Storyline detection and tracking using Dynamic Latent Dirichlet Allocation

Daniel Brüggermann, Yannik Hermey, Carsten Orth,
Darius Schneider, Stefan Selzer, Gerasimos Spanakis [*]
Department of Data Science and Knowledge Engineering
Maastricht University
Maastricht, Netherlands, 6200MD
{d.bruggermann, y.hermey, c.orth, d.schneider,
s.selzer, jerry.spanakis}@maastrichtuniversity.nl

Abstract

In this paper we consider the problem of detecting and tracking storylines over time using news text corpora. World wide web creates vast amounts of information and handling, managing and utilizing this information is difficult without having systems that are able to identify trends, arcs and stories and how they evolve through time. The proposed approach utilizes a dynamic version of Latent Dirichlet Allocation (DLDA) over discrete time steps and makes it possible to identify topics within storylines as they appear and track them through time. Moreover, a graphical tool for visualizing topics and changes is implemented and allows for easy navigation through the topics and their corresponding documents. Experimental analysis on Reuters RCV1 corpus reveals that the proposed approach can be effectively used as a tool for identifying turning points in storylines and their evolutions while at the same time allowing for an efficient visualization.

1 Introduction

Growth of internet came along with an increasingly complex amount of text data from emails, news sources, forums, etc. As a consequence, it is impossible for individuals to keep track of all relevant storylines and moreover to detect changes in emerging trends or topics.

Many stakeholders (companies, individuals, policy makers, etc.) would be interested to harness the amount of free text data available in the web in order to develop intelligent algorithms that are able to react to emerging topics as fast as possible and at the same time track existing topics over long time spans. There are many techniques about topic extraction like Nonnegative Matrix Factorization (NMF) (Sra and Dhillon, 2005) or Latent Dirichlet Allocation (LDA) (Blei et al., 2003) but there are not many extensions to dynamic data handling. Time dependent modeling of documents can be computationally expensive and complex (Allan et al., 1998) and despite the fact that such approaches can be effective, none of these effectively handles the visualization issue which can make results more intuitive. Thus, effective approaches in terms of both computation and visualization of the results need to be pursued.

[*] Authors contributed equally to the manuscript, thus appear in alphabetical order. Correspondence to: jerry.spanakis@maastrichtuniversity.nl

Proceedings of 2nd Workshop on Computing News Storylines, pages 9–19,
Austin, TX, November 5, 2016. ©2016 Association for Computational Linguistics

This research work aims at implementing a technique to present stories and their changes from a news items flow by detecting and tracking topics through time. Results will be visualized and evaluated using the (fully annotated and immediately available) RCV1 Reuters corpus (810.000 documents) which is partly utilized in this work. The remainder of the paper is organized as follows. Section 2 presents an overview of current research work in the area. The proposed approach is described in Section 3, while experimental results are presented in Section 4. Finally, Section 5 concludes the paper and presents future improvement work.

2 Related Work

Topic detection and tracking is a long studied task (Fiscus and Doddington, 2002) and many approaches have already been attempted. Non-negative matrix factorization is used in the field of text mining to factorize and thereby decrease the dimension of a large matrix (Lee and Seung, 1999). For topic detection, the original matrix can be composed of terms represented in the rows and documents represented in the columns, while the cell values represent the TF-IDF value (Sparck Jones, 1972) of each term in each document. As TF-IDF values cannot be negative, the algorithm's requirement of a matrix with only non-negative values is fulfilled. Ranking the terms of a topic by their matrix value reveals the most relevant terms that can make up the description of this topic. In a similar way, documents of a topic can be ranked as well. This makes it possible to visualize topics according to their importance amongst all documents (Godfrey et al., 2014).

There exist only few approaches so far that applied NMF for dynamically changing text data, i.e. when detecting and tracking topics over time. Although the original data size can be too large for matrix factorization, there already exist variants of the algorithm using an dynamic approach, processing data in chunks (Wang et al., 2011). (Cao et al., 2007) use an online NMF algorithm that applies the factorization to the data of each time step and then updates the matrix bases from the previous calculations accordingly by some metric. However, both these algorithms are not able to detect emerging topics. (Saha and Sindhwani, 2012) defines an evolving set and an emerging set of topics within the NMF algorithm and appends the matrices accordingly in both dimensions whenever a new time step is considered. Topics are only detected when they emerge rapidly, and removing topics that are not relevant anymore is not discussed (the matrices increase gradually). (Tannenbaum et al., 2015) introduces a sliding window over the time steps. First, NMF is applied on a certain time step, and then the discovered topics are assigned to the topic model defined by the previous time steps, if possible. If they do not fit into the model, they are added to the emerging set of topics, which are added to the model as soon as there are enough documents that cover this new topic. Within the emerging set, the texts are categorized into new topics using hierarchical clustering.

All these works have several drawbacks. First, they mostly focus on sources like social media (Yang and Leskovec, 2011), (Paul and Girju, 2009), thus the magnitude of their data is several orders smaller than ours. Moreover, temporal dimension introduces further complexity due to the need for additional distributions or function that characterize this dynamic change (Hong et al., 2011).

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a generative probabilistic mixture model for topic detection. In such a model, words define a vocabulary and topics are represented by a probabilistic distribution of words from this vocabulary. Each word may be part of several topic representations. LDA assumes that each document from a collection of documents is generated from a probabilistic distribution of topics. Bayes' Theorem in combination with a Dirichlet distribution as prior distribution are used to approximate the true posterior distribution. The probability space defined by the probabilities of the words and topics is multidimensional which is represented by a multinomial distribution. For the a priori estimation the conjugate distribution is needed, which corresponds to a Dirichlet distribution in this case. Information gain is used as measure for the difference between two iterated probability distributions and thereby acts as convergence criterion.

LDA has been extended in order to handle documents over long periods and many variations exist. Other approaches have been proposed as well (Banerjee and Basu, 2007) but scalability is an issue and visualization is not feasible. A milestone in the area was the work of (Wang and McCallum, 2006) since they associated a beta distribution over time with each topic to capture its popularity. There are also nonparametric models developed either using Gaussian mixture distributions (Ahmed and Xing, 2012) or utilizing Markovian assumptions (Dubey et al., 2013). These models are very effective but it is very difficult to choose a good distribution over time that allows both flexible changes and effective inferences. Disadvantage of these methods is that they either exhibit limited forms of temporal variation, or require computationally expensive inference methods.

There are extensions of the LDA model towards topic tracking over time such as (Wei et al., 2007). But according to (Wang et al., 2008), these methods deal with constant topics and the timestamps are used for better discovery. Opposed to that, our approach utilizes a dynamic model of LDA (Blei and Lafferty, 2006) that after examining the generated distributions for changes is able to detect turning points or storyline arcs. Finally, results are visualized using a stacked graph modeling and can be explored in an intuitive way by relating one topic to another.

LDA was selected due to the fact that topic modeling provides a powerful tool to uncover the thematic structure of large document collections. Moreover, the dynamic version of it (DLDA) offers the possibility of analyzing the topic distributions per time and provide insights on their changes and evolutions. Pre-selecting the number of topics is a known disadvantage of traditional LDA models, however experiments show that evolution of topics can still be identified between consecutive time steps. Selecting the initial number of topics relies on user requirements and on how much detail in the storylines (and their changes) is desired.

3 The proposed approach

3.1 Preprocessing

The preprocessing steps are separated into two major parts. First, the article text is extracted from the original documents and then the text is analyzed using natural language processing techniques to generate a meaningful vocabulary for the topic extraction. Then, the main natural language processing of the article text, namely the tokenization, named entity recognition (NER) and lemmatization, is performed using the Stanford CoreNLP (Manning et al., 2014). The text is split into single tokens and then these are filtered according to the categories, that the named entity recognition assigned to them. As numbers are not very descriptive for topics, the named entity recognition is used to exclude all tokens categorized in number-related categories, precisely those of the categories "DATE", "DURATION", "MONEY", "NUMBER", "ORDINAL", "PERCENT", "TIME" and "SET".

Lemmatization is used to normalize the tokens without loosing informational detail. Standard stemming algorithms aggressively reduce words to a common base even if these words are actually of different meaning thus they are not considered here (e.g. there is a difference between "marketing" and "markets"). On the other hand, lemmatization only removes inflectional endings and returns a dictionary form of the token.

The Stanford parser is highly context dependent and does not always categorize words correctly. As there are a lot of number-related words left, an additional step of removing such words is performed by regex-cleaning. This step also removes any words containing special characters human language words normally do not contain.

The next normalization step involves removing dashes and concatenating combined words as well as spell correction. As the news articles contain a lot of proper names and improperly resolving ambiguities can lead to loss of information, spell correction is done very carefully. The Levenshtein distance is used (Navarro, 2001) to correct those words of distance one who do not reveal ambiguities when compared to the entries of the official Hunspell dictionary. Named identities are excluded, as they cannot be correctly processed automatically. As the spell correction is computationally expensive, care is also taken, that it is only performed, when it makes sense. A preliminary, much faster test for existing equal words of the same length in the dictionary is preformed beforehand. A comparison is only done when the length of the strings differs by no more than the tested distance, which is 1 in this case. Last but not least, the spell correction is done as almost final step, after all the other refinements are applied.

Finally, the remaining list is filtered using a stopword list, that contains the most common words like "the" and "and". Such words do not contribute to reasonable meaning of the article and are not useful to identify topic content.

3.2 Dynamic Latent Dirichlet Allocation (DLDA)

The Dynamic LDA model is adopted and used on topics aggregated in time epochs and a state space model handles transitions of the topics from one epoch to another. A gaussian probabilistic model to obtain the posterior probabilities on the evolving topics along the time line is added as additional dimension. Figure 1 shows a graphical representation of the dynamic topic model.

DLDA (as LDA) needs to know the number of topics in advance. That depends on the user and the number of stories that we could like to be detected. For example, the RCV1 corpus has 103 actually used annotated topics, plus a large amount of unlabeled documents, so the parameter for the extraction is set to 104 topics. This corresponds to the 103 annotated topics and one additional "ERROR" topic for the unlabeled documents. Goal for this was to as accurately cover the original categories of the corpus, although more experiments with less topics (15, 30 and 60) were conducted. Moreover, the timestep has to be determined at this point. This again can be set to any time unit. For example, the RCV1 corpus used here (July and August of 1996) contains 42 days which makes exactly 6

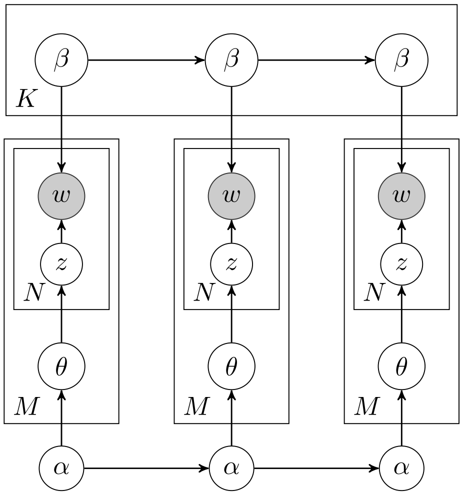


Figure 1: Plate diagram representing the dynamic topic model (for three time slices) as a Bayesian network. The model for each time slice corresponds to the original LDA process. Additionally, each topic's parameters α and β evolve over time (Blei and Lafferty, 2006)

weeks of time. The dynamic topic model is accordingly applied to 6 time steps corresponding to the 6 weeks of the data set.

3.3 Topic emergence and storyline detection

DLDA produces a list of topic distributions per time step. Topics appear not to evolve in a great degree and this trend is reflected by the word distributions. Inspecting them in detail reveals little difference among the word distributions for the time steps of each topic. Figure 2 shows the word distribution scores for the time steps 0 and 1 and the difference between them for a topic from the RCV1 corpus. The number of topics in the dynamic topic model is fixed and the computation infers the topics through a probabilistic distribution. This does not produce dynamic topics (appearing or disappearing) but instead, the word distributions for one topic could be used to capture gradual changes gradually over time and detect a new turning point (or arc) in the storyline of this topic.

To identify such turning points and changes inside the word distributions, the second step of the two folded approach consists of applying a similarity measure to identify time steps, where the word distributions change enough to identify a new arc within current topic. Cosine similarity is used in this case to measure differences in the distributions from time step to time step.

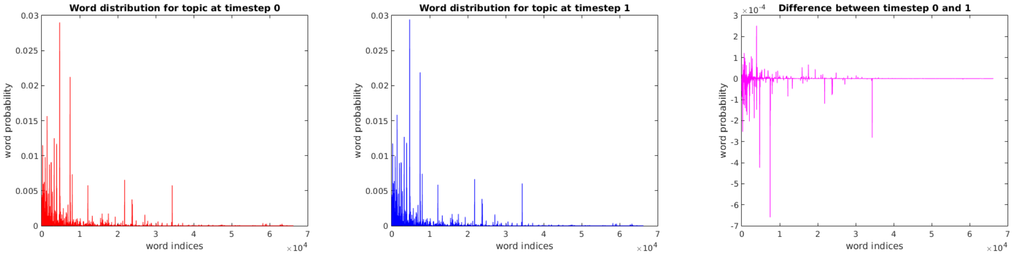
$$diff_i = ||TD_i(t) - TD_i(t-1)|| \qquad (1)$$

where:

- i refers to current topic,
- $TD_i(t)$ refers to the topic distribution at current time-step t,
- $TD_i(t-1)$ refers to the topic distribution at previous time-step $t-1$

A turning point is identified if $diff_i$ is larger than a threshold which can be selected by the user (see next Section for more details on this). This is interpreted as a change to the topic distribution, which means that significant events within the topic are noticed, and add new information to the storyline. These changes in the storylines can also be visualized by a topic river like the one in Figure 3. Peaks (like for example the yellow peak at the 3rd time-step reveal important changes in the storyline development and thus can be used to monitor the storyline. It is therefore assumed that each topic corresponds to one storyline.

Moreover, storyline aggregation can be performed using the same similarity measure as before. Points of aggregation, where previously



(a) Time step 0 (b) Time step 1 (c) Difference from 0 to 1

Figure 2: Example word distributions for neighboring time steps of one topic

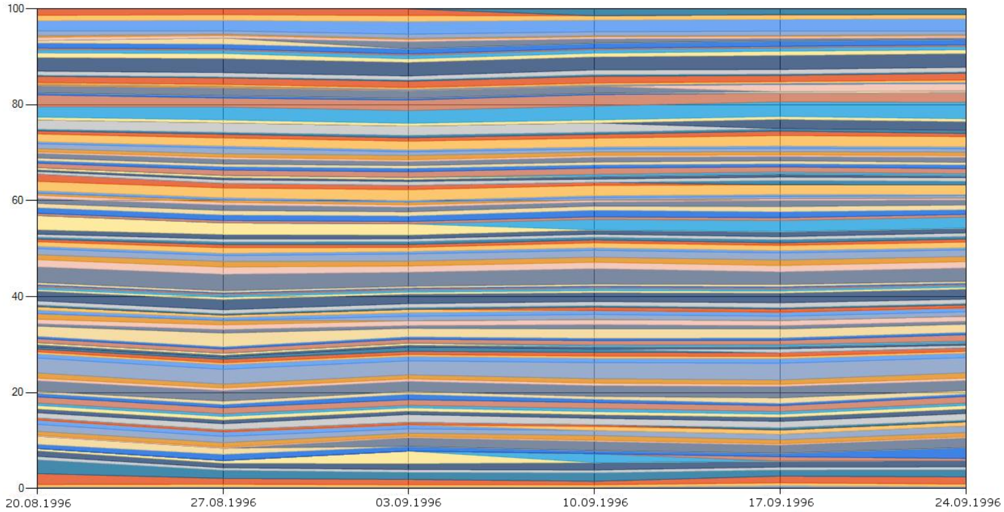


Figure 3: Topic rivers for August and September 1996 for emerging topics

separate topics should become one, are computed this way. As DLDA once more does a good job in clustering, the distance between different topics is rather high.

4 Experiments

4.1 The dataset

The process described in Section 3.1 is applied to the text content of the news articles from August and September 1996 of the RCV1 corpus (Lewis et al., 2004) to obtain a vocabulary containing terms, that are as meaningful and descriptive as possible while eliminating as much noise, consisting of not descriptive or ambiguous terms, as possible. The first two months of the RCV1 corpus contain 83.650 documents, which is about 10% of the corpus documents overall. Table 1 shows the results of the reduction of the number of terms from 308.854 distinct terms of about 16 million words overall to a final vocabulary size of 131.202.

Preprocessing leads to a reduction to 42% of the number of distinct tokens. Most important reduction in the vocabulary size comes from the NER category removals (almost half) which

Overall terms	16.467.261	
Distinct tokens	308854	100.00%
NER category removals	133976	43.38%
Lemmatization removals	17372	5.62%
Regex cleanup removals	18962	6.14%
Spellcheck cleanup removals	6768	2.19%
Stopword removals	574	0.19%
Final vocabulary size	131202	42.48%

Table 1: Terms and vocabulary for documents of August and September 1996 of RCV1

contributes most to keeping words and tokens that contribute most to the topics/storylines descriptions.

August and September 1996 contain newtimes from 42 days, thus 6 (weekly) timesteps are computed. Table 2 shows, that the documents are almost evenly distributed among the weeks.

Aug 20th-26th	Aug 27th-Sep 2nd	Sep 3rd-9th	Sep 10th-16th	Sep 17th-23rd	Sep 24th-30th
12807	12800	13953	14606	14487	14997

Table 2: Number of documents per week in August and September 1996 of RVC1

4.2 Storyline detection

The dynamic topic model partially identifies and reveals events of late summer 1996. Table 3 shows some of the identified events. The top 10 words of the topics' word distributions already give a precise overview of the topics' contents.

These topics describe events over a the period of two months and their change during the examined time frame (2 months) can be further explored in order to derive useful information for their evolution. This is done by comparing the topic distributions of consecutive weeks using Equation 1 and then turning points can be revealed if the following inequality is justified:

Child abuse in Belgium	Tropical storm Edouard	Peace talks in Palestina	Kurdish war in Iraq
child	storm	israel	iraq
police	hurricane	peace	iraqi
woman	north	israeli	iran
death	wind	netanyahu	kurdish
family	west	minister	turkey
girl	mph	palestinian	northern
murder	mile	arafat	arbil
dutroux	coast	talk	baghdad
body	move	government	force
sex	flood	west	united

Table 3: Extracted topics reveal events from August and September 1996

$$diff_i >= thres \tag{2}$$

where $thres$ is a user-defined threshold which is set to the first quartile (Q1) (i.e. the middle number between the smallest and the median) of all $diff_i$ values for all topics T at the first time step (i.e. $diff_i(1)$). This is justified due to the fact that depending on the corpus collection used, topic cohesion can vary from experiment to experiment. Other values were also tried (median, mean, 3rd quartile) but they proved to show very few changes in the storylines.

Table 4 shows the differences in the top 20 words of the word distributions for one example topic (about Iraq). Inspecting the top articles for this topic reveals an evolvement of the story behind the topic, as the main articles in the first weeks talk about the threat imposed by Iraqi forces and air strike battles, while the last weeks talk about concrete U.S. troop deployment in Kuwait. Table 5 presents the headlines of the corresponding articles for verification. While the first weeks the similarity between the distribution is almost identical (less than 0.01 difference), difference between week 3 and week 4 is significant (more than 0.02) and thus reflects this “turning point” within the same topic.

week 1	week 4	week 5
iraq	iraq	iraq
missile	missile	gulf
attack	gulf	kuwait
saudi	iraqi	military
iraqi	military	missile
military	kuwait	iraqi
gulf	attack	united
force	united	force
united	force	saudi
war	zone	zone
defense	saddam	troops
air	defense	war
kuwait	war	attack
zone	saudi	defense
arab	air	washington
official	strike	arab
arabia	southern	official
strike	official	air
saddam	troops	saddam
southern	washington	arabia

Table 4: Word distribution top word differences for Iraq topic

week 1 - 3	week 4	week 5 - 6
Perry cites two incidents in Iraq no-fly zone.	Iraq fires at U.S. jets, U.S. bombers move closer.	U.S. boosts Kuwait defence by deploying Patriots.
U.S. warns it will protect pilots over Iraq.	U.S. gets Kuwaiti approval for troops deployment.	U.S. ground troops set to fly to Gulf.
Defiant Saddam urges his warplanes to resist U.S.	Kuwait agrees new troop U.S. deployment.	U.S. carrier enters Gulf, troops land in Kuwait.
Saddam urges his warplanes and gunners to resist.	Iraq says fired missiles at US and allied planes.	U.S. sends last of 3,000 ground troops to Gulf.
U.S. launches new attack on Iraq - officials.	Iraq fires at U.S. jets, U.S. bombers move closer.	U.S. declines to rule out Iraq strikes.

Table 5: Article headlines for top documents of Iraq topic

Moreover, visualization works in a way that similar topics are on top of each other in the graph. Exploration of nearby topics can reveal further events within similar storylines. Table 6 shows the cosine similarity between two very similar topics (Iraq and Kurdish civil war) along the time line, while Table 7 gives an overview of the topic contents, represented by the top 20 words for each topic, at the time point with the highest similarity. The highest value for the cosine similarity, namely 0.613, can be found at time step 3 for two topics talking about the conflicts, the Iraq was involved in late 1996. Given these thresholds, both topics could be clustered further to a more general topic about Iraq politics, thus allowing for detecting the general storyline concept or the trend around these issues (if similarity threshold is high, then the current trend for these topics is low).

week 1	week 2	week 3	week 4	week 5	week 6
0.559	0.594	0.613	0.568	0.472	0.397

Table 6: Topic cosine similarities for both topics, Iraq and Kurdish Civil War, for each time step

Iraq topic	Kurdish civil war topic
iraq	iraq
missile	iraqi
attack	kurdish
iraqi	iran
military	northern
gulf	turkey
united	turkish
saddam	arbil
force	baghdad
zone	kdp
strike	united
kuwait	kurdistan
air	saddam
saudi	iranian
defense	puk
war	force
southern	official
baghdad	troops
action	border
official	kurds

Table 7: Word distribution top word similarities for both topics, Iraq and Kurdish Civil War, at week 3

Finally, an example of some topics of summer 1996 and their presence (in terms of percentage of documents that the equivalent topic distribution is non-zero) is shown in Figure 4. One can identify topics that are recurring and present turning points (like the “Russia-1”) which has two major hits or topics that have more bursty presence (like the “Olivetti” case in Italy or the “Tennis Open”). It should also be noticed the effect of topics that cover different stories under the same arc (e.g. the “plane crash” topics is already present in the news (referring mostly to TWA800 flight accident but it becomes more prevalent once a new plane crash in Russia (Vnukovo2801 flight) occurs, which also boosts the “Russia-1” since they are overlapping). These experiments reveal the ability of the system to identify turning points in storylines and track their presence and evolvement.

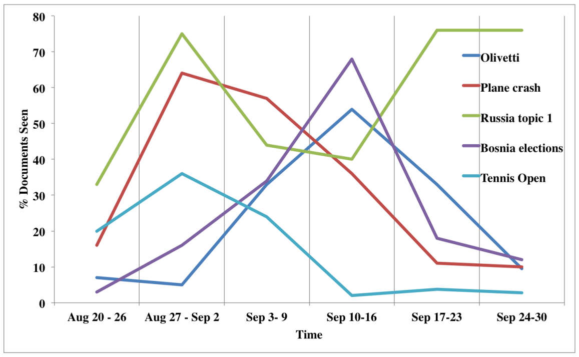


Figure 4: Emerging topics and turning points example

5 Discussion and Conclusion

This paper presented a Dynamic Latent Dirichlet Allocation for detecting storylines and monitor their development through time by revealing trends and similarities between evolved topics. The proposed approach was applied to news items of 6 weeks in August and September 1996 of the Reuters corpus RCV1. After applying careful preprocessing, it was possible to identify some of the main events happening at that time (e.g. the Kurdish civil war or the horrible crimes in Belgium). In order to identify details and possible turning points of a topic, a second step of comparing the word distributions inside each topic at each time step is added. Similarly, topics can also be aggregated revealing trends and arcs under the same storyline. Moreover, “burstiness” of topics can be detected and used for identifying new or recurring events.

Results from the RCV1 corpus subset reveal the possibilities of monitoring storylines and their evolvement through time and the opportunities for detecting turning points or identifying several sub-stories. Visualization of the results and the interaction with the stacked graph provide a framework for better monitoring the storylines. Further work involves the application of the model to the whole RCV1 corpus, as well as to the actual Reuters 2015 archive and develop a formal way to identify turning points and aggregate similar topics under a storyline. Moreover, evaluation of the output using human storyline evaluations will further improve model coherence and interpretation as well as validate the effect of the approach as to if identified storylines were correctly detected by the algorithm

References

Amr Ahmed and Eric P Xing. 2012. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*.

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report.

Arindam Banerjee and Sugato Basu. 2007. Topic models over text streams: A study of batch and online unsupervised learning. In *SDM*, volume 7, pages 437–442. SIAM.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Bin Cao, Dou Shen, Jian-Tao Sun, Xuanhui Wang, Qiang Yang, and Zheng Chen. 2007. Detect and track latent factors with online nonnegative matrix factorization. *IJCAI*, page 26892694.

Avinava Dubey, Ahmed Hefny, Sinead Williamson, and Eric P Xing. 2013. A nonparametric mixture model for topic modeling over time. In *SDM*, pages 530–538. SIAM.

Jonathan G. Fiscus and George R. Doddington. 2002. Topic detection and tracking. chapter Topic Detection and Tracking Evaluation Overview, pages 17–31. Kluwer Academic Publishers, Norwell, MA, USA.

Daniel Godfrey, Caley Johns, Carl Meyer, Shaina Race, and Carol Sadek. 2014. A case study in text mining: Interpreting twitter data from world cup tweets. *arXiv preprint arXiv:1408.5427*.

Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. 2011. A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 832–840. ACM.

D. D. Lee and H. S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, page 788791.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res.*, 5:361–397, December.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, March.

Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1408–1417. Association for Computational Linguistics.

Ankan Saha and Vikas Sindhwani. 2012. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. *Proceedings of the fifth ACM international conference on Web search and data mining*, page 693702.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Suvrit Sra and Inderjit S Dhillon. 2005. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in neural information processing systems*, pages 283–290.

Michael Tannenbaum, Andrej Fischer, and Johannes C. Scholtes. 2015. Dynamic Topic Detection and Tracking using Non-negative Matrix Factorization. In *Proceedings of the 27th Benelux Artificial Intelligence Conference (BNAIC)*. BNAIC.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on*

Knowledge discovery and data mining, pages 424–433. ACM.

Chong Wang, David M. Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In David A. McAllester and Petri Myllymki, editors, *UAI*, pages 579–586. AUAI Press.

Fei Wang, Ping Li, and Arnd Christian König. 2011. Efficient document clustering via online nonnegative matrix factorizations. In *SDM*, volume 11, pages 908–919. SIAM.

Xing Wei, Jimeng Sun, and Xuerui Wang. 2007. Dynamic mixture models for multiple time-series. In Manuela M. Veloso, editor, *IJCAI*, pages 2909–2914.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM.