Human Behavior, Development and Society
ISSN 2651-1762, Volume 20, No 2, June 2019

Trending Research Topics Detection over Time Using the Latent Dirichlet Allocation Model

Pimpa Cheewaprakobkit, Asia-Pacific International University, Thailand

Abstract

Topic modeling has become an extraordinary tool for analyzing large, unclassified documents to detect patterns that use similar words. It also permits the discovery of hidden themes that pervade the collection, allowing annotation according to those themes, and then using these to summarize and search the text. This paper focuses on Latent Dirichlet Allocation, which is one of the most widely used methods for topic modeling. Trend detection and evolution of topics relating to research topics was attempted from the Journal of the Modern Language Association of America. The study also identified those documents that explained topics according to time series and citation. The experimental data consisted of 5,605 articles from the Journal between the years 1889–2007 held in the Journal Storage digital library. The results show that the Latent Dirichlet Allocation model can effectively detect distinct topics and those documents that explained these topics over time.

Keywords: *Topic model, Latent Dirichlet Allocation, trending topics detection, topic evolution*

Introduction

In recent years, electronic document collections have rapidly increased, such as articles and news found on websites and blogs. The challenge for researchers is to find automatic techniques for analyzing and summarizing these document collections. The topic model is a powerful method for discovering hidden knowledge and analyzing the data from different points of view, especially using semantic analysis. It continues to be developed and is used widely in tasks such as information retrieval, exploratory and predictive analysis of text, classification, clustering, and summaries of documents. Landauer, Foltz, and Laham (1998) introduced Latent Semantic Analysis (LSA), which was the first probabilistic topic model. The model was initially used for text classification and word analysis in document collections, and later was transformed into a different form of LSA. Hofmann (2001) proposed the new model be called Probabilistic Latent Semantic Analysis (PLSA), but this model still encountered many problems. Therefore, Blei, Ng, and Jordan (2003) improved the previous model and proposed the Latent Dirichlet Allocation model (LDA) to overcome problems faced with PLSA (Al-Khateeb, 2014).

Latent Dirichlet Allocation is a famous model that is highly effective in exploring hidden topics in data warehouses (Yang, Downey, & Boyd-Graber, 2015). The LDA approach is an unsupervised learning method for finding distinct topics that occur in collections of documents. In each document, there is a mixture of different topics. In each topic, there are different words, and each word is connected to those topics (Al-Khateeb, 2014). Initially the LDA model focused on analyzing overviews of document collections. Now it is used more to identify hidden topics and explore words that are frequently used by mapping the timestamps from each document, and estimating the probability of all words that are related to these hidden topics (Bolelli, Ertekin, & Giles, 2009).

Studying the trending topics model without considering the time will cause errors due to obsolete information. As time goes by, the trends found in documents change (Meng, Zhang, & Guo, 2012). The content may be the same, but the meaning may change. For example, if we would like to know which product is the best-selling online in Twitter, the results might be out of date, since the LDA model did not notice that the information was 10 years old (Alghamdi & Alfalqi, 2015). It is essential to model the trend of topics so that people can see how topics have evolved.

Literature Review

Latent Dirichlet Allocation is a useful model that overcomes the overfitting problem of PLSA, which has received much attention on many applications (Zeng, 2012). Topic models are mostly used

in social networks, computer vision, and population genetics (Fei–Fei & Perona, 2005; Luo, Stenger, Zhao, & Kim, 2015). The Latent Dirichlet Allocation model is also used for in–depth business studies of customer behavior, comments, and other feedback to predict customer characteristics and what kind of products customers might be interested in in the future (Christidis, Apostolou, & Mentzas, 2010). Topics over Time (Wang & McCallum, 2006) is another LDA that jointly models both word co-occurrence patterns and localization in continuous time. Wang and McCallum's approach to Non–Markov Continuous–Time models was different. Each word is associated with timestamps; the algorithm does not discretize time. Therefore, it will generate a distinct timestamp for each word in the document. This model focuses on the evolution of topics in isolation; it does not consider information from the literature.

Blei and Lafferty (2006) introduced the Dynamic Topic Model, which is an extension of LDA. The method analyzes the evolution of topics or unobserved topics over time by the capture of topic dynamics through a learning and iteration process. It was applied to Science magazine articles from 1881–2000 to track how articles have changed over time. Another extension of LDA is the Author–Topic Model introduced by Rosen-Zvi, Chemudugunta, Griffiths, Smyth, and Steyvers (2010). This model works with topic distributions that correspond to each author in the corpora. This method aims to discover the author's research interests from a collection of documents that includes authorship information. Hisano, Sornette, Mizuno, Ohnishi, and Watanabe (2013) also used LDA to show how business news affects stock market activity. The flow of news can explain abnormal market activity and help to predict trading. Words that represent topic distribution extract essential pieces of information that influence the stock market.

Many topic models have been introduced. Hoffman, Blei, and Bach (2010) developed an online Variational Bayes algorithm for LDA analysis. The Gibbs sampling is an approximate posterior inference that can analyze massive collections of documents by using the previously learned model, resampling topics and assignments to new documents. In this paper, topic trends over time detection was performed using the LDA model. The results are presented using the frequent words and papers that are highly associated with a topic expressed in a topic proportion histogram.

Purpose of the Study

The purpose of this study was to detect the trend of topics relating to research topics found in an online database. The study was implemented using an LDA model, along with outputs of words and papers most highly associated with a topic. In this way, it is possible to see topic trends, how this topic has evolved, and lists of papers, limitations of each paper, as well as methods and strategies that previous researchers had studied, all of which will be very useful for future researchers.

Methodology and Methods

Topic Modeling

Topic models are a well-known and significant part of modern machine learning techniques. They are frequently used in text mining, computer science, and other fields. The algorithms of these models create a useful perspective of a vast collection of documents that allows the establishment of relationships among these documents (Jelodar, Wang, Yuan, & Feng, 2017; Liu, Tang, Dong, Yao, & Zhou, 2016). Topic modeling algorithms (Blei, 2012) are an efficient way to find themes or topics in documents, an approach which automatically teases out the topics. These algorithms are used on the corpora of textual data, grouping documents, and in collections to establish semantically-meaningful groupings. The assumption is that each document exhibits multiple topics. Figure 1 shows the intuition behind latent Dirichlet allocation (Nastase, 2012).

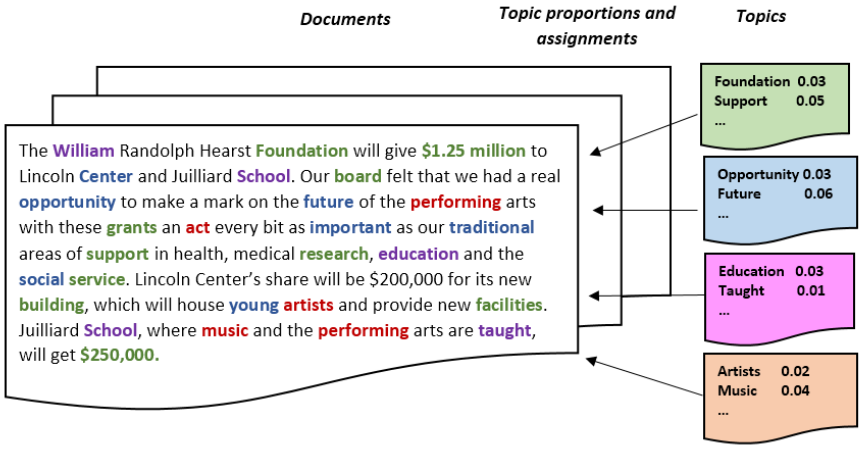


Figure 1. The Intuition behind the Latent Dirichlet Allocation Model

The document illustrated above contains a number of topics, and the model can be set up to determine the number of topics recognized in the document. Each topic represents a distribution incorporating words that are highlighted in different colors. On the right, topic proportions are shown for this document. For example, there are four topics in this document, and on the right side, each topic will contain the words and probability values. For example, topics 1 to N each have nominated words 1 to K. The question is, "What level of probability should be attached to each word?" The total probability value for each topic naturally must be equal to 1.0. As seen in this Figure, the first topic group (green color) contains the word "Foundation" and has a probability value of 0.03.

A document is usually represented by a Bag of Words as shown in Table 1 (Liu, et al., 2016). For example, there are four words (foundation, opportunity, education, and music) and six documents (d_1 to d_6). The word count in the cell under topic three and column d_4 is five, meaning that the frequency of the word "education" in document d_4 is 5.0. The Bag of Words represents the corpus.

Table 1. An Example of a Bag of Words

Topic	Document					
	d_1	d_2	d_3	d_4	d_5	d_6
Foundation	3	0	4	0	0	0
Opportunity	0	6	0	0	0	0
Education	0	2	0	5	0	0
Music	0	0	2	5	0	0

Model Training

The topic models represent an unsupervised class of machine learning algorithms that does not require labeling or any descriptions of a document for model training processes. Topic modeling aims to discover the abstract "topics," themes, or semantic types of topics in a collection document by analyzing the words and grouping them into topics to see their usage in each group. However, topic names are assigned based on groups of key words.

The Generative Process

Since documents exhibit multiple topics, therefore there is need to simulate the generative process of documents to detect a word distribution over each topic and topic distribution over each document. The process is to randomly choose a topic from the distribution over latent topics, where a distribution over words characterizes a topic, and then randomly choose the word from the corresponding topic (Blei, 2012).

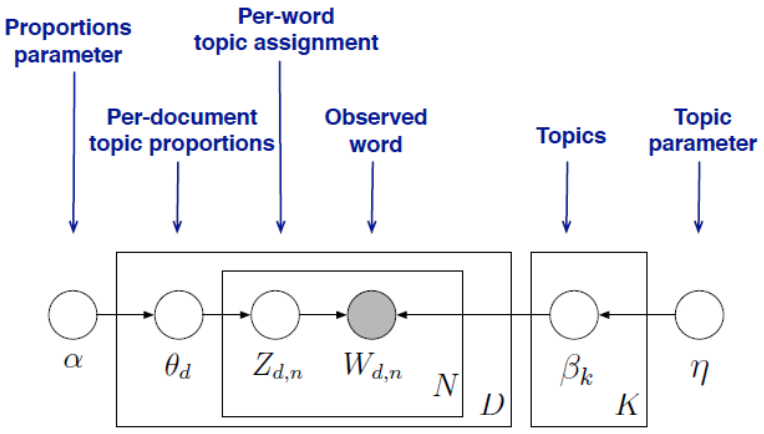


Figure 2. A Graphical Model of LDA (Blei, 2012)

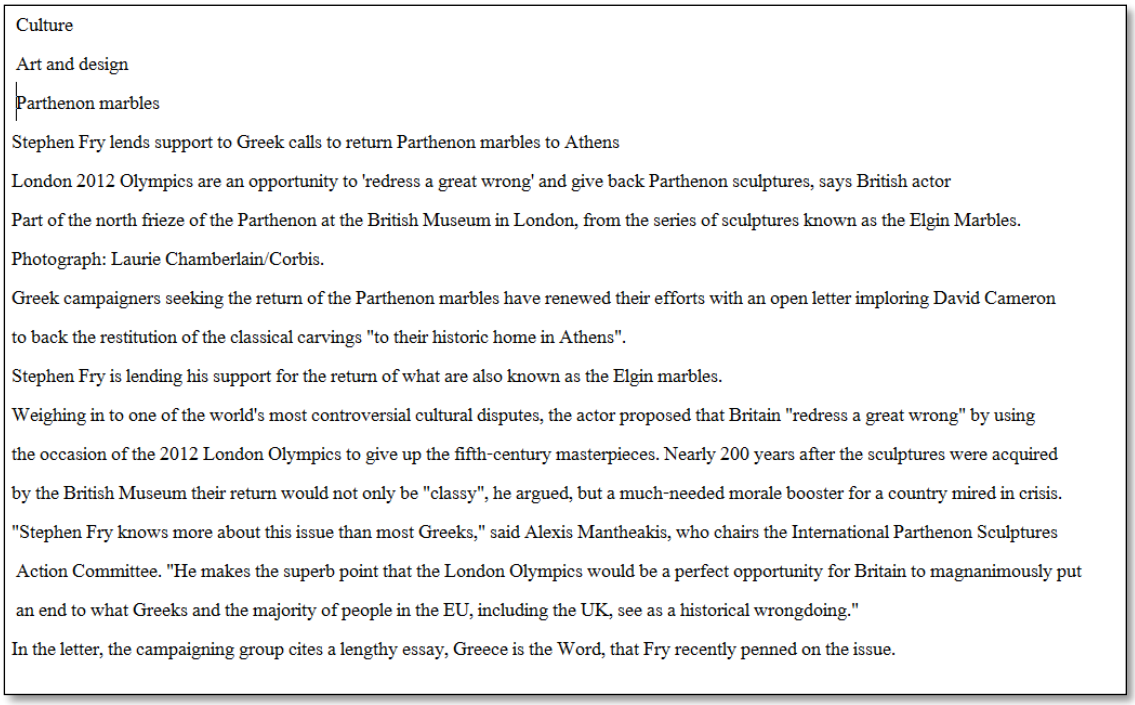
The Graphical Model of LDA in Figure 2 can be understood as follows:

1. Document D is represented as a mixture over latent topics K.
2. W (the gray circle, an observed variable) means the word that appears in the document. There are N words in the document.
3. Z (the white circle, a latent variable) refers to the topic of the word in each document. There are K topics.
4. θ is a proportion topic for each document, or the probability of each document and topic (such as document 1 to D in each document, with topic 1 to K in the probability). The total probability of each topic must be equal to 1.
5. α is a proportional parameter of that controls the number of topics found in each document, meaning that if alpha is high, the number of topics in the document will also be high, but if alpha is low, then similar words in the document will be grouped into a relatively smaller number of topics.
6. β is a proportional parameter or probability for each topic and word (such as topic 1 to K, or word 1 to D), and its probability value.
7. η is a word distribution control in each topic means that if η is high, the word distribution in the topic will be very consistent, but if the η is low, the word distribution in the topic will fall somewhere.

Latent Dirichlet Allocation is a commonly-used algorithm for topic modeling. It represents documents as random mixtures of topics. The LDA model generates automatic summaries of topics in terms of probability distribution over words for each topic, and for each document over a topic. The LDA model assumes that documents on roughly similar topics will use the same groups of words. Therefore, it identifies group of words in the corpus that appear frequently or occur together within the document.

Data Collection

The article file was in text format and consisted of 5,605 articles from the years 1889–2007 found in the Journal of the Modern Language Association of America, which was supplied by the Journal Storage digital library. Each data file was restricted to full length articles with more than 2000 words. Figure 3 shows a sample of raw data.



Culture

Art and design

Parthenon marbles

Stephen Fry lends support to Greek calls to return Parthenon marbles to Athens

London 2012 Olympics are an opportunity to 'redress a great wrong' and give back Parthenon sculptures, says British actor

Part of the north frieze of the Parthenon at the British Museum in London, from the series of sculptures known as the Elgin Marbles.

Photograph: Laurie Chamberlain/Corbis.

Greek campaigners seeking the return of the Parthenon marbles have renewed their efforts with an open letter imploring David Cameron to back the restitution of the classical carvings "to their historic home in Athens".

Stephen Fry is lending his support for the return of what are also known as the Elgin marbles.

Weighing in to one of the world's most controversial cultural disputes, the actor proposed that Britain "redress a great wrong" by using the occasion of the 2012 London Olympics to give up the fifth-century masterpieces. Nearly 200 years after the sculptures were acquired by the British Museum their return would not only be "classy", he argued, but a much-needed morale booster for a country mired in crisis.

"Stephen Fry knows more about this issue than most Greeks," said Alexis Mantheakis, who chairs the International Parthenon Sculptures Action Committee. "He makes the superb point that the London Olympics would be a perfect opportunity for Britain to magnanimously put an end to what Greeks and the majority of people in the EU, including the UK, see as a historical wrongdoing."

In the letter, the campaigning group cites a lengthy essay, Greece is the Word, that Fry recently penned on the issue.

Figure 3. Sample of Raw Data

Software Used

There are many tools and software used for training an LDA model, such as the Stanford Topic Modeling Toolbox, and Matrix Laboratory (MATLAB). In this study, we used MALLET, a Java "Machine Learning for Language Toolkit". It is open-source software; its script was licensed from Massachusetts Institute of Technology (MIT) (Goldstone, 2013), and later modified by the researcher.

Data Pre-Processing

When pre-processing the data, two steps were involved (Aiello et al., 2013) in order to reduce the number of errors in the raw data collected from users.

- Tokenization: Raw data can contain all types of errors in punctuation, hyphenation, abbreviations and wrong spelling. The tokenization process removes stop words, uninformative words, and punctuation. Uninformative word lists can be quite long, depending on the domain producing the text in question.

- Stemming: This represents the process of reducing the number of words by reduction to root words. For example, the words stemmer, stemming, stemmed, and stem become normalized to one common form, so it will be “stem.” The data are then ready, and can be trained and tested.

Findings and Results

The data was run through the LDA model, and 64 topics were extracted. Table 2 shows four topics out of the 64 topics, along with the top 10 words within each topic. The topic labels were assigned based on top key words.

This study attempted to classify words and documents through patterns of co-occurring words. A group of words was chosen that were assigned topic labels such as “University,” “Language,” “Social,” and “Home” to see the trends of topic evolution. Table 3 shows the most frequently used words and documents that were associated with the topic “University,” the percentage (%) of the number of tokens (views on Journal Storage), the number of tokens found in the documents, and the year.

Table 2. Example of Top 10 Most Frequent Words from Four Topics

Topics	"University"	"Language"	"Social"	"Home"
Words	university	languages	social	house
	students	English	political	family
	research	study	society	father
	graduate	foreign	public	marriage
	academic	literature	economic	wife
	education	present	community	love
	teaching	speech	moral	husband
	scholars	meaning	revolution	daughter
	professional	word	value	children
	intellectual	structure	culture	girl

Table 3. Example of Top Words Defining Topics, and Citation Documents Related to Each Topic

Label	University
Top Words	University, students, research, graduate, academic, education, teaching, scholars, professional, intellectual

	Top 10 Related Citation Documents	Tokens	Number of Tokens (Views on Journal Storage)	Percent (%)	Year
1.	Cohen, Walter. "The Economics of Doctoral Education in Literature." *PMLA* 115, no. 5 (October 2000): 1164–1187.	3940	5661	69.6%	2000
2.	Ferguson, Margaret. "Come Tell Me How You Live." *PMLA* 115, no. 5 (October 2000): 1246–1250.	738	1065	69.3%	2000
3.	Showalter, Elaine. "Presidential Address 1998: Regeneration." *PMLA* 114, no. 3 (May 1999): 318–328.	5272	7744	68.1%	1999
4.	Guillory, John. "The System of Graduate Education." *PMLA* 115, no. 5 (October 2000): 1154–1163.	1661	2487	66.8%	2000
5.	Abram, Michael E. "Graduate Student Assistants and Collective Bargaining: What Model?" *PMLA* 115, no. 5 (October 2000): 1188–1191.	597	898	66.5%	2000
6.	Feal, Rosemary G. "Report of the Executive Director." *PMLA* 118, no. 3 (May 2003): 618–639.	2263	3443	65.7%	2003
7.	Bartholomae, David. "Composition, 1900–2000." *PMLA* 115, no. 7 (December 2000): 1950–1954.	686	1057	64.9%	2000
8.	Stimpson, Catharine R. "Myths of Transformation: Realities of Change." *PMLA* 115, no. 5 (October 2000): 1142–1153.	1598	2494	64.1%	2000
9.	Hohendahl, Peter Uwe. "After Three Decades of Crisis: What Is the Purpose of a PhD Program in Foreign Languages?" *PMLA* 115, no. 5 (October 2000): 1228–1238.	1705	2662	64.0%	2000
10.	Liu, Catherine, Fay Beauchamp, and Gerald Graff. "Regeneration in the Humanities." *PMLA* 115, no. 1 (January 2000): 90-93.	731	1148	63.7%	2000

Figure 4 shows a histogram of popularity trends of sample topics discovered for each year from 1889–2007. It can be seen that the popularity of university topics have been increasing over

these years especially in the year 2000, which was highest. After that year, the trend gradually began to decrease.

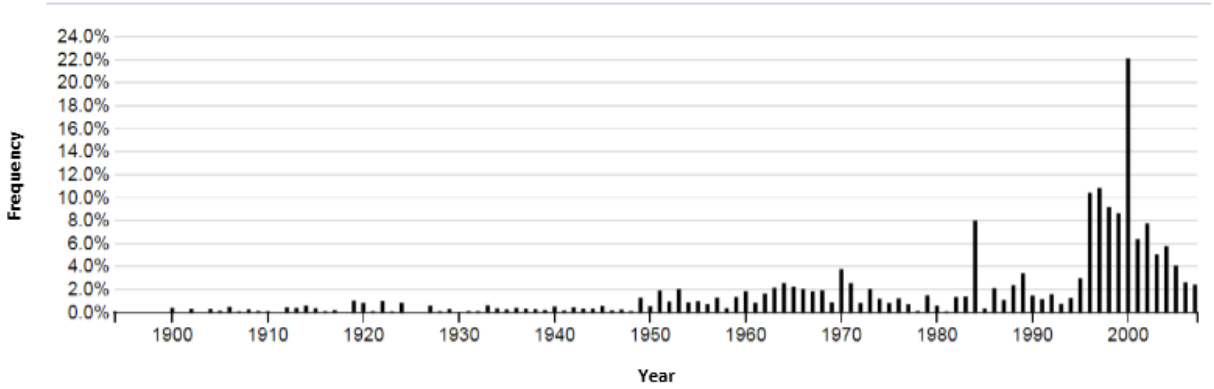


Figure 4. Topic Proportion Histogram from the LDA Model

Limitation and Implication

It is an ongoing research problem to find the best evaluation method for topic modeling using the Latent Dirichlet Allocation model. Because the model's outputs represent the probability of a word in a given topic, we can rely on human judgment by looking at the relevant words obtained for each topic. Our evaluation is based on whether or not the words are relevant to a topic, taken as a whole, and produce a semantically coherent topic such that the topic can be given a label. There are existing methods of evaluating LDA, such as measuring perplexity and other complex measurements, but they may be less relevant to evaluating the output of a topic model than human judgment.

Another limitation is to choose a suitable number of topics for LDA. Mostly we looked at goodness-of-fit and fitting time by calculating the perplexity of a held-out set of documents. The perplexity values are used to measure the effectiveness of the model that describes a set of documents or predicts a sample. However, the number of topics can be adjusted depending on the data type and content structure.

Conclusions

In this paper, a Latent Dirichlet Allocation model was used to analyze trending topics over time by looking at the probability of top key words from each topic, and monitoring their popularity and change over time. The LDA model is a useful method because it enhances a researcher's ability to interpret information.

However, in our findings, we were unable to conclude that a semantic label can be given for the topics and the associated words outputted by the model. In order to obtain more concrete topics for implementation of LDA, future work will aim at experimenting with various methods of preprocessing data, and with adding word features and n-gram models to discover topics.

References

Aiello, L., Petkos, G., Martin, C., Corney, D., Papadopoulos, S. Skraba, R., (. . .) Jaimes, A. (2013). Sensing trending topics in Twitter. *IEEE Transactions on Multimedia*, *15*(6), 1268–1282.

Alghamdi, R., & Alfalqi, K. (2015). A Survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, *6*(1), 147–153.

Al-Khateeb, S. (2014). *Topic modeling for associated press articles using Latent Dirichlet Allocation [LDA].* Retrieved from https://pdfs.semanticscholar.org/4eeb/900c36a8a9714cceb8a3ccc6ec8a307d2170.pdf

Blei, D., & Lafferty, J. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*. ACM, New York, 113–120. doi: 10.1145/ 1143844.1143859

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. doi:10.1145/2133806.2133826

Blei, M., Ng, Y., & Jordan, I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bolelli, L., Ertekin, S., & Giles, C. (2009). Topic and trend detection in text collections using Latent Dirichlet Allocation. *Advances in Information Retrieval,* 5478, 776–780. doi:10.1007/978-3-642-00958-7_84

Christidis, K., Apostolou, D., & Mentzas, G. (2010). Exploring customer preferences with probabilistic topics models. Retrieved from http://imu.ntua.gr/sites/default/files/biblio/Papers/exploring-customer-preferences-with-probabilistic-topics-models.pdf

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *Proceeding of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'05), 2, 524–531. doi: 10.1109/CVPR.2005.16

Goldstone, A. (2013). *Dfr-Browser license* (Licensed under the MIT license). Retrieved April 24, 2019, from https://github.com/agoldst/dfr-browser/find/master

Hisano, R., Sornette, D., Mizuno, T., Ohnishi, T., & Watanabe, T. (2013). High quality topic extraction from business news explains abnormal financial market volatility. *PLoS ONE, 8*(6), e64846. doi.org/10.1371/journal.pone. 0064846

Hoffman, M., Blei, D., & Bach, F. (2010). Online learning for Latent Dirichlet Allocation. *Proceeding of the 23rd International Conference on Neural Information Processing Systems (NIPS-10),* 856–864.

Hofmann, T. (2001). Unsupervised learning by probabilistic Latent semantic analysis. *Machine Learning*, *42*(1–2), 177–196. doi: 10.1023/A:1007617005950.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2017). Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 1–43.* Retrieved from https:// link.springer.com/article/10.1007/s11042-018-6894-4

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in Bioinformatics. *SpringerPlus*, *5*(1):1608. doi: 10.1186/s40064-016-3252-8.

Luo, W., Stenger, B., Zhao, X., & Kim, T-K. (2015). Automatic topic discovery for multi-object tracking. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Texas,* 3820–3826. Retrieved from https:// dl.acm.org/citation.cfm?id=2888246

Meng, C., Zhang, M., & Guo, W. (2012). *Evolution of movie topics over time*. Retrieved from http://cs229.stanford. edu/proj2012/MengZhangGuo-EvolutionofMovieTopicsOverTime.pdf

Nastase, V. (2012). Introduction to topic models. Retrieved from http://www.cl.uni-heidelberg.de/courses/ss12/ topicmodels/intro.pdf

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, *28*(1), 1–38.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, ACM Press. 424–433. doi: 10.1145/1150402.1150450

Yang, Y., Downey, D., & Boyd-Graber, J. (2015). Efficient methods for incorporating knowledge into topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 308-317. Retrieved from https://www.aclweb.org/anthology/D/ D15/D15-1037.pdf

Zeng, J. (2012). A topic modeling toolbox using belief propagation. *Journal of Machine Learning Research*, *13*(1), 2233–2236.