

# Towards a Topic Discovery and Tracking System with Application to News Items

Daniel Brüggermann, Yannik Hermeij, Carsten Orth, Darius Schneider, Stefan Selzer, and Gerasimos Spanakis<sup>(✉)</sup>

Department of Data Science and Knowledge Engineering,  
Maastricht University, Maastricht 6200MD, Netherlands  
{d.bruggermann,y.hermeij,c.orth,d.schneider,  
s.selzer,jerry.spanakis}@maastrichtuniversity.nl

**Abstract.** Rapid proliferation of the World Wide Web led to an enormous increase in the availability of textual corpora. In this paper, the problem of topic detection and tracking is considered with application to news items. The proposed approach explores two algorithms (Non-Negative Matrix Factorization and 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, emphasis is given to the visualization and interaction with the results through the implementation of a graphical tool (regardless the approach). Experimental analysis on Reuters RCV1 corpus and the Reuters 2015 archive reveals that explored approaches can be effectively used as tools for identifying topic appearances and their evolutions while at the same time allowing for an efficient visualization.

**Keywords:** Topic detection · Topic tracking · Non-negative matrix factorization · Dynamic latent dirichlet allocation

## 1 Introduction

With the rapid progress of computer technology in recent years, there has been an enormous increase in the availability of electronic information on the Internet. However, these vast amounts of data (emails, news sources, forums, etc.) are practically impossible to be harnessed by any human user in their primitive forms. The problem will only deteriorate without the development of approaches to deal with this information overload.

Different stakeholders (companies, individuals, policy makers, etc.) would be interested to harness the amount of free text data available in the web in order to make timely decisions, monitor information and support further actions. Topic detection and tracking in large textual streams (e.g. news items over long time spans) can assist in these efforts as it will reveal trends and evolutions in the underlying storylines. There are many techniques about topic extraction like

---

Authors contributed equally to the manuscript, thus appear in alphabetical order.

Nonnegative Matrix Factorization (NMF) [18] or Latent Dirichlet Allocation (LDA) [5] but there are not many extensions to dynamic data handling. Time dependent modeling of documents can be computationally expensive and complex [2] 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, there is much space for developing more effective approaches in terms of both computation and visualization of the results.

This research work aims at exploring how machine learning techniques can be utilized in order to detect topics (and subsequently trends or storylines) from a news items flow through time. Results are visualized and evaluated using the (fully annotated and freely available) RCV1 Reuters corpus (810.000 documents) and the actual Reuters 2015 Reuters archive (obtained by the authors). Section 2 presents an overview of current research work in the area. The NMF based approach is described in Sect. 3 and the DLDA approach in Sect. 4. Section 5 presents the experimental results and finally, Sect. 6 concludes the paper.

## 2 Related Work

Topic detection and tracking is a long studied task [8] 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 [11]. 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 a dynamic approach, processing data in chunks [21]. [6] 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. [17] 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). [19] 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 [14, 24], 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 [10].

Latent Dirichlet Allocation (LDA) [5] 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.

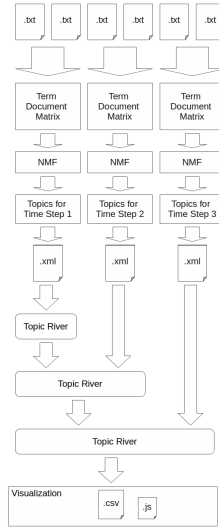
LDA has been extended in order to handle documents over long periods and many variations exist. Other approaches have been proposed as well [3] but scalability is an issue and visualization is not feasible. A milestone in the area was the work of [22] since they associated a beta distribution over time with each topic to capture its popularity. There are also nonparametric models developed either using Gaussian mixture distributions [1] or utilizing Markovian assumptions [7]. 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 [23]. But according to [20], 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 [4] that after examining the generated distributions for changes that denote turning points or arcs. Finally, results are visualized using a stacked graph modeling and can be explored in an intuitive way by relating one topic to another.

### 3 The Proposed NMF-Based Pipeline

Instead of modifying the NMF algorithm, this section elaborates on an algorithm that applies NMF on time steps with arbitrary granularity, and then uses a metric to evaluate the similarity between topics generated for these time steps. This way, topics can also fade away if they do not have a representative at a certain time step, and new topics can emerge if a topic is not similar enough to any existing topic. The similarity metric is intuitively easy to understand and less complex than the related work. Moreover, results can be visualized at any step. The process of NMF over time is described in Fig. 1 and is described by the following phases:

- Phase 1:
  - Define the time step unit (day, week, etc.) and extract texts from this equivalent time step.
  - Define the number of topics to be generated in each time step (a very low number might lead to a mix of multiple topics with insufficient separation whereas a very high number might create topics too specific and increase computational time)
- Phase 2:
  - Run NMF algorithm for each time step  $s$  ( $[T_s \times D_s] \approx [T_s \times H_s][H_s \times D_s]$ ) which decomposes the term-document matrix ( $T \times D$ ) to discover the hidden dimensions or topics ( $H$ )



**Fig. 1.** Work flow of the NMF Code. Example with 6 text files and 3 time steps.

- Convert the term-topic ( $H_s^T$ ) and topic-document ( $H_s^D$ ) matrices to ranked topic instances which store the term and document list (ranking is done by the values of the equivalent matrices)
  - Normalize values over all terms/documents in order to compare NMF results from discrete time steps (vocabulary and number of documents can change per time step)
- Phase 3:

- Use a metric to identify if a time step’s topic matches any topic from the preceding time steps: A topic is defined by its terms, and following Zipf’s law [15], only a low number of terms have significant relevance to a topic. Thus, comparing the top- $N$ -ranked terms (where  $N$  is set to 20 but can be adjusted by the user) of two topics, and defining a threshold  $\epsilon$  for the similarity to be enough to define one of the topic as successor of the other, creates a *topic wave* over time ( $\epsilon$  was set to 0.5 but can be adjusted by the user). Comparison between any topics  $H1$  and  $H2$  in any two timesteps  $s_1$  and  $s_2$  is possible through the  $H_s^T$  vectors and can be computed as:

$$dist(H1, H2) = \sum_i \sqrt{(H1_{s1,i}^T - H2_{s2,i}^T)^2}$$

where  $i$  denotes the common terms between the two topics examined.

A new topic wave begins at this time step if no previous topic is similar enough (thus if  $dist > \epsilon$ ). Topic waves fade away if no topic is found at a time step that is similar enough to the topic wave’s most relevant terms. It has to be noticed that a topic wave can reappear; topics of a time step are always compared against all existing waves, not just the ones that appeared in the last time step. This intuitively makes sense, e.g. for

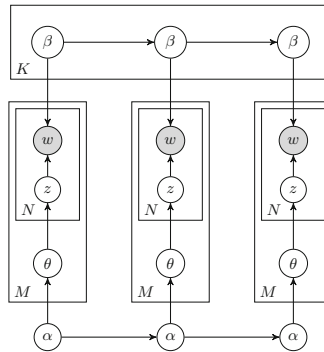
recurrent topics like the FIFA World Cup, that are barely relevant outside of their time frame.

- Topic waves from all time steps generate a *topic river* which then can be used to visualize the result. The relevance of all topic waves at a given time step sum up to 100%, and the sum of all topic waves over all time steps defines the topic river.
- Repeat Phases 2 and 3 for the equivalent number of time steps

## 4 The Proposed DLDA Approach

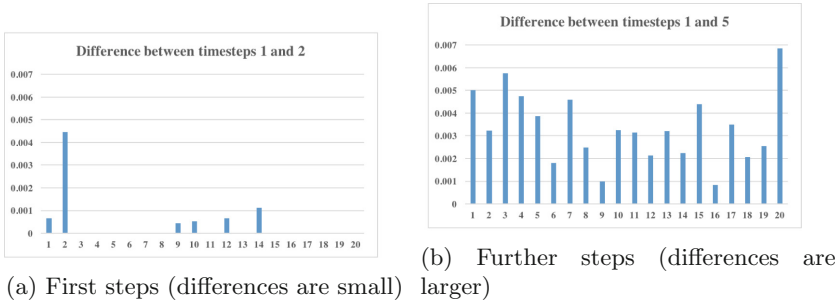
### 4.1 Dynamic Latent Dirichlet Allocation (DLDA)

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 2 shows a graphical representation of the dynamic topic model.



**Fig. 2.** 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 ( $z$  denotes the word distribution,  $\theta$  denotes the topic distribution,  $N$  is the total number of words per document,  $M$  is the total number of documents and  $K$  is the total number of topics. Additionally, each topic’s parameters  $\alpha$  and  $\beta$  evolve over time [4])

DLDA (as well 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. Moreover, the timestep has to be determined at this point. This again can be set to any time unit. For example,



**Fig. 3.** Example differences in word distributions for a topic using top-20 words

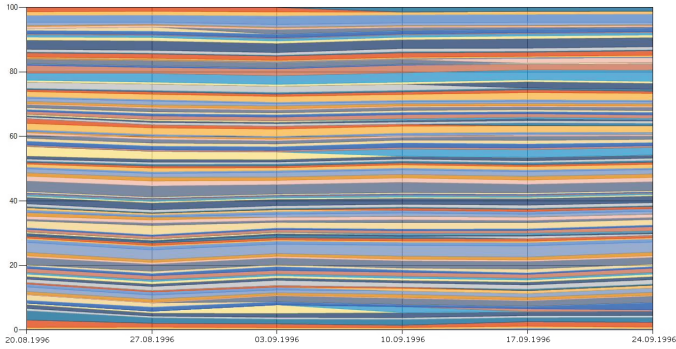
the RCV1 corpus used here (July and August of 1996) contains 42 days which makes exactly 6 weeks of time. The dynamic topic model is accordingly applied to 6 time steps corresponding to the 6 weeks of the data set.

## 4.2 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 3 shows the word distribution score differences for two different timesteps and for the same topic. 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. This means that when a change in the topic distribution is not that big to create a new topic, then it is checked whether it could be that a sub-topic or a change under the same topic can be assessed.

If a significant change is identified using this similarity measure, then significant events within the topics (or storylines) are noticed. These changes in the storylines can be detected and visualized by a topic river like the one in Fig. 4. 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.



**Fig. 4.** Topic rivers for August and September 1996 for emerging topics: x-axis represents time units (weeks) and y-axis represents normalized topic prevalence (sums to 100) (Color figure online)

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

## 5 Experiments

### 5.1 Datasets and Preprocessing

The Reuters Corpus RCV1 is used as a benchmark dataset [12]. It consists of roughly 800.000 news articles from the time of August 1996 to August 1997. The topic assignments for the articles make use of 103 out of 126 originally provided topics. These topics form a tree hierarchy, thus documents typically have multiple labels. Moreover, an actual copy from Reuters 2015 archive (roughly 330.000 news articles) was extracted<sup>1</sup> using a webcrawler containing the URL, time, headline, description, contents and tags.

Both algorithms (NMF and DLDA) can be very computationally expensive in terms of time and memory. The size of the vocabulary should be reduced to handle memory issues, preferably without loss of quality. The steps followed for preprocessing are: stemming (using Porter’s Stemmer [16]), stopword removal, americanization (reducing words to american spelled ones utilizing code from Stanford Core NLP’s tool [13]), lemmatization (also based on Stanford Core NLP tool), low occurrence removal (deleting terms used by less than five documents). This process leads to a significant reduction in the vocabulary, i.e. from 308.854 distinct tokens we are left with a final vocabulary of 131.202 unique words.

Both algorithms rely on the selection of number of topics and the time unit. In our case week is selected as time unit and number of topics was either compared

<sup>1</sup> <http://www.reuters.com/resources/archive/us/>.

to the original Reuters categorization or to an arbitrary number (e.g. 30). In any case, selection of these parameters relies on the use case of the system (a smaller number of topics will lead to more changes in the topic distributions) but developing automated approaches for these parameters is one of the future work directions.

## 5.2 Implementation and Visualization Details

All topics of a time step are stored in a Java class and serialized as an XML file. This way, the most time-intensive part of the algorithm can be separated from generating visualization data. The XML files can be loaded into the program as soon as all desired data files are processed. Data is visualized using the .NET framework visualization library<sup>2</sup> in a form of a normalized stacked area graph: For each topic there is a value (Y-axis) for each time step (X-axis) that represents the prevalence of that topic. These values are normalized for each time unit and then displayed in a graph, where similar topics are next to each other. Graph is interactive in terms of offering a zoom function, tooltips to analyze topics in a certain range and a connection to view relevant documents.

## 5.3 NMF Results

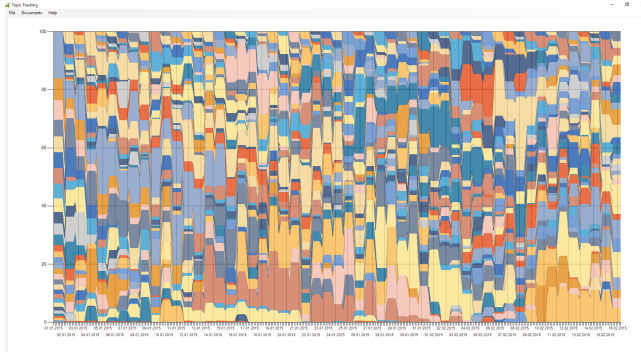
With the NMF algorithm, we were able to calculate topics from the annotated Reuters Corpus of the year 20.08.1996–19.08.1997. We chose “week” as a time unit (7 days of text data). We also calculated topics for the whole year of 2015 (around 336.000 articles), which is not annotated, but represents more recent data which can be easier checked manually for coverage of all recent events from that year. We generated 30 topics for week-wise intervals and visualized the results using the implemented charting tool and results are presented in Fig. 5.

Table 1 shows different prominent topics of 2015, that were reflected in topics generated by our algorithm. Table 2 shows an example for a (short) topic wave, present in four time steps. The first two topics in January follow the terror attack in Paris on the magazine staff of *Charlie Hebdo* at January 7th. The third topic is in February and covers a terror attack of the terror group Houthis, located in Yemen. Finally, the last topic represents a terror attack on tourists in Tunisia in June 2015. The similarity between the topics within the topic wave can thus be confirmed by the actual events.

An overview of all topics, reveals that the most relevant topics usually are of business-related nature, made up of general financial terms like “stock market”, “import” and “share”, and not very specific to the corresponding time period. This appears to be an attribute of the test data, the Reuters Corpus, that has a dedicated section of articles that contain financial content, thus creating high TF-IDF values for these terms that are often used in this small section of documents. The same problem on a smaller scale exists for weather forecasts.

<sup>2</sup> <https://msdn.microsoft.com/en-us/library/dd456632.aspx>.





**Fig. 5.** Visualization of 30 topics from year 2015: x-axis represents time units (weeks) and y-axis represents normalized topic prevalence (sums to 100)

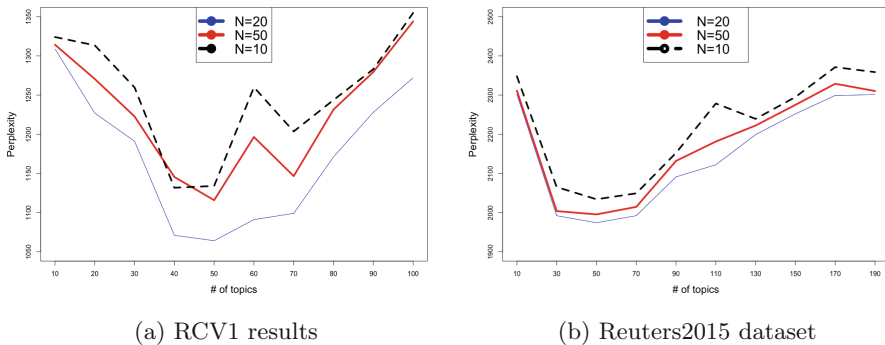
**Table 1.** Topics and their most relevant terms in 2015 (topic labels are manually selected)

|                         |                       |                 |                |                     |
|-------------------------|-----------------------|-----------------|----------------|---------------------|
| Greece Financial Crisis | Charlie Hebdo Attacks | FiFA Corruption | Ukraine Crisis | VW Emission Scandal |
| greeks                  | attacker              | fifa            | ukraine        | vws                 |
| greece                  | paris                 | soccer          | russia         | piech               |
| syryza                  | charlie               | cups            | russians       | winterkorn          |
| tsipras                 | hebdo                 | blatter         | sanctions      | ceos                |
| euros                   | police                | corruption      | moscow         | porsche             |

**Table 2.** Evolution of a topic: terror attacks in Paris, Jemen, Tunisia

|          |           |          |           |
|----------|-----------|----------|-----------|
| Jan 08   | Jan 15    | Feb 12   | Jun 25    |
| attacker | police    | attacker | tunisia   |
| paris    | charlie   | houthis  | attacker  |
| hebdo    | french    | police   | tourists  |
| police   | attacker  | boko     | tunisians |
| french   | hebdo     | haram    | hotels    |
| islamic  | paris     | killed   | tourism   |
| france   | muslims   | islamic  | beaches   |
| killed   | cartoons  | yemen    | islamists |
| muslims  | islamists | militant | killed    |
| kouachis | killed    | libya    | sousse    |

Following common practice for comparing topic models, we use perplexity of the held-out test data as our goodness-of-fit measure. Perplexity is defined as:  $exp\left(-\frac{\sum_{d=1}^D \sum_{i=1}^{N_d} \log p(w_{d,i}|M)}{\sum_{d=1}^D N_d}\right)$  where  $w_{d,i}$  represents the  $i$ -th term in document  $d$ ,  $M$  is the model and  $N_d$  is the number of words in document  $d$ . In order to allow comparisons, all models were trained using a real-world scenario (using news items of the first 8 months of 2015 as training set and the last 4 months as testing set). As the number of topics is increased beyond that minimum, overfitting appears to set in, as was also observed in other literature work [9]. Results can be found in Fig. 6 where also the effect of parameter  $N$  (number of top-words considered for comparison between topics) is presented and the choice of  $N = 20$  is justified. The effect of  $N$  and  $\epsilon$  was explored using a grid search approach for different values: low value of  $\epsilon$  leads to many new topics appearing at each time step which might reduce perplexity but results are not robust whereas high value of  $\epsilon$  leads to stable topic rivers where not many new topics appear and that increases perplexity.  $\epsilon$  value of 0.5 (combined with  $N = 20$ ) was found to produce the best results.



**Fig. 6.** Perplexity results for NMF-based algorithm

In order to evaluate the retrieval performance, a matching between the topics extracted by NMF and the RCV1 categories is attempted. Both NMF and RCV1 provide multi-labeled instances (each document is “assigned” to multiple topics), which leads to difficulties in assessing performance in a comparative way but treating each label for each document as a separate instance overcomes this issue. Then, precision and recall for each pair of NMF and RCV1-annotated topics are computed. The F-score is 0.496, derived from a precision of 0.521 and recall of 0.473, denoting that 103 topics over time does not lead to satisfactory results (as also suggested from Fig. 6). Further reducing the number of topics, leads to an increase in F-score of around 20% considering a merging of similar RCV1 annotated categories which gives promising directions for the proposed method.

## 5.4 LDA Results for Story Detection

Due to time limitations, the dynamic topic model was applied to only two months of the RCV1 corpus (Summer 1996). It was able to partially identify 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.

**Table 3.** Extracted topics reveal events from August and September 1996

|                        |                        |                          |                     |
|------------------------|------------------------|--------------------------|---------------------|
| 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              |

These topics describe events over a the period of two months and their change during the examined time frame (2months) can be further explored in order to derive useful information for their evolution. 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 evolution 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. Exploring the differences in the topic distributions between these weeks can help identify this change in the events. 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.

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. Following the same process as for NMF, we evaluated the performance against RCV1 predefined categories and that gave an F-score of 0.288, derived from a precision of 0.432 and recall of 0.216. Low scores (both for NMF and DLDA) are explained due to the fact that RCV1 corpus contains many unlabeled documents and that the nature of two tasks (classification versus topic detection) is different.

**Table 4.** Word distribution top word differences for Iraq 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 5.** Article headlines for top documents of 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           |

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 Fig. 7. One can identify topics that are recurring and present turning points (like the “Russia-1”) which has two major hits or topics that have

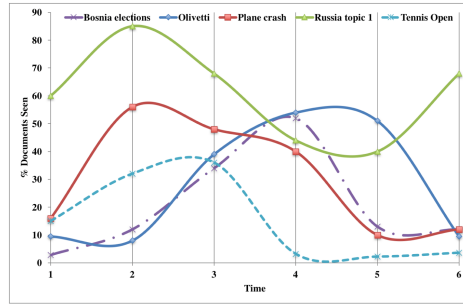


Fig. 7. Emerging topics and turning points example

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 evolution.

## 6 Discussion and Conclusion

This paper presented two algorithmic approaches to the topic detection and tracking from news items corpora. The first approach applied NMF on separate time steps of the data and then connected using a similarity metric. By using extensive cleaning of the vocabulary during pre-processing, a fast data processing algorithm was developed and the proposed pipeline is more simplified than literature work but still effective. The generated topics can easily be identified by their most relevant terms and associated with events happening in the corresponding time period. The second approach utilized a dynamic version of LDA and was able to identify some of the main events happening at the summer of 1996 (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.

Summarizing the two approaches, the DLDA method produces topics, that are on a more generalized level. A high frequency of topic evolution can not be seen here but the algorithm can detect turning points within the developing stories. On the other hand, the NMF approach is significantly faster and can effectively detect new stories, track them until they disappear (and detect them again if they re-appear). Changes in the topics appear to be fast enough regardless the threshold set for detecting changes showing the evolution here is large.

Further evaluation of two approaches against each other is needed in order to identify these differences in a microscopical level.

Results reveal the possibilities of monitoring storylines and their evolution 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. Besides the detailed comparison of the two algorithms, future work involves developing approaches that will validate the results and allow for more comprehensive and sophisticated topic descriptions. Moreover, automated approaches for determining the number of topics and the time unit will be explored. This will give the opportunity to create stories or summaries based on the topic waves created by the system.

## References

1. Ahmed, A., Xing, E.P.: Timeline: a dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. arXiv preprint [arXiv:1203.3463](https://arxiv.org/abs/1203.3463) (2012)
2. Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study final report (1998)
3. Banerjee, A., Basu, S.: Topic models over text streams: a study of batch and online unsupervised learning. In: SDM, vol. 7, pp. 437–442. SIAM (2007)
4. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 113–120. ACM, New York (2006)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
6. Cao, B., Shen, D., Sun, J.-T., Wang, X., Yang, Q., Chen, Z.: Detect and track latent factors with online nonnegative matrix factorization. In: IJCAI, pp. 2689–2694 (2007)
7. Dubey, A., Hefny, A., Williamson, S., Xing, E.P.: A nonparametric mixture model for topic modeling over time. In: SDM, pp. 530–538. SIAM (2013)
8. Fiscus, J.G., Doddington, G.R.: Topic detection and tracking evaluation overview. In: Topic Detection and Tracking, pp. 17–31. Kluwer Academic Publishers, Norwell (2002)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proc. Nat. Acad. Sci.* **101**(suppl 1), 5228–5235 (2004)
10. Hong, L., Dom, B., Gurumurthy, S., Tsioutsoulouklis, K.: A time-dependent topic model for multiple text streams. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 832–840. ACM (2011)
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999)
12. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: RCV1: a new benchmark collection for text categorization research. *J. Mach. Learn. Res.* **5**, 361–397 (2004)
13. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60 (2014)

14. Paul, M., Girju, R.: 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, vol. 3, pp. 1408–1417. Association for Computational Linguistics (2009)
15. Piantadosi, S.T.: Zipfs word frequency law in natural language: a critical review and future directions. *Psychon. Bull. Rev.* **21**(5), 1112–1130 (2014)
16. Porter, M.F.: An algorithm for suffix stripping. *Program* **14**(3), 130–137 (1980)
17. Saha, A., Sindhvani, V.: Learning evolving and emerging topics in social media: a dynamic NMF approach with temporal regularization. In: Proceedings of the fifth ACM International conference on Web Search and Data Mining, pp. 693–702 (2012)
18. Sra, S., Dhillon, I.S.: Generalized nonnegative matrix approximations with bregman divergences. In: *Advances in Neural Information Processing Systems*, pp. 283–290 (2005)
19. Tannenbaum, M., Fischer, A., Scholtes, J.C.: Dynamic topic detection and tracking using non-negative matrix factorization. In: Proceedings of the 27th Benelux Artificial Intelligence Conference (BNAIC). BNAIC (2015)
20. Wang, C., Blei, D.M., Heckerman, D.: Continuous time dynamic topic models. In: McAllester, D.A., Myllymki, P. (eds.), *UAI*, pp. 579–586. AUAI Press (2008)
21. Wang, F., Li, P., Christian König, A.: Efficient document clustering via online nonnegative matrix factorizations. In: *SDM*, vol. 11, pp. 908–919. SIAM (2011)
22. Wang, X., McCallum, A.: 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, pp. 424–433. ACM (2006)
23. Wei, X., Sun, J., Wang, X.: Dynamic mixture models for multiple time-series. In: Veloso, M.M. (ed.) *IJCAI*, pp. 2909–2914 (2007)
24. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 177–186. ACM (2011)