset option active liability employer define cost class benefit use court make allocation
100%	Chevron Wins Dismissal of ERISA Challenge
100%	Chevron Wins Dismissal of Amended Complaint Regarding Fund Choices
100%	Participant Challenges Prudential and Morningstar Allocation Solution
100%	Self-Dealing Suit Challenges Fees and Fund Monitoring
100%	Understanding Share Classes in DC Plan Funds
100%	Understanding Mutual Fund Share Classes
100%	Making Sure Plan Fees Are Reasonable
100%	What Plan Sponsors Should Know About the Final Fiduciary Rule

Topic 2	team management join esg investment manager equity analyst appoint senior asset head director global manage swiss research portfolio firm rbc experience responsible client year cap serve role work mandate service
100%	Wednesday people roundup
100%	Eaton Vance expands global equity team
100%	Davy Asset Management expands European operations with senior hire
100%	Matrix hires UK real estate team from JP Morgan Cazenove
100%	Rockefeller Capital Management appoints Head of Institutional Distribution
100%	Thomas Weisel expands research and private client services
100%	Deutsche Asset Management makes new management appointments
100%	Edge Asset Management hires Cliff Remily and Toby Jayne

Topic 3	index market etf cap equity return fund stock year small large emerge asset month performance billion quarter bond msci exposure sector high russell factor investor low volatility growth global yield
100%	Target Maturity Funds Have Tough Second Quarter
100%	Target Maturity Funds Have Tough Second Quarter
100%	U.S. Large Caps Flex Muscle in Russell Index Rebalancing
100%	Target-Date Funds Extend Performance Winning Streak
100%	ETFs Enjoy March Inflows of $20B
100%	ETFs Enjoy $20M March Inflows
100%	Actively Managed Funds Fail to Beat Benchmarks
100%	Passively Managed Funds Trounce Actively Managed Funds

Topic 4	fund investment strategy portfolio manager cap small launch management invest investor manage equity growth company asset value long income market mutual provide capital opportunity team client seek focus new offer
100%	Henderson Group announces proposed acquisition of Gartmore
100%	Putnam Investments to launch suite of multi-cap equity funds
100%	Jupiter plans launch of Emerging & Frontier Income Trust
100%	Janus Capital Group launches Asian and Japanese equity funds
100%	Abhay Deshpande launches Centerstone Investors
100%	Aristotle launches Aristotle Value Equity Fund Class I
100%	Putnam to Launch Multi-Cap Equity Funds
100%	Schroders launches first fund investing purely in onshore China

Topic 5	bank say year market company percent firm business make investor time capital big think like good trade deal private look billion buy trading financial need price stock new come research
100%	May Day II, (Institutional Investor, February 1999)
100%	Liquidnet's Merrin Wants Main Street to Dump Wall Street
100%	Cash management strategy debate Cash management in a world of risk and complexity
100%	Greek Banks Lure Foreign Investors Betting on a Turnaround
100%	Russia debate Russia pushes on with financial markets developments
100%	Death of the IPO
100%	Russian Woodlands Are a New Green Frontier
100%	Germany's Helaba to the Rescue

Figure D.2: Top 30 words and top 8 articles for the topics in configuration 2.

D.3 Configuration 3

Topic 1	market year sector equity say growth investor company european rate return high stock economy rise manager yield price economic earnings bond europe china term strong expect dividend japan valuation remain
100%	Multi
100%	Snapshot Europe shines but any setback could be fierce
100%	Syz comments on Trump victory
100%	UK equity markets resilient in 2014
100%	Barings sees greater signs of recovery in Western economies
100%	Investment Managers Dim on U.S. Economic Outlook
100%	Analysts caution against confidence in inflation dip
100%	Fund Selector Stuck in a holding pattern

Topic 2	bank firm market say trading business trade research client capital deal company banking new year million ipo private finance analyst liquidity exchange make work loan big need investor buy want
100%	Nasdaq and AX Trading Look at Block Trade Alternative To HFT
100%	2 Firms Share Title of America's Top Corporate Access Provider
100%	Five Questions CA Cheuvreux's Ian Peacock on HFT Anxiety
100%	Flight Path for the SEC's Tick-Size Pilot
100%	2015 All-America Research Team Welcomes 30 Newcomers
98%	May Day II, (Institutional Investor, February 1999)
98%	Viet Capital blazes a trail
98%	Back to the Future for Small-Company Capital

Topic 3	plan pension retirement percent fund investment say fee participant sponsor hedge make asset share newsdash fiduciary use money endowment cost think employee time target option pay year risk liability return
100%	Tune Up Your DC Plan in 2014
100%	BlackRock CEO Mulls Retirement in Twitter Era
100%	BlackRock CEO Mulls Retirement in Twitter Era
100%	Are Multiemployer Plans Understating Their Liabilities
100%	What Plan Sponsors Should Know About the Final Fiduciary Rule
100%	PSNC 2013 Up at Night
100%	PSNC 2013 Up at Night
99%	United Technologies CIO Robin Diamonte Has Lifetime Income Plans

Topic 4	index etf cap market fund vanguard russell equity billion msci large stock return small performance factor emerge inflow beta asset total month quarter exposure weight category ishares benchmark bond active
100%	World's largest stock and bond funds report lower expense ratios
100%	DC Participants Less Active Traders in March
100%	Mercer Finds Equity Markets End 2009 Strong
100%	No Strong Participant Reaction to Market Swings
100%	S&P Dow Jones Indices continues South Africa expansion
100%	Vanguard to launch two dividend oriented funds and ETFs
100%	2014 Closed With Light DC Plan Trading
100%	DC Plan Trading Activity Picked Up in January

Topic 5	fund investment management manager portfolio strategy asset cap manage equity team small company value investor invest launch global growth client capital new firm long join focus say mutual income provide
100%	Mairs & Power Mutual Funds announce co-portfolio manager and officer changes
100%	Sentry Investments adds two senior portfolio managers
100%	Balter converts London and NYC-based hedge funds to liquid alts mutual funds
100%	Neuberger Berman introduces Absolute Return Multi-Manager Fund
100%	Franklin Templeton proposes changes for two Bissett Balanced Fund mandates
100%	TA Associates backs buyout of Goldman Sachs Aussie investment platform
100%	Ankur Crawford joins Patrick Kelly as Portfolio Manager on Alger SICAV
100%	Putnam Investments to launch suite of multi-cap equity funds

Figure D.3: Top 30 words and top 8 articles for the topics in configuration 3.

D.4 Configuration 4

Topic 1	market year sector growth equity company high stock european rate investor rise return economy say cap yield price manager small earnings dividend economic europe fund strong term month income ftse
100%	Snapshot Europe shines but any setback could be fierce
100%	UK equity markets resilient in 2014
100%	Fund Selector Markets are finely balanced
100%	Fund Selector Stuck in a holding pattern
100%	Rising valuations stoke caution but sterling weakness to support sector
100%	Is the FTSE no longer benefiting from pound weakness
100%	Five macroeconomic factors driving European equities
100%	Neil Wilkinson waits on small caps

Topic 2	trading market ipo exchange trade trader russia russian company volume say bat order firm listing broker stock moscow new commission block liquidnet vector brokerage nasdaq electronic execution deal maker technology
100%	Flight Path for the SEC’s Tick-Size Pilot
97%	BATS Tries to Reboot Its IPO
95%	Nasdaq and AX Trading Look at Block Trade Alternative To HFT
94%	May Day II, (Institutional Investor, February 1999)
91%	JP Morgan starts trading in SLS
91%	The Tick Size Pilot Key Trading Considerations
86%	BCS Global Markets completes first IPO
85%	Liquidnet's Merrin Wants Main Street to Dump Wall Street

Topic 3	percent pension newsdash retirement employee employer activist plan say new worker endowment state board school health university cio benefit public kemna callan pay proxy retiree hedge wisconsin year saving financial
93%	Hewlett-Packard the Latest to Bow to Shareholder Pressure
92%	Western Union Ups the Ante in Proxy Access Battles
92%	2011 NewsDash Archive List
85%	BlackRock CEO Mulls Retirement in Twitter Era
85%	BlackRock CEO Mulls Retirement in Twitter Era
82%	Wisconsin's Public Pension Works to Spread the Cheddar
79%	Taft-Hartley blues
76%	PSNC 2013 Up at Night

Topic 4	index etf cap msci russell market weight billion ishares inflow emerge small large stock spdr factor exposure etfs volatility global beta total track month outflow performance capitalization dow category mid
99%	SsgA Introduces Low Volatility ETFs
99%	SsgA Introduces Low Volatility ETFs
99%	MSCI Launches New Indexes for Developed Markets
99%	ProShares Launches Daily 3x and -3x ETFs
99%	MSCI Unveils Micro Cap Indices
90%	Russell Investments Launches 10 Factor ETFs
90%	Rydex Launches Two New S&P Equal Weight ETFs
89%	Rydex Launches Two S&P Equal Weight ETFs

Topic 5	fund investment management portfolio team cap manager small manage strategy equity asset company launch growth value global capital join invest investor new client long focus mutual experience provide income firm
100%	Ankur Crawford joins Patrick Kelly as Portfolio Manager on Alger SICAV
100%	Putnam Investments to launch suite of multi-cap equity funds
100%	Pegasus UCITS Fund restructures and rebrands as Tosca Micro Cap UCITS Fund
100%	CI Investments re-opens Cambridge Canadian Growth Companies Fund
100%	Ranger enters MF marketplace with launch of two new funds
100%	Sentry Investments and Sun Life Global Investments expand partnership with three new funds
100%	Value Line renames two funds as ‘focused’ funds
100%	Abhay Deshpande launches Centerstone Investors

Figure D.4: Top 30 words and top 8 articles for the topics in configuration 4.

Topic 6	plan vanguard fund fee share expense sponsor retirement participant class option hancock investment fiduciary cost plaintiff john offer complaint ratio nextpage mutual defendant charge menu duty available erisa fidelity collective
100%	Kraft Suit Plaintiffs Denied Class Status on Remaining Claim
100%	John Hancock Establishes New Series of R Share Classes
99%	John Hancock mutual funds launches new R6 share class
95%	John Hancock Establishes New Series of R Share Classes
93%	ING U.S. Unveils R6 Shares
91%	ING U.S. Unveils R6 Shares
90%	Union Fund Hit With Excessive Fee Suit
86%	Participant Challenges Prudential and Morningstar Allocation Solution

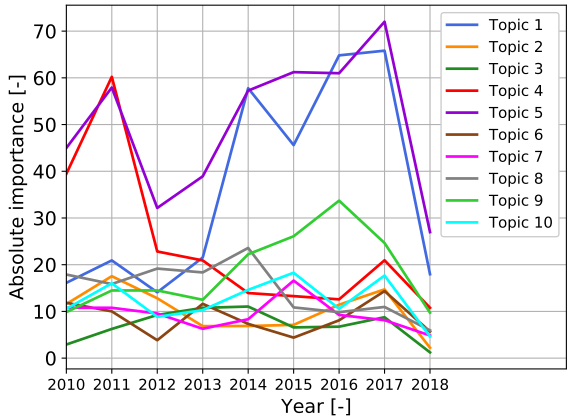
Topic 7	bank banking loan finance business lending bnp paribas year credit smes corporates lender capital client euromoney analyst billion runner trade america say financing customer crisis financial need work cash street
96%	Germany's Helaba to the Rescue
95%	2015 All-America Research Team How the Firms Fared
94%	2015 All-America Research Team Welcomes 30 Newcomers
93%	Spain ICO fills the gap
93%	Italy ECB's first merger brings more worry
90%	J.P. Morgan's Joseph Greff Joins All-America Hall of Fame
90%	2015 All-America Research Team Key Facts and Figures
88%	Santander said to have hired RBS trio for corporate FX sales

Topic 8	quarter return fund target plan asset equity date bond year fixed allocation income average participant tdfs increase maturity performance flow gso gain median commodity outperform aon contribution end hewitt class
100%	Target Maturity Funds Bounce Back at End of Year
100%	Target Maturity Funds Bounce Back at End of Year
100%	April Sees Decreased Funding for Corporate DB Plans
100%	July Signals Positive Trend for Pension Plan Sponsors
100%	July Signals Positive Trend for Pension Plan Sponsors
100%	An April Drop in Corporate DB Funding
99%	Target-Date Funds Up in 3Q
99%	Corporate Pensions Funding Dips Further in July

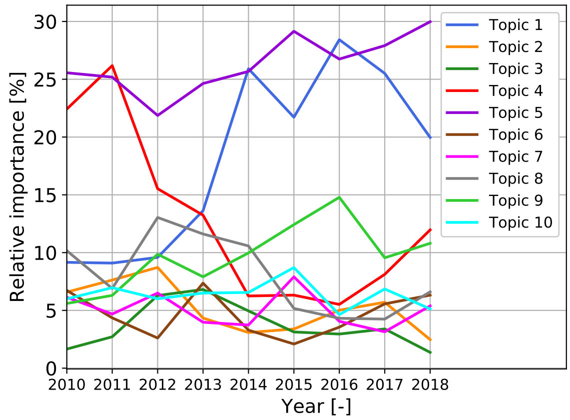
Topic 9	manager investor think active use risk say investment portfolio make research time way firm lot strategy different want like need look money good company client thing know try people factor
97%	Book Excerpt Charles Ellis and the Index Revolution
91%	Dimensional Fund Advisors Grapples With Its Future
91%	Alternative beta strategies can enhance HF allocations
87%	JOBS Act May Employ the Unscrupulous
87%	Alternative beta strategies can enhance HF allocations
86%	Smart questions
86%	Four asset managers 'may have broken UK competition law' [updated]
85%	Surge in platform options opens up investor choice

Topic 10	china private fund billion equity hedge hong kong chinese asset market million say percent capital emerge firm year management asia partner deal debt swiss investor mandate bram pension shanghai investment
90%	Swiss pension fund tenders $10m factoring mandate using IPE Quest
90%	China Set to Become Net Exporter of Capital
89%	Pension fund seeks Swiss smallmid cap managers via IPE Quest
88%	Swiss pension fund tenders small-cap mandate worth up to $900m
87%	Private Equity Wire Global Awards 2017 - The winners
86%	Squadron Launches $150M Asia PE Fund
84%	Switzerland's AHV tenders CHF500m global equity portfolio
80%	Abraaj Buys Amundi North Africa PE Platform

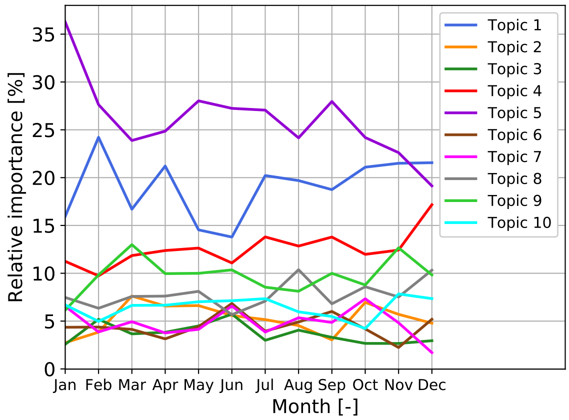
Figure D.4: Top 30 words and top 8 articles for the topics in configuration 4 (cont.).



(a) Absolute importance.



(b) Relative importance.



(c) Relative (average) importance per month.

Topic 1	Equity market (economy)
Topic 2	Trading
Topic 3	Pension
Topic 4	Indexes and ETFs
Topic 5	Fund management and fund launches
Topic 6	Retirement plan products
Topic 7	Banking, analyst research
Topic 8	Performance
Topic 9	Investment strategy
Topic 10	Asia, private equity, searches for asset managers via IPE Quest

(d) Topic labels.

Figure D.5: Importance of 10 topics per year or per month.

Magazine	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Euromoney	16%	19%	3%	1%	3%	1%	35%	4%	9%	10%
Institutional Investor	16%	11%	12%	7%	9%	1%	9%	3%	21%	12%
Financial Adviser	59%	2%	1%	2%	20%	1%	1%	2%	8%	3%
AlphaQ	33%	0%	1%	8%	32%	4%	0%	1%	15%	7%
Institutional Asset Manager	12%	12%	1%	12%	35%	3%	6%	3%	10%	7%
Wealth Adviser	15%	4%	1%	13%	50%	4%	2%	2%	5%	5%
Investment & Pension Europe	12%	4%	10%	3%	14%	1%	4%	8%	17%	28%
PlanSponsor	5%	3%	4%	26%	15%	11%	2%	23%	8%	3%
PlanAdviser	9%	3%	3%	25%	16%	10%	2%	22%	8%	2%

Table D.1: Topic coverage in each magazine.

D.5 Configuration 5

Topic 1	market sector year growth equity european company high stock rate investor rise economy price say yield cap return earnings dividend manager economic small europe strong ftse valuation term income fund
100%	Snapshot Europe shines but any setback could be fierce
100%	UK equity markets resilient in 2014
100%	Fund Selector Markets are finely balanced
100%	Fund Selector Stuck in a holding pattern
100%	Rising valuations stoke caution but sterling weakness to support sector
100%	Is the FTSE no longer benefiting from pound weakness
100%	Five macroeconomic factors driving European equities
100%	Neil Wilkinson waits on small caps

Topic 2	trading market ipo trade exchange trader russian company russia say volume firm bat stock order broker block new moscow listing commission brokerage liquidnet deal technology nasdaq execution share electronic research
100%	Flight Path for the SEC's Tick-Size Pilot
98%	Nasdaq and AX Trading Look at Block Trade Alternative To HFT
96%	BATS Tries to Reboot Its IPO
94%	May Day II, (Institutional Investor, February 1999)
93%	The Tick Size Pilot Key Trading Considerations
89%	JP Morgan starts trading in SLS
88%	BCS Global Markets completes first IPO
84%	How the ETF Market Quickly Got Over Its “Knightmare”

Topic 3	percent pension newsdash retirement employee activist say employer new worker state board endowment health school plan university cio benefit public pay kemna callan financial year retiree proxy wisconsin union care
100%	Hewlett-Packard the Latest to Bow to Shareholder Pressure
90%	Western Union Ups the Ante in Proxy Access Battles
88%	2011 NewsDash Archive List
84%	BlackRock CEO Mulls Retirement in Twitter Era
84%	BlackRock CEO Mulls Retirement in Twitter Era
81%	Wisconsin's Public Pension Works to Spread the Cheddar
75%	PSNC 2013 Up at Night
74%	PSNC 2013 Up at Night

Topic 4	index etf cap market msci russell billion emerge inflow weight small ishares large stock exposure factor spdr global volatility etfs beta month total outflow track category capitalization performance mid list
99%	SsgA Introduces Low Volatility ETFs
99%	SsgA Introduces Low Volatility ETFs
99%	MSCI Launches New Indexes for Developed Markets
99%	ProShares Launches Daily 3x and -3x ETFs
99%	MSCI Unveils Micro Cap Indices
98%	ETFs Increase by $12B in November
98%	ETFs Increase by $12B in November
97%	ETFs Increase by $26B in October

Topic 5	fund investment management team portfolio manager cap small manage strategy equity asset company launch growth value global join capital invest investor new long focus client experience mutual firm director provide
100%	Pegasus UCITS Fund restructures and rebrands as Tosca Micro Cap UCITS Fund
100%	Sentry Investments and Sun Life Global Investments expand partnership with three new funds
100%	Volantis moves to Lombard Odier Investment Managers
100%	Elessar Investment Management team joins Emerald Advisers
100%	Cove Street Capital launches Value Strategies
100%	Pzena Investment Management enters retail market with expanded leadership team
100%	VAM Funds enters South African market
100%	Bridgehouse Asset Managers launches in Canada

Figure D.6: Top 30 words and top 8 articles for the topics in configuration 5.

Topic 6	plan fund fee vanguard share sponsor retirement participant expense investment class option hancock offer fiduciary cost adviser john mutual fidelity plaintiff nextpage complaint tax charge esg ratio defendant available include
100%	Kraft Suit Plaintiffs Denied Class Status on Remaining Claim
100%	John Hancock Establishes New Series of R Share Classes
99%	John Hancock mutual funds launches new R6 share class
99%	John Hancock Establishes New Series of R Share Classes
94%	ING U.S. Unveils R6 Shares
92%	ING U.S. Unveils R6 Shares
91%	Participant Challenges Prudential and Morningstar Allocation Solution
91%	Neuberger Berman introduces retirement share class for seven mutual funds

Topic 7	bank banking loan finance business bnp lending paribas year client smes corporates lender credit euromoney capital analyst runner america billion trade germany say customer banco cash need european deutsche crisis
95%	Germany's Helaba to the Rescue
94%	2015 All-America Research Team How the Firms Fared
94%	2015 All-America Research Team Welcomes 30 Newcomers
90%	2015 All-America Research Team Key Facts and Figures
87%	Awards for Excellence 2016 Fine-tuned BNP Paribas excels at the business of banking
86%	Trade finance survey 2010 In world trade, banks turn out not to be the villains
86%	Italy ECB's first merger brings more worry
86%	Santander said to have hired RBS trio for corporate FX sales

Topic 8	return quarter fund equity asset target plan allocation bond year date average fixed performance income increase tdfs maturity gain end median outperform participant high passive liability active large class period
100%	Target Maturity Funds Bounce Back at End of Year
100%	November Sees Increase for Corporate Pension Funding
100%	Corporate Pension Funding Up in November
100%	April Sees Decreased Funding for Corporate DB Plans
100%	July Signals Positive Trend for Pension Plan Sponsors
100%	July Signals Positive Trend for Pension Plan Sponsors
100%	An April Drop in Corporate DB Funding
100%	DB Plan Liabilities Declined in June

Topic 9	manager think investor active use risk portfolio investment research say make time way lot different strategy firm look want thing good try know like factor need idea money people don
92%	Book Excerpt Charles Ellis and the Index Revolution
89%	Alternative beta strategies can enhance HF allocations
86%	Dimensional Fund Advisors Grapples With Its Future
84%	Smart questions
84%	Alternative beta strategies can enhance HF allocations
81%	Four asset managers 'may have broken UK competition law' [updated]
80%	Surge in platform options opens up investor choice
80%	UK Railways Pension Scheme takes big strides in risk-factor equities

Topic 10	china private fund hedge equity billion hong kong capital firm market say chinese asset percent year management million deal partner emerge asia gso debt investor investment mandate manager swiss bram
93%	China Set to Become Net Exporter of Capital
88%	Swiss pension fund tenders $10m factoring mandate using IPE Quest
88%	Private Equity Wire Global Awards 2017 - The winners
87%	Squadron Launches $150M Asia PE Fund
83%	Swiss pension fund tenders small-cap mandate worth up to $900m
81%	Pension fund seeks Swiss smallmid cap managers via IPE Quest
80%	Pension fund tenders Asia-Pacific, EM mandates using IPE Quest
80%	Abraaj Buys Amundi North Africa PE Platform

Figure D.6: Top 30 words and top 8 articles for the topics in configuration 5 (cont.).

D.6 Configuration 6

Topic 1	market sector year growth equity company high european investor rise stock rate economy say price cap yield dividend earnings manager return economic small ftse europe term strong valuation expect remain
100%	Snapshot Europe shines but any setback could be fierce
100%	Fund Selector Stuck in a holding pattern
100%	Rising valuations stoke caution but sterling weakness to support sector
100%	Is the FTSE no longer benefiting from pound weakness
100%	Five macroeconomic factors driving European equities
100%	Neil Wilkinson waits on small caps
99%	City Financial's Mark Harris warns over 'fragile nature' of US recovery
98%	UK equity markets resilient in 2014

Topic 2	etf market trading index exchange cap trade weight factor stock small volatility emerge beta exposure bat nasdaq nyse launch volume smart low capitalization investor vector amundi etfs new large trader
99%	IndexIQ launches first ETF to focus on emerging market mid-cap stocks
99%	IndexIQ Releases Emerging Markets Mid Cap ETF
99%	IndexIQ Releases Emerging Markets Mid Cap ETF
99%	Van Eck Launches ETF Offering Access to German Small-Caps
99%	Van Eck Launches German ETF
98%	Van Eck Launches Russia ETF
95%	iShares launches emerging markets small cap fund
93%	Market Vectors Launches New ETF

Topic 3	percent pension newsdash retirement activist worker employer employee new board state say health kemna cio endowment school wisconsin public walker benefit dutch proxy court chalkstream retiree proposal ppaca plan committee
86%	Hewlett-Packard the Latest to Bow to Shareholder Pressure
86%	Western Union Ups the Ante in Proxy Access Battles
79%	2011 NewsDash Archive List
78%	BlackRock CEO Mulls Retirement in Twitter Era
78%	BlackRock CEO Mulls Retirement in Twitter Era
74%	Wisconsin's Public Pension Works to Spread the Cheddar
67%	Optimism Is Growing That Abenomics Will Succeed in Japan
63%	Top Returns at Midsize Endowments Challenge the Yale Model

Topic 4	index esg russell msci sri wilshire dow jones environmental aon cap hewitt sustainable sustainability measure social acwi market company frontier represent asean nuance benchmark solactive governance clarington usa china seng
99%	MSCI launches 12 new China indexes
99%	Index Family Available to FactSet Clients
99%	MSCI Launches Overseas China Indices
99%	MSCI Launches Overseas China Indices
99%	MSCI Launches New Indexes for Developed Markets
99%	MSCI Unveils Micro Cap Indices
88%	FTSE Licenses Index for New Asian ETF
85%	MSCI renames South East Asia Indexes as MSCI ASEAN Indexes

Topic 5	fund investment management manager portfolio cap small strategy team equity manage asset company growth value launch global invest capital investor long new focus client mutual join provide experience income mid
100%	TA Associates backs buyout of Goldman Sachs Aussie investment platform
100%	Putnam Investments to launch suite of multi-cap equity funds
100%	TA Associates backs MBO of GSAM's Australian investment capabilities and fund platform
100%	Pegasus UCITS Fund restructures and rebrands as Tosca Micro Cap UCITS Fund
100%	Sentry Investments and Sun Life Global Investments expand partnership with three new funds
100%	Abhay Deshpande launches Centerstone Investors
100%	Volantis moves to Lombard Odier Investment Managers
100%	Putnam to Launch Multi-Cap Equity Funds

Figure D.7: Top 30 words and top 8 articles for the topics in configuration 6.

Topic 6	vanguard hancock john expense etf tax ratio index firearm fund explorer international ast timessquare dividend basis ing mcnabb transamerica municipal chf deltashares share usaa bii plurimi wellington biotech annuity brf
100%	Vanguard Adds Index Funds and ETFs
99%	Vanguard Introduces Index Funds and ETFs Based on S&P Benchmarks
92%	Vanguard launches International High Dividend Yield Index Fund and International Dividend Appreciation Index Fund
92%	Vanguard launches suite of Russell-based index funds and ETFs
91%	Vanguard to launch two dividend oriented funds and ETFs
84%	Vanguard Unveils Seven IndexETF Offerings
73%	Five Vanguard Index Funds transition to CRSP indices
73%	Vanguard reports third round of expense ratio reductions

Topic 7	bank loan finance italy germany lender banking european german lending spain italian helaba hungary trade eurobank debt greek billion smes piraeus greece france unicredit spanish cee landesbanks austria poland government
80%	Greek Banks Lure Foreign Investors Betting on a Turnaround
80%	Germany’s Helaba to the Rescue
71%	Hungary special report 2015 Good times are here again
60%	Deutsche Bank Repurchases ELEMENTS ETNs
59%	FX people moves DB sells into Nordics
55%	Three ELEMENTS ETNs Set For Redemption
54%	Truffle Capital appoints Olivier Streichenberger as listed securities manager
53%	French debt binge turns spotlight on buy-outs

Topic 8	return quarter equity asset allocation year bond fixed fund target average plan income performance gain real maturity rate end median outperform increase liability brazil estate period high class pension duration
100%	Target Maturity Fund Performance Climbs Back Up in 2009
100%	Target-Maturity Fund Performance Climbs Back Up in 2009
99%	Target-Date Funds Up in 3Q
99%	Corporate Pensions Funding Dips Further in July
99%	More Conservative Pension Allocation Fares Better
99%	Target-Date Funds Up in 3Q
99%	Target-Date Fund Returns Up in 3Q
97%	Target Maturity Funds Bounce Back at End of Year

Topic 9	investor manager portfolio think research active strategy firm say stock hedge use make investment return time like factor risk beta try way idea market good percent long thing smart money
96%	Book Excerpt Charles Ellis and the Index Revolution
91%	Four asset managers 'may have broken UK competition law' [updated]
90%	Fund selector Behind behavioural finance
90%	Dimensional Fund Advisors Grapples With Its Future
89%	Alternative beta strategies can enhance HF allocations
87%	Alternative beta strategies can enhance HF allocations
85%	Smart questions
85%	Mifid II ruffles fund research practices

Topic 10	china private market deal ipo billion hong kong say capital russian chinese million raise firm company percent russia fund gso year investor moscow asia shanghai bram equity hedge emerge partner
89%	China’s IPO Flurry
87%	Russia equity markets back in business as IPO trio find demand
85%	Private Equity Wire Global Awards 2017 - The winners
82%	Blackstone Group’s GSO Capital Lenders of Last Resort
81%	Russia Contrasting fortunes for Russian share deals
78%	RussiaCIS Russian IPOs wait for a propitious 2011
77%	An Upstart Start-up Takes on Jim Cramer
77%	Brazilian companies move to streamline IPO leads

Figure D.7: Top 30 words and top 8 articles for the topics in configuration 6 (cont.).

Topic 11	bank client business need euromoney say bnp paribas want corporates customer liquidity technology make cash people banking work lot just way look service don big financial corporate credit relationship capital
100%	Cash management strategy debate Cash management in a world of risk and complexity
98%	Liquidity management debate Liquidity management in an age of anxiety
97%	Transaction services guide 2014 Corporate clients demand more
96%	Cash management debate Show me the money
95%	World's best bank for corporates BNP Paribas
94%	US regional banks BP's woes spill over into US banks
94%	BP's woes spill over into US banks
90%	Peer-to-peer FX providers pitch corporates

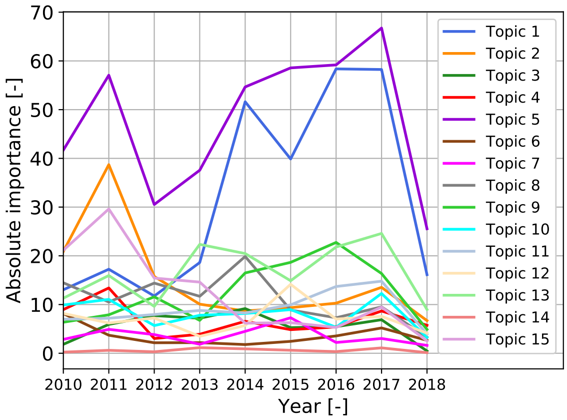
Topic 12	analyst firm research evercore morgan team america year runner merrill join liquidnet university merrin client lynch work independent goldman senior sale partner ubs hire service york degree stanley wall coverage
100%	J.P. Morgan's Joseph Greff Joins All-America Hall of Fame
100%	2015 All-America Research Team Key Facts and Figures
100%	2015 All-America Research Team The Top-Ranked Analysts
100%	2015 All-America Research Team Welcomes 30 Newcomers
98%	2015 All-America Research Team How the Firms Fared
81%	Matrix hires UK real estate team from JP Morgan Cazenove
80%	National Bank Names Energy Analyst
75%	Bank of America Merrill Lynch Leads 2016 All-Europe Sales Team

Topic 13	plan fund fee participant sponsor retirement investment active share option class target date fiduciary passive asset fidelity adviser use nextpage contribution say cost offer expense define mutual plaintiff complaint charge
100%	Chevron Wins Dismissal of ERISA Challenge
100%	Charts and Graphs Are Good, but Don't Change Target-Date Fund Names
100%	Court Buys Retail vs. Institutional Share Fee Claims
100%	The Investment Menu Trends Sponsors Are Talking About
100%	PSNC 2016 DC Plan Investment Menu Trends
100%	White Labeling DC Plan Investments May Offer Advantages
100%	Morgan Stanley Facing Excessive-Fee, Self-Dealing Lawsuit
98%	Self-Dealing Suit Challenges Fees and Fund Monitoring

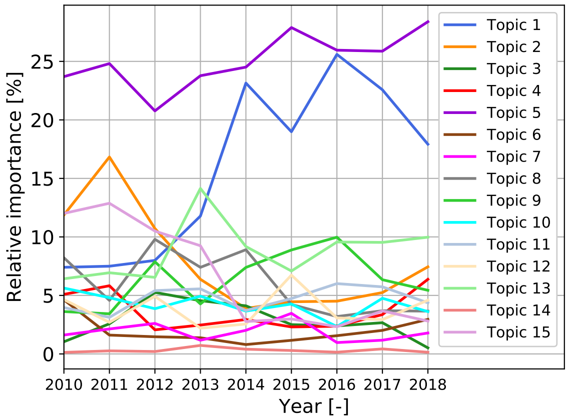
Topic 14	alphadex franklin bullishness bissett goalmaker templeton ifunds rollins ave maria ifas dashboard schneider family ibillionaire redwood albion fma polley factsheet kempen guinther tapestry mers ibln imo camelot ifsl billionaire stewart
28%	Franklin Templeton to increase fee transparency for Canadian mutual fund investors
27%	LTA cut may spur interest in Venture Capital Trusts
25%	Ave Maria Mutual Funds surpasses USD1bn AUM
23%	Family firms likely to outperform the market
22%	Franklin Templeton proposes changes for two Bissett Balanced Fund mandates
22%	ETF Tracking Billionaire Buys Genius or Sucker's Play
22%	UK economic recovery strengthens IFA support for small business investing
17%	Kempen Capital Management new head of family office business

Topic 15	billion inflow etf outflow month million flow ishares net category cap gold saw respectively market spdr asset commodity bond large decrease spy aggregate report etps june qqq total morningstar positive
100%	ETFs See Inflows of $12B in April
100%	U.S. Stock Flows Fall in March
100%	U.S. Stock Fund Flows Fall Off in March
99%	ETF Assets Increased $50B in July
99%	ETFs Pull In More Than $42B in July
99%	ETFs Increase by $12B in November
99%	ETFs Increase by $12B in November
99%	ETFs Increase $32B in September

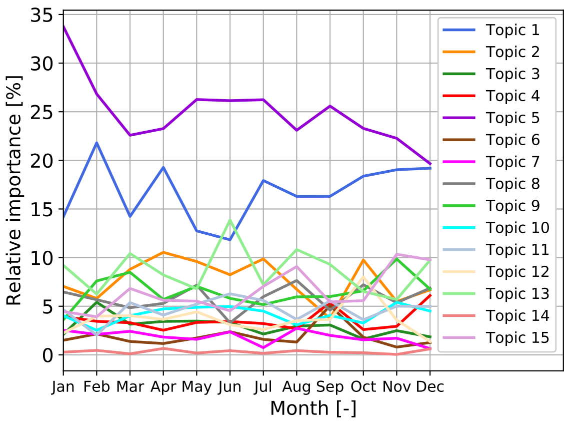
Figure D.7: Top 30 words and top 8 articles for the topics in configuration 6 (cont.).



(a) Absolute importance.



(b) Relative importance.



(c) Relative (average) importance per month.

Topic 1	Equity market (economy)
Topic 2	ETF launches
Topic 3	Pension
Topic 4	Indexes and ethical investing
Topic 5	Fund management and fund launches
Topic 6	Vanguard and John Hancock
Topic 7	European banking
Topic 8	Performance
Topic 9	Investment strategy
Topic 10	Emerging markets, IPO, private equity
Topic 11	Corporate banking
Topic 12	Analyst research
Topic 13	Retirement plan products
Topic 14	Articles with an exclusive frequent word
Topic 15	Fund flows

(d) Topic labels.

Figure D.8: Importance of 15 topics per year or per month.

Magazine	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
Euromoney	13%	4%	1%	1%	3%	1%	11%	4%	5%	18%	32%	5%	0%	0%	2%
Institutional Investor	14%	8%	11%	1%	9%	1%	3%	3%	19%	10%	6%	8%	4%	0%	3%
Financial Adviser	54%	1%	1%	1%	20%	0%	1%	2%	5%	2%	4%	2%	5%	0%	1%
AlphaQ	30%	1%	1%	11%	34%	4%	0%	4%	4%	5%	3%	0%	3%	0%	1%
Institutional Asset Manager	10%	7%	1%	9%	32%	2%	3%	3%	7%	5%	9%	6%	2%	0%	3%
Wealth Adviser	13%	10%	0%	3%	48%	3%	1%	2%	3%	3%	2%	4%	4%	0%	3%
Investment & Pension Europe	9%	3%	7%	1%	13%	0%	6%	12%	10%	4%	7%	2%	23%	0%	1%
PlanSponsor	4%	10%	2%	7%	14%	2%	0%	15%	3%	1%	2%	2%	20%	0%	17%
PlanAdviser	9%	9%	2%	5%	14%	2%	0%	12%	4%	0%	1%	1%	19%	0%	20%

Topic 1	Equity market (economy)
Topic 2	ETF launches
Topic 3	Pension
Topic 4	Indexes and ethical investing
Topic 5	Fund management and fund launches
Topic 6	Vanguard and John Hancock
Topic 7	European banking
Topic 8	Performance
Topic 9	Investment strategy
Topic 10	Emerging markets, IPO, private equity
Topic 11	Corporate banking
Topic 12	Analyst research
Topic 13	Retirement plan products
Topic 14	Articles with an exclusive frequent word
Topic 15	Fund flows

Table D.2: Topic coverage in each magazine.

Bibliography

Abner, D. J. (2016). *The ETF handbook. How to value and trade exchange-traded funds*. John Wiley & Sons, second edition.

AlphaQ (2015). Editorial of AlphaQ (April 2015). Accessed online at https://www.institutionalassetmanager.co.uk/sites/default/files/1504_AlphaQ_2.pdf on July 3rd, 2020.

Amenc, N., Deguest, R., Goltz, F., Lodh, A., and Martellini, L. (2014). Risk allocation, factor investing and smart beta: reconciling innovations in equity portfolio construction. Accessed online at https://risk.edhec.edu/sites/risk/files/pdf/RISKReview.2014-07-09.0733/attachments/EDHEC-Risk_Publication_Risk_Allocation_Factor_Investing_Smart_Beta.pdf on June 24th, 2020.

Amenc, N., Goltz, F., Le Sourd, V., and Lodh, A. (2015). Alternative equity beta investing: a survey. Accessed online at https://risk.edhec.edu/sites/risk/files/pdf/RISKReview.2015-03-26.2929/attachments/EDHEC_Publication_Alternative_Equity_Beta_Investing_Survey.pdf on July 9th, 2020.

Amenc, N. and Le Sourd, V. (2003). *Portfolio theory and performance analysis*. John Wiley & Sons.

Atkins, A., Niranjan, M., and Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2):120–137.

Aziz, S., Dowling, M. M., Hammami, H., and Piepenbrink, A. (2019). Machine learning in finance: a topic modeling approach. Accessed online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3327277 on July 16th, 2020.

Ball, C., Hoberg, G., and Maksimovic, V. (2015). Disclosure, business change and earnings quality. Accessed online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2260371 on July 18th, 2020.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9(1):3–18.

Bao, Y. and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, 60(6):1351–1616.

Barberis, N. and Shleifer, A. (2003). Style investing. *Journal of Financial Economics*, 68(2):161–199.

Bender, J., Briand, R., Melas, D., and Subramanian, R. A. (2013). Foundations of factor investing. Accessed online at https://www.msci.com/documents/1296102/1336482/Foundations_of_Factor_Investing.pdf/004e02ad-6f98-4730-90e0-ea14515ff3dc on June 24th, 2020.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

Blei, D. (2009). Topic models. Machine learning summer school (MLSS), Cambridge. Accessed online at http://videolectures.net/mlss09uk_blei_tm/ in June/July 2020.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2009). Topic models. In Srivastava, A. N. and Sahami, M., editors, *Text mining. Classification, clustering and applications*, pages 71–94. Chapman & Hall/CRC.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Bodie, Z., Kane, A., and Marcus, A. J. (2018). *Investments*. McGraw-Hill Education, eleventh edition.

Boyd-Graber, J. (2018). INST 414 - Advanced data science. University of Maryland. College of Information Studies. Accessed online at http://users.umiacs.umd.edu/~jbg/teaching/INST_414/ in June/July 2020.

Boyd-Graber, J., Hu, Y., and Mimno, D. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Brealey, R. A., Myers, S. C., and Allen, F. (2011). *Principles of corporate finance*. McGraw-Hill/Irwin, tenth edition.

Cao, C., Farnsworth, G., Liang, B., and Lo, A. W. (2017). Return smoothing, liquidity costs, and investor flows: evidence from a separate account platform. *Management science*, 63(7):2049–2395.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: how humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 288–296.

Clifford, C. P., Fulkerson, J. A., and Jordan, B. D. (2014). What drives ETF flows? *The Financial Review*, 49(3):619–642.

Coelho, L. P. and Richert, W. (2015). *Building machine learning systems with Python*. Packt Publishing, second edition.

Davis Evans, A. (2009). A requiem for the retail investor? *Virginia Law Review*, 95(4):1105–1129.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: evidence from latent Dirichlet allocation. *Journal of Accounting and Economics*, 64(2–3):221–245.

Elton, E. J., Gruber, M. J., Brown, S. J., and Goetzmann, W. N. (2014). *Modern portfolio theory and investment analysis*. John Wiley & Sons, ninth edition.

ETFGI (2017). ETFGI reports assets invested in smart beta equity ETFs/ETPs listed globally have increased 18.3% in 2017 to reach a new record of US$630 Bn at the end of August. Accessed online at https://etfgi.com/news/press-releases/2017/09/etfgi-reports-assets-invested-smart-beta-equity-etfsetps-listed on June 29th, 2020.

Euromoney (2020). About Euromoney. Accessed online at https://www.euromoney.com/about-us on July 3rd, 2020.

Euromoney Institutional Investor (2020). Who we are. Accessed online at https://www.euromoneyplc.com/about-us/who-we-are on July 3rd, 2020.

Fama, E. F. and French, K. R. (1992). The cross-section of expected returns. *The Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (1993). Common risk factors in the return on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fang, L. H., Peress, J., and Zheng, L. (2014). Does media coverage of stocks affect mutual funds' trading and performance. *The Review of Financial Studies*, 27(12):3441–3466.

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5).

Fellbaum, C. (1998). *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

Feuerriegel, S. and Pröllochs, N. (2018). Investor reaction to financial disclosures across topics: an application of latent Dirichlet allocation. *Decision Sciences*. Special issue available at https://doi.org/10.1111/deci.12346.

Feuerriegel, S., Ratku, A., and Neumann, D. (2016). Analysis of how underlying topics in financial news affect stock prices using latent Dirichlet allocation. In *49th Hawaii International Conference on System Sciences (HICSS)*, pages 1072–1081, Koloa, HI, USA. IEEE Computer Society.

French, K. (2020). Professor Kenneth French's data library. Accessed online at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html on July 7th, 2020.

FTAdviser (2020). About us. Accessed online at https://www.ftadviser.com/about-us/ on July 3rd, 2020.

Gelman, A., Carlin, J., Stern, H., Rubin, D., Dunson, D., and Vehtari, A. (2020). *Bayesian data analysis*. Electronic version available at http://www.stat.columbia.edu/~gelman/book/BDA3.pdf, third edition.

Ghayur, K., Heaney, R. G., and Platt, S. C. (2019). *Equity smart beta and factor investing for practitioners*. John Wiley & Sons.

Gillain, C. (2020). *Data analytics and financial market anomalies (provisional)*. PhD thesis, HEC-Ecole de gestion de l'Université de Liège. In progress.

Gillain, C., Ittoo, A., and Lambert, M. (2019). News-induced style seasonality. In *36th International Conference of the French Finance Association*, Québec, Canada.

Gillain, C., Ittoo, A., and Lambert, M. (2020a). Investment style coverage in institutional media. Private communication.

Gillain, C., Ittoo, A., and Lambert, M. (2020b). Style coverage in institutional media. In *12th Annual Hedge Fund Research Conference*, Paris, France.

Gillain, C. and Lambert, M. (2020). Size coverage in institutional media. Asset & risk management workshop.

Girolami, M. and Kabán, A. (2003). On an equivalence between PLSI and LDA. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*, pages 433–434.

Global Fund Media (2020). Our publications. Accessed online at https://www.ftadviser.com/about-us/ on July 3rd, 2020.

Goltz, F. and Le Sourd, V. (2015). Investor interest in and requirements for smart beta ETFs. Accessed online at https://risk.edhec.edu/sites/risk/files/pdf/RISKReview.2015-06-23.3723/attachments/EDHEC_Publication_Smart_Beta_ETFs_Survey.pdf on June 24th, 2020.

Grafe, P. (2010). Topic modeling in financial documents. Accessed online at http://cs229.stanford.edu/proj2010/Grafe-TopicModelingInFinancialDocuments.pdf on July 16th, 2020.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(suppl 1):5228–5235.

Hanley, K. W. and Hoberg, G. (2019). Dynamic interpretation of emerging risks in the financial sector. *The Review of Financial Studies*, 32(12):4543–4603.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hoberg, G. and Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure. *Journal of Corporate Finance*, 43:58–85.

Hoffman, M., Blei, D. M., and Bach, F. R. (2010). Online learning for latent Dirichlet allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 856–864. Curran Associates, Inc.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. (2018). Analyst information discovery and interpretation roles: a topic modeling approach. *Management Science*, 64(6):2473–2972.

Institutional Asset Manager (2020). About Institutional Asset Manager. Accessed online at https://www.institutionalassetmanager.co.uk/about/iam on July 3rd, 2020.

Institutional Investor (2020). About us. Accessed online at https://www.institutionalinvestor.com/about-us on July 3rd, 2020.

IPE International Publishers Ltd (2020). Company overview. Accessed online at https://www.ipe.com/company-overviews on July 3rd, 2020.

Israelsen, R. D. (2014). Tell it like it is: disclosed risk and factor portfolios. Accessed online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2504522 on June 24th, 2020.

ISS (2020a). About ISS. Accessed online at https://www.issgovernance.com/about/about-iss/ on July 3rd, 2020.

ISS (2020b). Media solutions. Accessed online at https://www.issgovernance.com/media on July 3rd, 2020.

Jenkinson, T., Jones, H., and Martinez, J. V. (2016). Picking winners? Investment consultants' recommendations of fund managers. *The Journal of Finance*, 71(5):2333–2370.

Jeynes, M. (2014). Fidelity replaces manager of European Opps fund. Accessed online at https://www.ftadviser.com/2014/06/30/investments/europe/fidelity-replaces-manager-of-european-opps-fund-OE2qDj0tjzHaO1HycNyJHM/article.html on July 13rd, 2020.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., and Ramakrishnan, N. (2013). Forex-foreteller: currency trend modeling using news articles. In *Proceedings of the 19th ACM SIGKDD international conference in knowledge discovery and data mining*, pages 1470–1473.

Jurafsky, D. and Manning, C. (2012). Natural language processing. Accessed online at https://www.youtube.com/playlist?list=PLoROMvodv4rOFZnDyrlW3-nI7tMLtmiJZ in June/July 2020.

Jurafsky, D. and Martin, J. H. (2019). *Speech and language processing*. Accessed online at https://web.stanford.edu/~jurafsky/slp3/ on June 24th, 2020, third edition.

Kaplan, S. and Vakili, K. (2015). The double-edge sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36(10):1435–1457.

Kula, G., Raab, M., and Stahn, S. (2017). *Beyond smart beta: index investment strategies for active portfolio management*. John Wiley & Sons.

Larsen, V. H. and Thorsrud, L. A. (2017). Asset returns, news topics, and media effects. Working paper of Norges Bank Research accessed online at https://static.norges-bank.no/contentassets/b19465af1c824d478493f22c4623f731/working_paper_17_17.pdf on July 18th, 2020.

Le Sourd, V. and Martellini, L. (2019). The EDHEC European ETF, smart beta and investing survey 2019. Accessed online at https://risk.edhec.edu/sites/risk/files/edhec_european_etf_smart_beta_and_factor_investing_survey_2019_.pdf on June 28th, 2020.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pages 535–541, Denver, CO, United States.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47(1):13–37.

Lo, A. (2008). 15.401 Finance theory I. Massachusetts Institute of Technology: MIT OpenCourseWare accessed online at https://www.youtube.com/playlist?list=PLUl4u3cNGP63B2lDhyKOsImI7FjCf6eDW in June/July 2020.

Lopez-Lira, A. (2019). Risk factors that matter: textual analysis of risk disclosures for the cross-section of returns. Accessed online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3313663 on June 24th, 2020.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: a survey. *Journal of Accounting Research*, 54(4):1187–1230.

Lynch, A., Puckett, A., and Yan, X. S. (2014). Institutions and the turn-of-the-year effect: evidence from actual institutional trades. *Journal of Banking & Finance*, 49:56–68.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance*, 42(3):483–510.

Morningstar (2002). Fact sheet: the new Morningstar style box methodology. Accessed online at http://news.morningstar.com/pdfs/FactSheet_StyleBox_Final.pdf on June 24th, 2020.

Morningstar (2019). A global guide to strategic-beta exchange-traded products. Accessed online at https://www.morningstar.com/content/dam/marketing/shared/pdfs/Research/A_Global_Guide_To_Strategic_Beta_Exhange-Traded_Products.pdf on June 24th, 2020.

Moro, S., Cortez, P., and Rita, P. (2015). Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, 42(3):1314–1324.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica*, 34(4):768–783.

Murphy, K. P. (2012). *Machine learning. A probabilistic perspective*. The MIT Press.

Oliphant, T. E. (2006). *A guide to NumPy*. Trelgol Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information system*, 14(3):130–137.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Reilly, F. K. and Brown, K. C. (2011). *Investment analysis & portfolio management*. South-Western, Cengage Learning, tenth edition.

Rennie, J. (2008). 20 newsgroups. Accessed online at http://qwone.com/~jason/20Newsgroups/ on July 22, 2020.

Resnik, P. and Hardisty, E. (2010). Gibbs sampling for the uninitiated. Accessed online at http://hdl.handle.net/1903/10058 on July 2nd, 2020.

Richardson, L. (2020). Beautiful soup documentation. Accessed online at https://beautiful-soup-4.readthedocs.io/ on July 4th, 2020.

Riding, S. (2018). Smart beta moves into mainstream for large investors. *Financial Times*. Accessed online at https://www.ft.com/content/18497f00-c32b-11e8-84cd-9e601db069b8 on June 28th, 2020.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360.

Russell FTSE (2019). Smart beta: 2019 global survey findings from asset owners. Accessed online at https://www.ftserussell.com/index/spotlight/smart-beta-factor-indexes/smart-beta-survey on July 4th, 2020.

Russell Investments (2014). Smart beta guidebook. An overview of Russell smart beta indexes. Accessed online at https://russellinvestments.com/-/media/files/nz/insights/smart-beta-guidebook.pdf on July 9th, 2020.

Scrapinghub (2020). Scrapy. Accessed online at https://scrapy.org/ on July 4th, 2020.

Serrano, L. (2020). Latent Dirichlet allocation and Training latent Dirichlet allocation: Gibbs sampling. Accessed online at https://www.youtube.com/watch?v=T05t-SqKArY and https://www.youtube.com/watch?v=BaM1uiCpj_E in June/July 2020.

Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Sharpe, W. F. (1992). Asset allocation. Management style and performance measurement. *The Journal of Portfolio Management*, 18(2):7–19.

Sievert, C. and Shirley, K. E. (2014). LDAvis: a method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistic.

Sikes, S. A. (2014). The turn-of-the-year effect and tax-loss-selling by institutional investors. *Journal of Accounting and Economics*, 57(1):22–42.

Sirri, E. R. and Tufano, P. (1998). Costly search and mutual fund flows. *The Journal of Finance*, 53(5):1589–1622.

Small, M. (2018). Mind your P's and F's: don't confuse leveraged ETPs with ETFs. Accessed online at https://www.blackrockblog.com/2018/02/07/leveraged-etps-vs-etfs/ on July 4th, 2020.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, *Latent semantic analysis: a road to meaning*, pages 427–448. Lawrence Erlbaum.

Strang, G. (2016). *Introduction to linear algebra*. Wellesley-Cambridge Press.

Thomas, R. (2019). A code-first introduction to natural language processing. Accessed online at https://www.fast.ai/2019/07/08/fastai-nlp/ in June/July 2020.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies*, 25(2):65–86.

Treynor, J. L. (1962). Toward a theory of market value and risky assets. Accessed online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=628187 on July 3rd, 2020.

Vernimmen, P., Quiry, P., Dallocchio, M., Le Fur, Y., and Salvi, A. (2018). *Corporte finance. Theory and practice*. John Wiley & Sons, fifth edition.

Wealth Adviser (2020). About Wealth Adviser. Accessed online at https://www.wealthadviser.co/about/wa on July 3rd, 2020.

Zhai, C. (2016). Text mining and analytics. Coursera course accessed online at https://www.youtube.com/playlist?list=PLLssT5z_DsK8Xwnh_0bjN4KNT81bekvtt in June/July 2020.

Executive summary

Smart beta exchange-traded funds (ETFs) are increasingly popular investment products among institutional investors. These ETFs can be categorized into different styles depending on the systematic risk factors to which they provide exposure. Hence, the question arises whether certain topics within the news coverage of specific styles influence the investment decision and thereby fund flows towards respective smart beta ETFs. This thesis focuses on partially answering this question by identifying the major topics in investment style news and their importance measured by their frequency of occurrence.

Based on a review of topic models, which are machine learning methods to discover topics in large collections of documents, latent Dirichlet allocation (LDA) is selected to identify the topics in investment style news. Moreover, the *most extensive literature survey of LDA in finance* (to the best of our knowledge) is compiled in order to optimally apply this method.

Subsequently, the major topics in a *unique corpus, which has never before been investigated by topic models* (to the best of our knowledge), are identified by LDA. This corpus consists of 1720 articles related to small-cap investing from 9 magazines targeting institutional investors.

The 5 major topics are “equity market (economy)”, “analyst research, trading and banking”, “retirement planing”, “indexes, ETFs and performance” and “fund management and fund launches”. These topics either persist, disappear or specialize when the number of topics to identify is increased. Dominant topics of individual magazines correspond to those proposed by the corpus specialist and the short descriptions of the magazines. The dominant topic over time is “fund management and fund launches”, which follows a seasonal trend characterized by lower coverage at the end of the year and higher coverage in January, thus suggesting that changes of fund management and fund launches preferentially occur at the beginning of the year.

Since the topic proportions of each article are identified, the correlation between the importance of topics over time and corresponding fund flows can be studied in future research.

Keywords: style investing, news coverage, topic modeling, latent Dirichlet allocation