Lenz, David; Winker, Peter

Working Paper

Measuring the diffusion of innovations with paragraph vector topic models

MAGKS Joint Discussion Paper Series in Economics, No. 15-2018

Provided in Cooperation with:
Faculty of Business Administration and Economics, University of Marburg

Suggested Citation: Lenz, David; Winker, Peter (2018) : Measuring the diffusion of innovations with paragraph vector topic models, MAGKS Joint Discussion Paper Series in Economics, No. 15-2018, Philipps-University Marburg, School of Business and Economics, Marburg

This Version is available at:
https://hdl.handle.net/10419/200671

Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 15-2018

David Lenz and Peter Winker

Measuring the Diffusion of Innovations with Paragraph Vector Topic Models

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Measuring the Diffusion of Innovations with Paragraph Vector Topic Models[1]

David Lenz
Justus-Liebig-University Giessen
Licher Strasse 64
35394 Giessen, Germany
email: david.lenz@wirtschaft.uni-giessen.de

Peter Winker
Justus-Liebig-University Giessen
Licher Strasse 64
35394 Giessen, Germany
email: peter.winker@wirtschaft.uni-giessen.de

May 16, 2018

Abstract

Topic modeling became an intensively researched area in economics, mainly due to the ever increasing availability of huge digital text information and the improvements in methods to analyze these datasets. In natural language processing, topic modeling describes a set of methods to extract the latent topics from a collection of documents. Several new methods have recently been proposed to improve the topic generation process. However, examination of the generated topics is still mostly based on unsatisfactory practices, for example by looking only at the list of most frequent words for a topic. Our contribution is threefold: 1) We present a topic modeling approach based on neural embeddings and Gaussian mixture modeling, which is shown to generate coherent and meaningful topics. 2) We propose a novel "topic report" based on dimensionality reduction techniques and model generated document vector features which helps to easily identify topics and significantly reduces the required mental overhead. 3) Lastly, we demonstrate on a technology related newsticker corpus how our approach could be used by economists to tackle economic problems, for example to measure the diffusion of innovations.

[1] Funding of the project from BMBF (16IFI002) is gratefully acknowledged. We are indebted to the Heise Medien GmbH & Co. KG for providing access to the their archives and the permission to apply text mining methods on these data. We would like to thank Dr. Marc Strickert for insightful discussions, as well as for his support during the implementation of the Barnes-Hut t-SNE algorithm. We would also like to thank Viktoriia Naboka for her valuable research support.

1 Introduction

The amount of digital information available , e.g., through the world wide web, is growing rapidly. This provides new data sources for economic analysis, e.g., with regard to identifying and measuring innovation trends. In order to exploit this valuable information, there is a growing need for automated information retrieval from large text corpora (Miner et al., 2012). Meanwhile, great progress has been made in machine learning and neural network theory that led to the emergence of new methods to extract high quality features from text. Contrary to Varian (2014), who suggested that machine learning methods should be more widely known to and used by economists, very little research has been conducted in the fields of economics and econometrics that appropriately incorporates these new data sets and methods. The opportunities created by the ongoing digitization have not been properly acknowledged yet nor have they been extensively studied in the economic literature. For a few early exceptions see the economic use cases discussed in Section 2.

In natural language processing (NLP), topic modeling describes a set of methods to extract the latent topics from a collection of documents. The results yielded by topic models are typically 1) a list of topics, where each topic is associated with a list of words that are especially relevant in the context of the topic, and 2) a document-topic matrix, where every document in the text corpus is assigned with a probability of belonging to each of the topics present in the corpus. For many years, Latent Dirichlet Allocation (LDA, Blei et al., 2003) has been the algorithm of choice for modeling latent topics in text corpora. However, LDA only describes the statistical relationships between words in the text corpus based on co-occurrence probabilities, which might not be the best feature representation for text (Niu and Dai, 2015). Furthermore, LDA reportedly has long computation times, especially with large text corpora (Ai et al., 2016). Aggravating, the interpretation of topics generated by LDA is not always straightforward and the necessary mental effort to give meaning to the extracted words can be tideous and demanding work (Baldwin et al., 2017).

We propose a topic model architecture based on neural embedding methods, which is able to generate meaningful and coherent topics. In particular, we use Paragraph Vector (Doc2Vec, Le and Mikolov (2014)) to construct vector space representations of text documents and Gaussian mixture models (GMM) to cluster the resulting document vectors. This combination of embedding and clustering methods is called Paragraph Vector Topic Modelling (PVTM). The presented neural embedding method yields comparable results to traditional topic models, the difference is in how the problem is approached.

Our proposed method has several advantages over traditional topic modeling. The unsupervised learning algorithm utilized to learn the document

embeddings automatically arranges documents according to their high level semantic information without the need of prior assumptions. Besides topic modeling, the representations provided by the neural embedding algorithms can efficiently be used in other NLP tasks, for example sentiment analysis or document retrieval (Dai et al., 2015).

The introduced combination of methods also allows to generate broader topic descriptions compared to standard topic modeling. For example, the learned vector representations can be mapped into human interpretable 2D or 3D space using the fast Barnes-Hut tSNE algorithm (BHtSNE, van der Maaten, 2013), which scales the algorithm proposed by van der Maaten and Hinton (2008) to large data sets. Consequently, we propose a novel topic report to represent the identified topics, which significantly reduces the mental overhead required during the topic interpretation phase.

We demonstrate the applicability of our approach on a corpus of news articles from the German IT-publisher *heise news* from the last 20 years. Our results suggest that PVTMs are particularly well suited for topic modeling of this type of text data.

The remainder of this article is organized as follows. Section 2 acknowledges recent work in the fields of topic modeling architectures, topic interpretability and applications to economic problems. Section 3 details the methodological background for our analysis. The news ticker dataset and the experimental design are reviewed in Section 4. In Section 5 we discuss our findings and describe avenues for future research.

2 Related work

This section reviews relevant work in the fields of neural topic model architecturs, topic interpretability and topic modeling in economics.

Neural topic models

Neural topic models, which combine topic modeling with neural embedding methods, became an intensively researched topic lately. In many approaches the topics generated by LDA are combined with neural embeddings of words and/or documents, which has been shown to benefit and improve both sides (Shi et al., 2017). Liu et al. (2015) combine LDA generated topics with word embeddings to learn word embeddings for all possible word-topic combinations, which alleviates the common problem of words with different meanings. Topic2Vec (Niu and Dai, 2015) embeds LDA generated topics into a neural word embedding space and constructs the topic word list based on the cosine similarity between topics and words to improve the descriptiveness of the topic word list. Language models based on recurrent neural networks (RNN) have been studied as well (Tian et al., 2016; Dieng et al., 2016; Palangi et al., 2016). The methodologically most closely related approach

to ours is from Hashimoto et al. (2016), who identify relevant articles for systematic reviews using Doc2Vec to construct vector representations of documents and then cluster the document vectors using k-means. The resulting cluster centroids are interpreted as latent topics and topic probabilities for the documents are constructed using the distance between cluster centroids and document vectors. In comparison, while also relying on Doc2Vec to construct document embeddings, we use GMM soft-clustering to directly assess topic memberships.

Topic Interpretability

The reduction of the cognitive overhead during topic interpretation is an actively researched field. Early labeling approaches (Mei et al., 2007) minimized the Kullback-Leibler divergence (Kullback and Leibler, 1951) between word distributions of topic and label words, while other approaches label topics by summarization (Basave et al., 2014). Lau et al. (2011) demonstrated an automatic topic labeling method for topics generated by LDA models by querying wikipedia articles for the top terms in a topic, which has been extended by Bhatia et al. (2016) to combine embeddings of documents and words. Lau et al. (2017) combine sentence level language models with document context from topic models, which allows to generate topic sentences for easier topic interpretation.

Economic Topic Modeling

Hisano et al. (2013) use topic models to extract stock related topics from financial news, which are then used to predict abnormal returns. Mizuno et al. (2017) measure the novelty of financial news using topic models. Hansen et al. (2014) analyse how monetary policy is affected by increased transparency. Larsen and Thorsrud (2015) utilize the topics found in a Norwegian business newspaper to model the impact of news on the business cycle. Lüdering and Winker (2016) apply LDA to articles in the Journal of Economics and Statistics to study whether or not the scientific discussion of topics correlates with the actual development of economic key indicators. Wehrheim (2017) employs LDA to model the topics in the Journal of Economic History (JEH) between 1941 and 2016. Stathoulopoulos (2017) use doc2vec to learn vector representations of textual information for UK based companies to improve upon the Standard Industrial Classification (SIC) codes to classify companies into sectors.

3 Methodology

The following section introduces the PVTM methodology to generate topics and topic memberships for documents. As Doc2Vec borrows the main idea

from Word2Vec it is useful to discuss the Word2Vec mechanics for encoding single words before detailing the Doc2Vec method and the document clustering algorithm.

3.1 Neural Embeddings of Words and Documents

Neural network based embedding methods like Word2Vec (Mikolov et al., 2013a,b) play an increasingly vital role for encoding the semantic and syntactic meaning of words. Because similar words tend to appear in similar contexts (Harris, 1954), encoding words based on their local context captures interesting properties in the resulting vectors, which have been shown to represent the way in which we use these words. Intuitively, words that share many contexts are more similar than words that share fewer contexts. Word2Vec builds low-dimensional dense vector space representations which encode the syntactic and semantic meaning of a word in a given context. Doing vector calculations on the resulting word vectors yields interesting results, for example $v_{Paris} - v_{France} + v_{Italy} = v_{Rome}$ (Mikolov et al., 2013a), where v_{word} is the learned word vector for that word. These meaningful vector space representations can be used for a variety of NLP tasks, including topic modeling. Word2Vec comes in two architectural variants: Skip-Gram (SG) and Continuous Bag of Words. We discuss the SG architecture in more detail, as this is relevant for our employment of the Doc2Vec model.

Skip-Gram

During model training, the SG architecture iterates over a given text corpus in fixed-sized sliding windows and generates (*context — target*) word pairs. Assume the following 5-word window: *Innovation is good for business.* The context word w_c would be the middle word, *good*, and the target words $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ are *Innovation, is, for, business.* This results in 4 *(context — target)* pairs, *(good — Innovation)*, *(good — is)*, *(good — for)* and *(good — business)*. Given these word pairs, construction of the word embeddings is done as follows. Each word in the vocabulary is represented as a one-hot encoded sparse vector of size $V \times 1$, where V is the size of the vocabulary. In a one-hot encoding scheme, a 0 indicates the absence of a word while a 1 indicates the presence of a word. There are two weight matrices in the neural network model: the weight matrix $\mathbf{W}_i$ that connects the input and hidden layer and the weight matrix $\mathbf{W}_o$ that connects the hidden to the output layer[2].

[2] The number of neurons N in the hidden layer is determined by the desired dimensionality of the resulting word vectors N. Assuming we want to learn $N = 100$ dimensional word vectors from a text corpus with a vocabulary size of $V = 10.000$, the size of the hidden layer has to be set to $n = 100$ neurons. As a result, the weight matrix connecting input and hidden layer is of size $[10.000 \times 100]$, one row per word and one column per neuron.

Using as input the one-hot encoded representation of word w_c, $\mathbf{x}_c$, the hidden layer $\mathbf{h}$ is computed as

$$\mathbf{h} = \mathbf{W_i}^T \mathbf{x}_c = \mathbf{W_i}^T_{(k,\cdot)} := \mathbf{v}^T_{w_c} \tag{1}$$

where the superscript ^{T} indicates the transpose of a vector or matrix. The hidden layer uses a linear activation function, so the weights are passed unchanged to the output layer. Therefore, operation (1) basically extracts the k-th row from $\mathbf{W}_i$ to use it as a dense vector representation $\mathbf{v}^T_{w_c}$ for word w_c.

The weight matrix $\mathbf{W}_o$ from the hidden to the output layer is of size $N \times V$. Multiplying $\mathbf{W}_o$ with the one-hot representation of the target word results in a transposed word vector $\mathbf{v}^T_{w_t}$ of size $N \times 1$ which represents the target word.

$$\mathbf{W_o x}_t = \mathbf{W_o}_{(k,\cdot)} := \mathbf{v}^T_{w_t} \tag{2}$$

A softmax activation function (Bridle, 1990) is used in the output layer. The output of a single neuron in the output layer is the probability of target word w_t given the context word w_c, i.e.

$$p(w_t|w_c) = \frac{e^x}{\sum e^x}, \text{ with } x = \mathbf{v}^T_{w_c} \mathbf{v}^T_{w_t} \tag{3}$$

The n-dimensional word vector $\mathbf{v}^T_{w_i}$ is multiplied by the target word vector $\mathbf{v}^T_{w_o}$, afterwards the exponential-function is applied.

Given a sequence of training words $w_1, w_2, w_3, \cdots, w_T$, the objective of the word vector model is to maximize the average log probability, i.e.

$$\frac{1}{T} \sum_{t-k}^{T-k} \log p(w_t|w_{c1}, \cdots, w_{cT}) \tag{4}$$

During training on the text corpus, the backpropagation algorithm is used to iteratively update the weights in $\mathbf{W}_i$ and $\mathbf{W}_o$ until some convergence criteria is met. After training, the weights in $\mathbf{W}_i$ act as a lookup table for the word embeddings.

3.2 Paragraph Vector (Doc2Vec)

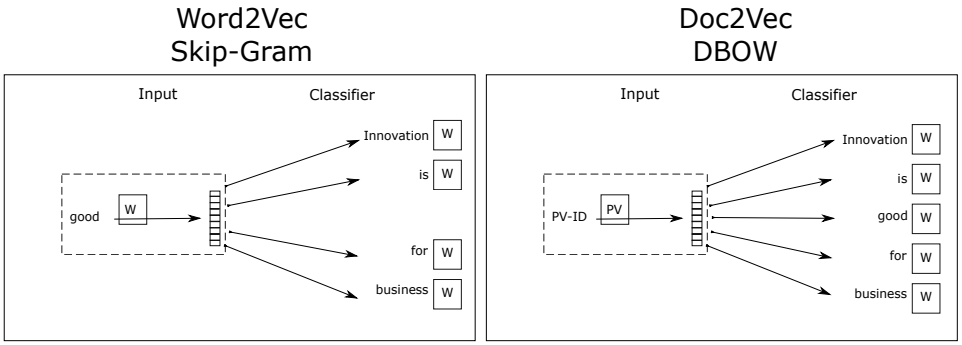
Paragraph Vector (PV or Doc2Vec, Le and Mikolov, 2014) describes a collection of methods to represent variable-length pieces of text as fixed sized, low dimensional but highly expressive dense feature vectors[3]. Doc2Vec expands the idea of word embeddings for longer pieces of texts. The document vectors learned during the training process capture latent document variables,

[3](The term "dense" is used to distinguish it from sparse representations such as Bag-of-Words representations of text.)

for instance the underlying semantic topic of a document. Doc2Vec also comes in two variants: Distributed Bag of Words (DBOW) and Distributed Memory (DM). In our experiments we focus on the DBOW methodology, as it has been shown to produce slightly better results compared to DM[4].

Doc2Vec-DBOW basically uses the Word2Vec-SG architecture but replaces the context word with a unique document vector. During training, the weights in the document vectors are learned in a similar fashion as the weights for words in the SG architecture. Figure 1 details the Doc2Vec-DBOW architecture in comparison to the Word2Vec-SG architecture.

Figure 1: Word2Vec Skip-Gram vs Doc2Vec DBOW



Notes: The SG architecture uses one-hot encoded word vectors as input to the neural network and the task is to predict the context words. With the Doc2Vec-DBOW architecture, the document vector of the current document is used as input and the task is to predict the context words.

3.3 Gaussian Mixture Clustering

LDA topic modeling uses clusters of important words to define topics, where different topics may share some words. In our approach, this is done by clustering the document vectors from Doc2Vec and finding the most relevant words per cluster. A Gaussian mixture model (GMM, see for example Reynolds, 2015) is a parametric probability density function represented as a weighted sum of Gaussian component densities (Sammut and Webb, 2017, p. 827). Gaussian Mixture Models employ the expectation maximization algorithm (EM, Dempster et al., 1977) to fit a mixture of Gaussian models to a given dataset and can be used to represent normally distributed subpopulations within an overall population.

GMMs have been used to track multiple objects in video sequences (Dadi et al., 2013), to extract features from speech data (Yu and Deng, 2014) and for speaker verification (Reynolds et al., 2000). Compared to frequently used clustering techniques such as k-means (Lloyd, 1982) or mean-shift (Fukunaga

[4]Though Le and Mikolov (2014) report that the DM architecture seems to perform better, subsequent research came to different conclusions (Lau and Baldwin, 2016).

and Hostetler, 1975), GMMs offer the advantage of soft-clustering the data. Soft clustering allows multiple cluster memberships per document, so each document can be represented as a probability distribution over the cluster memberships, which is quite similar to the results of LDA models. The result of the process is a matrix with one row per document and one column per identified cluster, where each entry represents the probability of belonging to a certain cluster. Given that Doc2Vec captures latent topics in the corpus, it is reasonable to suggest that clustering the resulting document vectors can be seen as identifying latent topics.

Particularly, given a D-dimensional document vector $\mathbf{x}$ and a pre-set number of Gaussian compoments M with mixture weights w_i, the Gaussian mixture model is defined as a weighted sum over the M Gaussian components, where the mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$:

$$P(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i g(\mathbf{x}|\mu_i, \Sigma_i) \tag{5}$$

Thereby, each mixture component $g(\mathbf{x}|\mu_i, \Sigma_i), \quad i = 1, \cdots, M$ is defined as a D-variate Gaussian function of the form

$$g(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_1^{-1} (\mathbf{x} - \mu_i) \right\} \tag{6}$$

with μ_i and Σ_i representing the mean vector and covariance matrix respectively. All component densities are collected in λ, which parameterizes the GMM by the mean vectors, covariance matrices and mixture weights, i.e.

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \quad i = 1, \cdots, M \tag{7}$$

The optimal parameter configuration λ is typically estimated through iteratively updating the model components to best fit the training data using the EM algorithm. Starting with an initial configuration λ, a new configuration $\bar{\lambda}$ is estimated such that $p(\mathbf{X}|\bar{\lambda}) \geq p(\mathbf{X}|\lambda)$. The initial configuration is computed using k-means. The mixture components are updated according to

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \lambda) \tag{8}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \lambda)\mathbf{x}_t}{\sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \lambda)} \tag{9}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \lambda)\mathbf{x}_t^2}{\sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \lambda)} - \bar{\mu}_i^2 \tag{10}$$

(8) is the update for weight w_i, the means are updated according to (9) and (10) details the variance re-estimation. σ_i^2, x_t and μ_i are elements of the

vectors $\sigma_i^2, \mathbf{x}_t$ and μ_i respectively. The a posteriori probability for component i is given by

$$p(i|\mathbf{x}_t, \lambda) = \frac{w_i g(\mathbf{x}_t \mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k g(\mathbf{x}_t \mu_k, \Sigma_k)} \tag{11}$$

The downside of GMMs is that one needs to specify the number of mixture components M beforehand, and the algorithm is always going to use all the components it has access to. This gives rise to the need of external validation methods. One way of evaluating the quality of a given GMM clustering is to use theoretical criteria like the Bayesian Information Criterion (BIC, Schwarz, 1978), which is the approach we are taking.

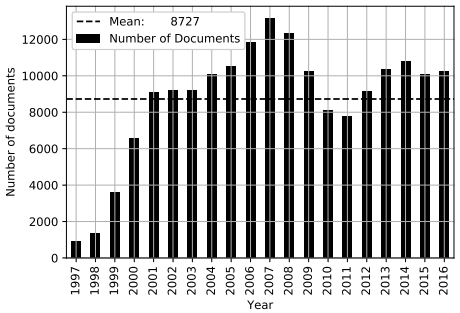
4 Application to news article corpus

4.1 Corpus

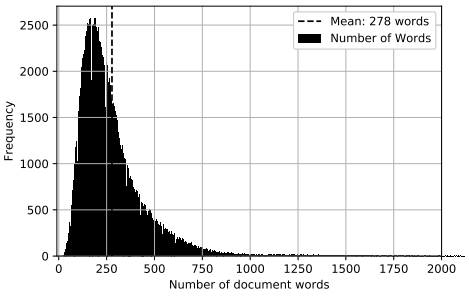
The dataset is formed by news articles from a single source, namely the newsticker archive from the German IT-magazine *Heise*[5] and consists of news articles in German language. Additionally, author and time information are available. The dataset dates back to 1997, covering a period of roughly 20 years. The total number of documents in the corpus is 174, 532 and the average number of documents per year is 8727, however the number of documents per year before the 2000s was considerably lower compared to subsequent periods. The average document consists of 278 words, while no document has less than 25 or more than 3919 words. Figure 2 details the number of documents per year and the number of words per document.

Figure 2: Text Corpus

(a) Corpus evolution over time



(b) Word frequency distribution



Given that the news are mostly IT related, we expect to mainly identify technology related topics. Using the available time information, we can

[5] `https://www.heise.de/newsticker/`

measure the relevance of topics over time, which we discuss in more detail during the topic report presentation.

4.2 Data Preprocessing & Parameter Settings

We removed all non-alphanumeric characters and lowercased the resulting words. Apart from this, no further preprocessing steps have been applied. Removal of stopwords is not done in advance but only before generation of the topic word lists.

Doc2Vec[6] is run for 10 epochs, where each epoch consists of going over all training examples once. The window size has been set to 5 words and the dimensionality of the resulting document vectors to 100. We used the BIC to find the optimal number of Gaussian mixture components and the best way to construct the covariance matrices Σ_i. For this, all possible combinations of K and Σ_i for $K \in \{10, 1000\}$ and $\Sigma_i \in \{diagonal, tied, full, spherical\}$ have been tested. As computing the GMM for all possible combinations takes quite some time, we used a two step procedure to find the optimal model parameters. We first iterated over the parameter space of K in steps of 50, i.e. $1, 51, 101$. The best result K^* was then used to construct a smaller search space $K \in \{K^* \pm 50\}$, which was searched in steps of 5. The optimal number of Gaussian mixture components K (=topics) was found to be 675 after this run which we kept as the final number of clusters. At every step during the parameter optimization procedure, all of the ways to construct the covariance matrices have been tested. The covariance matrices Σ_i of the GMM are constrained to be diagonal as this resulted in the lowest BIC scores for the dataset at hand.

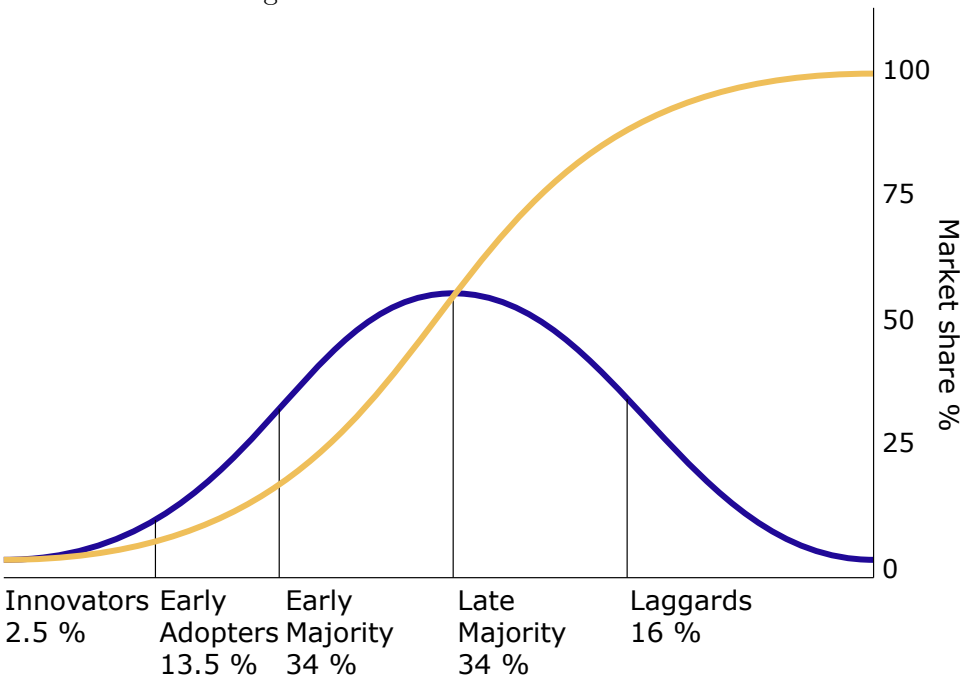
4.3 On the adoption of innovations

What is an innovation? Depending on the context and who you ask, innovation has different meanings. There is no final definition of what an innovation is (Baregheh et al., 2009). However, Rogers (2003) described innovation as an idea, practice, or object that is perceived as having new values by an individual or other unit of adoption. Given this formulation it becomes clear that not every topic is related to an innovation. Therefore, how can we identify the relevant innovation related topics without searching through all of them manually?

The innovation adoption curve is typically drawn as a hamp shaped line as shown in figure 3. Given a topics importance over time, we might be able to identify innovation topics by looking for similar patterns. When the relevance of a topic over time exhibits such a pattern, it might be innovation

[6]The gensim package (Řehůřek and Sojka, 2010) was used to train the Word2Vec and Doc2Vec models. To cluster the constructed document vectors we used the GMM implementation from Sklearn (Pedregosa et al., 2011).

Figure 3: The Diffusion of Innovations



related. Based on the time span the curve covers one could draw conclusions about the adoption time for this specific innovation. We present an example topic where this prototypical innovation curve is present in the next section. Future research could focus on developing methods to automatically identify innovation topics based on their importance timelines.

5 Results

5.1 Embedding visualization

Visualizing the similarity between documents and, on a higher level, between topics provides interesting insights into the structural relationships discovered by the model. To better understand the context of a given topic, dimensionality reduction techniques can be used to map the high-dimensional document vectors as well as the identified cluster centroids into human-interpretable space, i.e. 2D or 3D. t-distributed stochastic neighbor embedding (tSNE, van der Maaten and Hinton, 2008) is a dimensionality reduction technique that has been shown to produce significantly better visualizations than those provided by other techniques[7]. tSNE often preserves local structures, therefore it is useful for exploring local neighbor-

[7]Recently, LargeVis (Tang et al., 2016) has been shown to produce very good results on large datasets as well.

hoods and visually identifying clusters. To apply tSNE to large datasets, van der Maaten (2014) demonstrate how the runtime of tSNE can be improved from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, which they name Barnes-Hut tSNE and which is the method we use here. While interpreting tSNE maps one has to keep in mind that complicated relationships can be much more easily expressed in high-dimensional space compared to two dimensions and that some information about interactions between topics and documents is lost during dimensionality reduction.

We apply principal component analysis (PCA) to reduce the 100 dimensional document vectors to 50 dimensions and use Barnes-Hut tSNE to map from 50 into 2 dimensions. Figure 4 details the resulting 2-dimensional document embedding for all documents.

We annotated documents from 7 different topics with their document titles to demonstrate the relatedness of documents belonging to the same topic. The number of the topic to which the document has been associated is displayed in brackets behind the document title. Visual analysis of the cluster structures leads to the impression that the Doc2Vec algorithm extracts useful features from the documents, which the GMM algorithm utilized to cluster related documents into coherent topics.

5.2 Cluster Center

The topic vectors identified by the GMM can be used as an effective way to provide more context to a topic by mapping them into 2D. In Figure 5, we mapped the identified cluster centroids into 2D using tSNE. As this mapping is detached from the BHtSNE mapping in Figure 4 there is no resemblance between the points and clusters in the two images. Visual inspection allows to identify several cluster structures that could as well be clustered together, resulting in higher-level topics. We annotated some of the resulting high-level topic cluster by the most frequent word in the respective topic. The result of this process is shown in Figure 6. From the labels one may conclude that the topics in the identified high-level topics are related to one another.

5.3 Topic Report Style Sheet

The aim of the proposed topic style sheet is an improved and simplified interpretability of topics compared to traditional methods through reduced cognitive overhead. The topic report consists of one page per identified topic, where each page is composed of four components: 1) The top words represented as a wordcloud, 2) a timeline depicting the topic evolution over time, 3) textual article labels and 4) a tSNE map of the topic vectors. Example topic reports are shown in Figures 7 and 8.

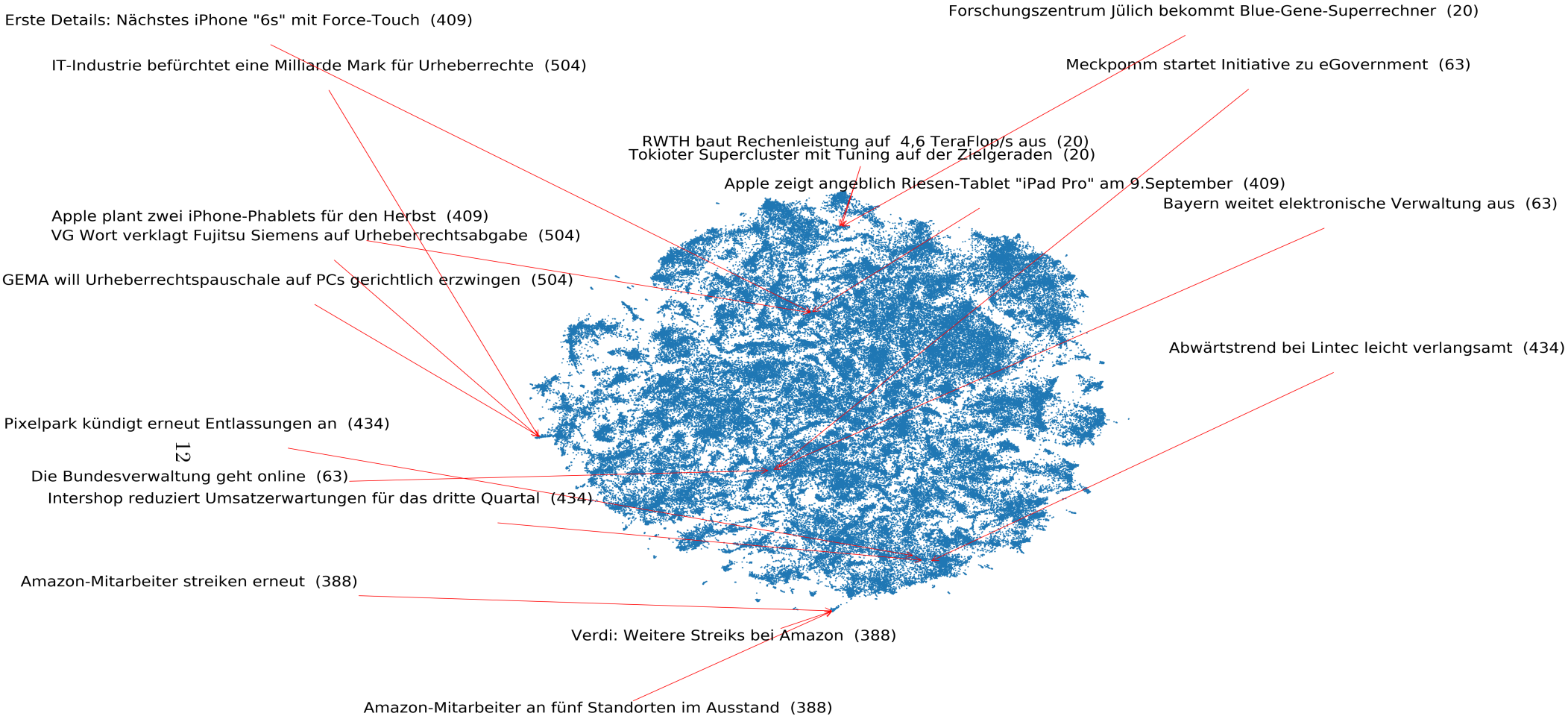
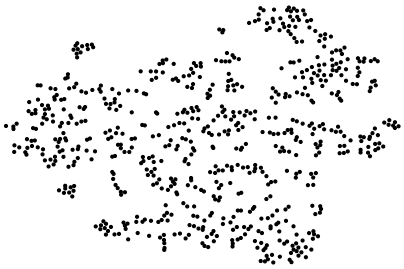


Figure 4: 2D Barnes-Hut tSNE map

Notes: BH-tSNE embedding of the 175k document vectors learned by the Doc2Vec architecture. We highlighted some documents belonging to 7 different topics with their article headers, three documents for each topic. The topic number is displayed in brackets after the title.

Figure 5: tSNE map of the identified cluster centroids



Top Words

In the upper left corner a wordcloud shows the most frequent words in the topic, with stopwords[8] removed. The font size of words correlates with their respective frequency.

Topic Evolution

Measurements of the topic importance over time are provided in the line plot in the lower left corner. We can monitor the diffusion of the identified topics by aggregating the topic probabilities per document per timestep. Topics that only become relevant after a certain point in time can be interpreted as representing something novel that does not fit into previous topics. Newly emerging topics might therefore be related to innovative processes, and we could capture the diffusion of innovations by measuring the relevance of the associated topic over time.

We differentiate two dimensions of importance, i.e., given a certain time interval, we quantify 1) the probability, that a topic appears in the text corpus and 2) the number of documents that are hard assigned to a topic. The hard assignment is done using a hard-vote mechanism where documents are allocated according to their highest membership probability, therefore in this setting each document can only belong to one topic T_i.

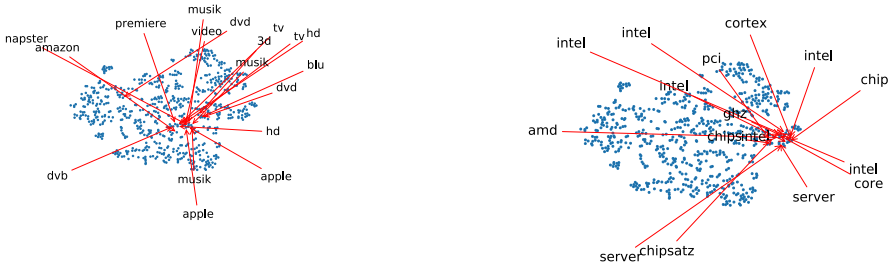
Given a text corpus at a certain time frame C_t, the probability for a single topic $p(T_i|C_t)$ is defined as the sum over the topic probabilities for all documents in the corpus in that time frame, divided by the number of

[8]The stopword-list is generated as follows. First, the stopwords from the python package `stop_words` (https://pypi.python.org/pypi/stop-words) are downloaded. Second, the stopwords from the python Natural Language Toolkit (NLTK, Loper and Bird, 2002) are added to the list. Third, we downloaded a list of common stopwords from https://github.com/6/stopwords-json. Lastly, we added some stopwords that we thought were still missing but relevant for our application. The complete procedure for generating the final list of stopwords is available on github: https://github.com/davidlenz/pvtm.

Figure 6: Example high-level topics identified using GMM clustering on the topic vectors

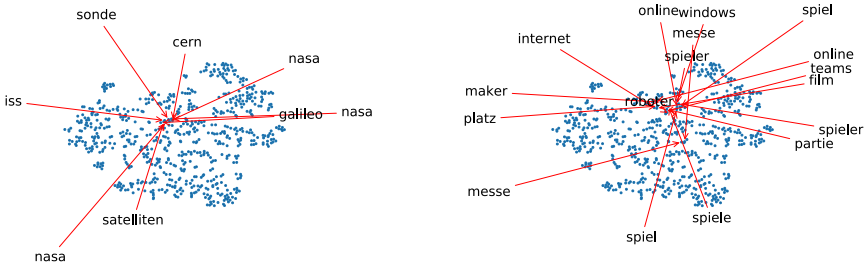
(a) High-level Topic: Multimedia

(b) High-level Topic: Processors



(c) High-level Topic: Space

(d) High-level Topic: Gaming



Notes: Each individual topic is annotated by the most frequent word in the topic.

documents D_t.

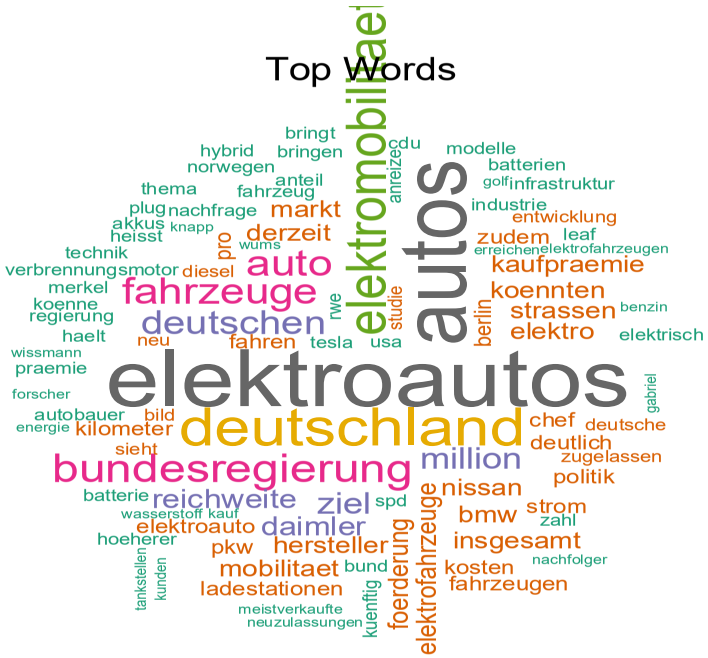
$$p(T_i|C_t) = \frac{\sum_{d \in C_t} p(T_i|d)}{D_t} \tag{12}$$

Generally speaking, if the two measurements are close together the hard-assigned documents tend to also have high probabilities for the respective topic. Differently moving lines on the other hand signalize that other topics also play an important part in the documents that are hard-assigned to the current topic. Depending on which measurement is relatively higher, the interpretation slightly changes. At every time frame, one of three possible events has to occur:

1. the lines of probability and number of documents match,
2. the probability is higher than the number of documents or
3. the probability is lower than the number of associated documents.

In case 1) we can assume that a topic is relatively closed in itself, meaning that if the topic occurs in a document it is likely that the topic is also the

Top Words



15

Topic Labels

1 - Fraunhofer-Studie: Eine Million Elektroautos 2020 möglich (representativeness: 0.833)

2 - Grüne: Spritschlucker-Fahrer sollten für E-Auto-Prämie zahlen (representativeness: 0.831)

3 - Bosch-Chef rechnet mit Rückgang der Akkukosten für E-Autos (representativeness: 0.825)

4 - Daimler: "Markt für E-Mobilität ist so weit" - neue Modelle kommen (representativeness: 0.818)

5 - Spritfresser und Diesel könnten E-Auto-Prämie finanzieren (representativeness: 0.816)

6 - Elektroautos: Was Subventionen in anderen Ländern bringen (representativeness: 0.815)

7 - Kaufprämien für E-Autos bleiben umstritten (representativeness: 0.813)

8 - Elektroautos: Regierung fordert Mitarbeit der Autoindustrie (representativeness: 0.812)

9 - Förderung der E-Mobilität: "Elektroautos brauchen mehr als nur Kaufprämie" (representativeness: 0.807)

10 - Porsche-Chef: Ladeinfrastruktur würde Elektroautos besser helfen als Kaufpr
ämie (representativeness: 0.805)

11 - Merkel: Entscheidung zu Elektroauto-Kaufprämie noch im April (representativeness: 0.805)

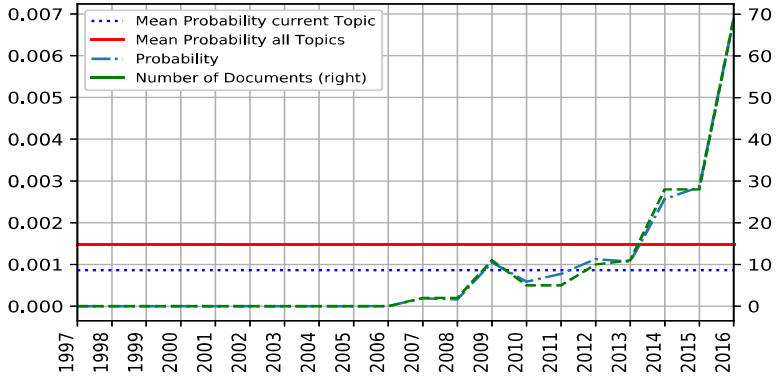
12 - Elektroautos: Bundesregierung führt 4000 Euro Kaufprämie ein (representativeness: 0.804)

13 - Öko-Institut und VDE sehen wirtschaftliche Kostenvorteile für gewerblich ge
nutzte Elektroautos (representativeness: 0.803)

14 - Elektromobilität: Bundesregierung will flächendeckendes Netz mit Schnelllad
esäulen (representativeness: 0.799)

15 - Bundesregierung lehnt Kaufprämie für Elektroautos weiter ab (representativeness: 0.798)

Topic Evolution



Topic Neighborhood

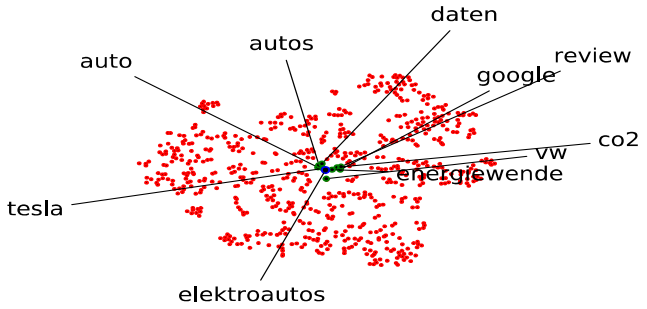


Figure 7: Topic 448

Top Words

16

Topic Labels

1 - Acer bringt 2011 Tablets mit 5, 7 und 10 Zoll (representativeness: 0.874)

2 - Hannspree zeigt 10-Zoll-Tablet mit Android (representativeness: 0.856)

3 - Das erste Tablet mit ARM-Prozessor und Windows RT (representativeness: 0.847)

4 - MSI stellt 12-Zoll-Subnotebook vor (representativeness: 0.845)

5 - Sony-Netbook kommt im August (representativeness: 0.829)

6 - Motorola Moto G: Live-Ticker zum billigen Android-Smartphone bei TechStage (representativeness: 0.828)

7 - Samsungs Jackentaschentablets (representativeness: 0.827)

8 - All-in-One-PCs mit Atom-Prozessor (representativeness: 0.822)

9 - Samsungs Windows-8-Tablets mit Tastatur und zusätzlichem Stift (representativeness: 0.822)

10 - Eee PCs mit Touch (representativeness: 0.818)

11 - Samsung hält ultramobilen Touchscreen-PCs die Treue (representativeness: 0.817)

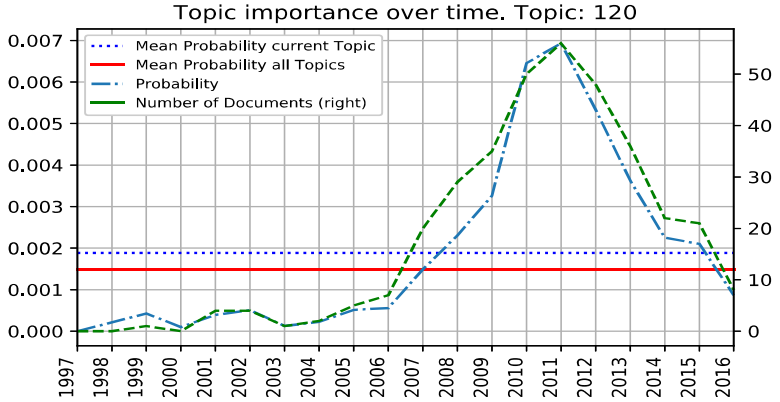
12 - Windows-Tablets: von 600 bis 1400 Euro (representativeness: 0.817)

13 - Tablet-Netbook-Hybrid mit Schiebetastatur und Edel-Subnotebook (representativeness: 0.815)

14 - Nvidia bestätigt Android-Tablet von Toshiba mit hoher Auflösung (representativeness: 0.815)

15 - Samsung zeigt Ultrabook und Windows/Android-Tablet mit ultrahoher Auflösung (representativeness: 0.813)

Topic Evolution



Topic Neighborhood

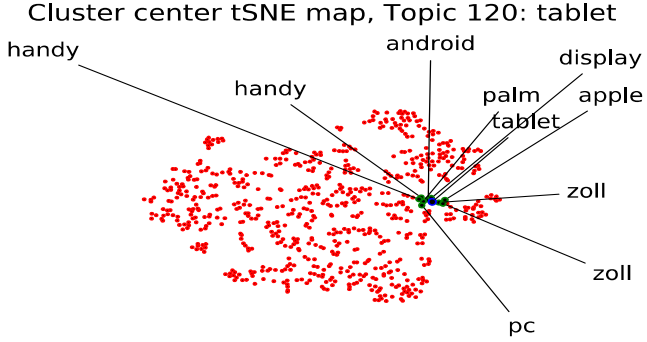


Figure 8: Topic 43

main topic in the document with a reasonably high probability. In case 2) a topic is more than average present in documents that are not primarily allocated to the topic, leading to a relatively higher probability of appearance compared to the absolute number of documents. A possible interpretation is that the topic is especially relevant in the context of other topics, for example when NVIDIA releases new GPU's (Topic: NVIDIA GPU) that are not mainly for gaming but deep learning (Topic: DEEP LEARNING) or crypto-currency mining (TOPIC: BITCOIN) instead. The NVIDIA GPU topic is relevant in both, the deep learning topic and the bitcoin mining topic, however it is not the main topic in the relevant documents, leading to a relatively higher probability than the total number of assigned topics.

Event 3) describes the other way around and allows the suggestion that even though a lot of documents were hard assigned to a topic it was not the sole topic of interest in the respective documents, however it was the most important topic. This could for example be the case for the deep learning topic, which also inherits the NVIDIA GPU topic but is still hard assigned to DEEP LEARNING, leading to a relatively higher number of hard assigned topics than the probability of occurrence.

The horizontal lines display the average probability of the topic in comparison to the average probability of all topics.

Textual Labels

As it has been shown that textual labels make topic interpretation easier for humans (Aletras et al., 2017), the 15 article headers of the most representative documents for a topic are provided in the upper right corner. Similarity is measured as the cosine similarity between cluster center and document vectors and is displayed as *representativeness*. Topics with a higher mean representativeness in the top headers might indicate tighter cluster structures since documents are closer to the cluster center. A possible interpretation is that the topic is less interfering with other topics, and documents from this topic tend to have a close focus on the topic, while topics with lower mean representativeness interact more with other topics. A promising alternative to the usage of article headlines could be using Wikipedia article headers as label candidates as demonstrated in Bhatia et al. (2016).

Neighboring Topics

We use the tSNE map of the cluster centroids to provide some local area context for a given topic through highlighting and labeling the 10 closest topics. The local context of a topic is found using a KDtree (Bentley, 1975) on the 2 dimensional tSNE map to perform nearest neighbor searches on the topics. Highlighting is done by means of using the most frequent word in a topic as the topic label.

Interpretation of example topic reports

We discuss two example topics. The full list of topics is available on github[9]

The topic in figure 7 is about electric cars. Some of the most important words are *electric-car* and *electro mobility*. The textual labels in form of document titles improve our understanding of the topic. *Electro* or *car* are words that appear in all article headlines, also every document header is clearly related to electric cars. The importance over time can be seen in the timeline. While there were some early related documents, the topic was not very important in the corpus before 2009 and only starts evolving around 2013, when the probability of seeing documents related to the topic increased significantly. Besides, we see from the nearby topics that the local neighborhood is inhabited by topics related to renewable energies, which further strengthens the idea we have about the topic. Compared to the prototypical shape of the innovation curve, the topic is at the beginning; the adoption process just started.

The topic in figure 8 is about tablets. *Tablet*, *android* and *windows* are amongst the top words for this topic. The neighboring topics are also related to electronic devices, for example *pc*, *display* or *handy*. The topic received major attention starting around 2006/2007 before peaking in 2010-2012. Since then the topic importance in the corpus has been decreasing. Given the topic timeline, the adoption to the tablet technology seems to be almost completed. While there are some kinks in the timeline, we can recognize the prototypical diffusion curve. The adoption process spans about 10 years.

Avenues for future research

From an economic point of view it would be interesting to know which entities[10] are being the main players in a topic. A simple possibility would be counting how often an entitiy has been mentioned. Going further, one could use sentiment analysis on document level to determine if entities have a positive or negative impact on the topic. Sentiment analysis could also be used to predict future appearance probabilities for topics based on the sentiment of documents. Through identification of first mentionings of firms we could classify them into categories such as innovators, early adopters, early majority, late majority and laggards. We could also interpret the probability measurements as public knowledge about new products or processes. Using live news feeds could offer the possibility to capture the diffusion of innovations with very little delay. Scaling up to more news sources could offer the potential to cover a larger share of the innovative activities that are going on.

[9] `https://github.com/davidlenz/pvtm/raw/master/heise_topics.zip`

[10] Named entities can be denoted with a proper name and are real-world objects, such as persons, locations, organizations, products etc.

6 Conclusion

There is an increasing interest in topic modeling, driven and ignited by the fast growing amount of textual data sources. In natural language processing, neural embedding methods have been shown to outperform standard methods on many tasks. They are therefore viable candidates for information retrieval from big text corpora, for example for topic modeling.

Our main contribution is threefold: 1) We propose Paragraph Vector Topic Modelling, which uses Doc2Vec to construct document vectors and Gaussian mixture clustering to cluster the resulting vectors into meaningful topics. 2) To make the interpretation of topics easier we use a novel topic report style sheet, partly based on features extracted from the learned document representations. 3) To show the applicability of our approach, we demonstrate the emergence of coherent topics from technology related news articles.

It became apparent that PVTM offers a useful alternative to LDA for large datasets. The topic vectors identified by the GMM were used as an effective way to provide local context to a topic by mapping document vectors into 2D using Barnes-Hut tSNE. The combination of top words, textual labels, evolution timelines and local context proves to be an effective way to interpret topics as the proposed topic report allowed to easily gain an accurate understanding of a topics content.

First empirical examples derived from an application of PVTM to technology newsticker data demonstrate the potential relevance for innovation economics. In particular, it enables the measurement of diffusion of innovations over time. It remains a task for further research to derive methods for prediction of diffusion and for the assessment of entities involved in innovative activities with regard to their stage of technology adoption. Taking the ongoing digitization into account, early identification and measurement of innovations will become more important. Therefore, it is planned to analyze to what extent the method is applicable also in a dynamic setting.

References

Ai, Q., L. Yang, J. Guo and W. B. Croft (2016). 'Analysis of the Paragraph Vector Model for Information Retrieval'. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, ICTIR '16. ACM, New York, NY, USA, S. 133–142.

Aletras, N., T. Baldwin, J. H. Lau and M. Stevenson (2017). 'Evaluating topic representations for exploring document collections'. *JASIST* 68(1), 154–167.

Baldwin, T., J. H. Lau, N. Aletras and I. Sorodoc (2017). 'Multimodal Topic Labelling'. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers. S. 701–706.

Baregheh, A., J. Rowley and S. Sambrook (2009). 'Towards a multidisciplinary definition of innovation'. *Management Decision* 47(8), 1323–1339.

Basave, A. E. C., Y. He and R. Xu (2014). 'Automatic Labelling of Topic Models Learned from Twitter by Summarisation'. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers. S. 618–624.

Bentley, J. L. (1975). 'Multidimensional Binary Search Trees Used for Associative Searching'. *Commun. ACM* 18(9), 509–517.

Bhatia, S., J. H. Lau and T. Baldwin (2016). 'Automatic Labelling of Topics with Neural Embeddings'. *CoRR* abs/1612.05340.

Blei, D. M., A. Y. Ng and M. I. Jordan (2003). 'Latent Dirichlet Allocation'. *J. Mach. Learn. Res.* 3, 993–1022.

Bridle, J. S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. Springer Berlin Heidelberg, Berlin, Heidelberg, S. 227–236.

Dadi, H., P. Venkatesh, P. Poornesh, N. Rao L and N. Kumar (2013). 'Tracking Multiple Moving Objects Using Gaussian Mixture Model' 3, 114–119.

Dai, A. M., C. Olah and Q. V. Le (2015). 'Document Embedding with Paragraph Vectors'. *CoRR* abs/1507.07998.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). 'Maximum Likelihood from Incomplete Data via the EM Algorithm'. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1), 1–38.

Dieng, A. B., C. Wang, J. Gao and J. W. Paisley (2016). 'TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency'. *CoRR* abs/1611.01702.

Fukunaga, K. and L. D. Hostetler (1975). 'The estimation of the gradient of a density function, with applications in pattern recognition'. *Information Theory, IEEE Transactions on* 21(1), 32–40.

Hansen, S., M. McMahon and A. Prat (2014). 'Transparency and Deliberation within the FOMC: a Computational Linguistics Approach'. Discussion Papers 1411, Centre for Macroeconomics (CFM).

Harris, Z. S. (1954). 'Distributional Structure'. *WORD* 10(2-3), 146–162.

Hashimoto, K., G. Kontonatsios, M. Miwa and S. Ananiadou (2016). 'Topic detection using Paragraph Vectors to support Active Learning in Systematic Reviews'. *Journal of Biomedical Informatics* 62, 59 – 65.

Hisano, R., D. Sornette, T. Mizuno, T. Ohnishi and T. Watanabe (2013). 'High Quality Topic Extraction from Business News Explains Abnormal Financial Market Volatility'. *PLOS ONE* 8(6), 1–12.

Kullback, S. and R. A. Leibler (1951). 'On information and sufficiency'. *The Annals of Mathematical Statistics* , 79–86.

Larsen, V. H. and L. A. Thorsrud (2015). 'The Value of News'. Working Papers No 6/2015, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

Lau, J. H. and T. Baldwin (2016). 'An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation'. *CoRR* abs/1607.05368.

Lau, J. H., T. Baldwin and T. Cohn (2017). 'Topically Driven Neural Language Model'. *CoRR* abs/1704.08012.

Lau, J. H., K. Grieser, D. Newman and T. Baldwin (2011). 'Automatic Labelling of Topic Models'. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11. Association for Computational Linguistics, Stroudsburg, PA, USA, S. 1536–1545.

Le, Q. V. and T. Mikolov (2014). 'Distributed Representations of Sentences and Documents.' In: ICML, volume 14. S. 1188–1196.

Liu, Y., Z. Liu, T.-S. Chua and M. Sun (2015). 'Topical Word Embeddings'. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15. AAAI Press, S. 2418–2424.

Lloyd, S. P. (1982). 'Least squares quantization in PCM'. *IEEE Trans. Information Theory* 28, 129–136.

Loper, E. and S. Bird (2002). 'NLTK: The Natural Language Toolkit'. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02. Association for Computational Linguistics, Stroudsburg, PA, USA, S. 63–70.

Lüdering, J. and P. Winker (2016). 'Forward or backward looking? The economic discourse and the observed reality'. *Jahrbücher für Nationalökonomie und Statistik* 236(4), 483–515.

Mei, Q., X. Shen and C. Zhai (2007). 'Automatic Labeling of Multinomial Topic Models'. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07. ACM, New York, NY, USA, S. 490–499.

Mikolov, T., K. Chen, G. Corrado and J. Dean (2013a). 'Efficient Estimation of Word Representations in Vector Space'. *CoRR* abs/1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado and J. Dean (2013b). 'Distributed Representations of Words and Phrases and their Compositionality'. *CoRR* abs/1310.4546.

Miner, G., D. Delen, J. Elder, A. Fast, T. Hill and R. A. Nisbet (2012). 'Chapter 4 - Applications and Use Cases for Text Mining'. In: G. Miner, D. Delen, J. Elder, A. Fast, T. Hill and R. A. Nisbet (Hg.), Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press, Boston, S. 53 – 72.

Mizuno, T., T. Ohnishi and T. Watanabe (2017). 'Novel and topical business news and their impact on stock market activity'. *EPJ Data Science* 6(1), 26.

Niu, L. and X. Dai (2015). 'Topic2Vec: Learning Distributed Representations of Topics'. *CoRR* abs/1506.08422.

Palangi, H., L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song and R. Ward (2016). 'Deep Sentence Embedding Using Long Short-term Memory Networks: Analysis and Application to Information Retrieval'. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 24(4), 694–707.

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

Řehůřek, R. and P. Sojka (2010). 'Software Framework for Topic Modelling with Large Corpora'. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, S. 45–50. http://is.muni.cz/publication/884893/en.

Reynolds, D. A. (2015). 'Gaussian Mixture Models'. In: Encyclopedia of Biometrics, Second Edition. S. 827–832.

Reynolds, D. A., T. F. Quatieri and R. B. Dunn (2000). 'Speaker Verification Using Adapted Gaussian Mixture Models'. *Digit. Signal Process.* 10(1), 19–41.

Rogers, E. M. (2003). Diffusion of innovations. Free Press, New York, NY [u.a.], Fifth edition.

Sammut, C. and G. I. Webb (2017). Encyclopedia of machine learning and data mining. Springer.

Schwarz, G. (1978). 'Estimating the Dimension of a Model'. *The Annals of Statistics* 6(2), 461–464.

Shi, B., W. Lam, S. Jameel, S. Schockaert and K. P. Lai (2017). 'Jointly Learning Word Embeddings and Latent Topics'. *CoRR* abs/1706.07276.

Stathoulopoulos, a. J. M.-G., Kostas (2017). 'Mapping Without A Map: Exploring The Uk Business Landscape Using Unsupervised Learning.' *SocArXiv* .

Tang, J., J. Liu, M. Zhang and Q. Mei (2016). 'Visualizing Large-scale and High-dimensional Data'. In: Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, S. 287–297.

Tian, F., B. Gao, D. He and T. Liu (2016). 'Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves'. *CoRR* abs/1604.02038.

van der Maaten, L. (2013). 'Barnes-Hut-SNE'. *CoRR* abs/1301.3342.

van der Maaten, L. (2014). 'Accelerating t-SNE using Tree-Based Algorithms'. *Journal of Machine Learning Research* 15, 3221–3245.

van der Maaten, L. and G. Hinton (2008). 'Visualizing High-Dimensional Data Using t-SNE'. *Journal of Machine Learning Research* 9, 2579–2605.

Varian, H. R. (2014). 'Big Data: New Tricks for Econometrics'. *Journal of Economic Perspectives* 28(2), 3–28.

Wehrheim, L. (2017). 'Economic History Goes Digital: Topic Modeling the Journal of Economic History'. *BGPE Discussion Paper Series* 177(ISSN 1863-5733), 146–162.

Yu, D. and L. Deng (2014). Automatic Speech Recognition: A Deep Learning Approach. Springer.