2017

Comparison of Two-pass Algorithms for Dynamic Topic Modelling Based on Matrix Decompositions

John Cardiff
Technological University Dublin, john.cardiff@tudublin.ie

Gabriella Skitalinskaya
Technological University Dublin

Mikhail Alexandrov
Autonomous University of Barcelona

Recommended Citation

SSkitalinskaya G., Alexandrov M., Cardiff J. (2018) Comparison of two-pass algorithms for dynamic topic modeling based on matrix decompositions. In: (Castro F., Miranda-Jiménez S., González-Mendoza M. (eds)) *Advances in Computational Intelligence. MICAI 2017. Lecture Notes in Computer Science*, vol 10633. Springer. doi:10.1007/978-3-030-02840-4_3

Comparison of two-pass algorithms for dynamic topic modeling based on matrix decompositions

Gabriella Skitalinskaya[1,2,4], Mikhail Alexandrov[3,4], and John Cardiff[1]

[1]Institute of Technology, Tallaght, Dublin
[2]Moscow Institute of Physics and Technology (State University), Russia
[3]Autonomous University of Barcelona, Spain
[4]Russian Presidential Academy of National Economy and Public Administration, Russia
gabriellasky@icloud.com; malexandrov@mail.ru; john.cardiff@ittdublin.ie

Abstract. In this paper we present a two-pass algorithm based on different matrix decompositions, such as LSI, PCA, ICA and NMF, which allows tracking of the evolution of topics over time. The proposed dynamic topic models as output give an easily interpreted overview of topics found in a sequentially organized set of documents that does not require further processing. Each topic is presented by a user-specified number of top-terms. Such an approach to topic modeling if applied to, for example, a news article data set, can be convenient and useful for economists, sociologists, political scientists. The proposed approach allows to achieve results comparable to those obtained using complex probabilistic models, such as LDA.

Keywords: dynamic topic modeling, matrix decomposition, latent Dirichlet allocation

1 Introduction

1.1 Problem setting

In recent years, there has been a sharp increase in the popularity and development of methods for extracting hidden topics in texts. Topic modeling is an approach that allows users to explore collections of text documents, search and analyse information based on topics covered in the documents. Algorithms of topic modeling allow the determination of topics that are covered in the collection of articles, and present the result in a way that enables simple navigation in the corpus of documents using the found topics. Dynamic topic models can track how a particular topic has changed over a certain period of time, for example in months or years, and how it is related to other topics. Thus, dynamic topic modeling can serve as an addition to static modeling, which is associated only with the identification of a set of topics outside the context of time.

In this paper, we propose a dynamic topic model with different matrix decompositions such as latent semantic indexing (LSI), principal component analysis (PCA), independent component analysis (ICA) and non-negative matrix factorization (NMF), which captures the evolution of topics in a sequentially organized corpus of documents. The proposed algorithm allows to achieve results comparable to those obtained using complex probabilistic models, such as latent Dirichlet allocation (LDA), but with less resources and faster. We demonstrate the algorithms applicability by analyzing news articles in the Russian language obtained from [1], that have been published during the year 2016. Under this model, articles are grouped by month, and from each months articles we retrieve a set of topics that evolve throughout the year.

The paper is organized as follows. Section 2 describes the proposed two-pass algorithm and model quality evaluation measures. Section 3 provides information on the dataset used for topic modeling. Sections 4 and 5 provide a comparison of results obtained by different methods during the first pass and second pass of the proposed approach. It is demonstrated how the proposed dynamic topic models allow the exploration of a large document collection distributed in time. Finally, Section 6 presents our conclusions.

1.2 State-of-the-art

Topic models are aimed at finding hidden semantic structures or themes in textual data that can be obtained from word coincidences in documents. These models date back to an early work on LSI [2], which proposes the decomposition of Term-Document matrices using singular value decomposition. Here each singular vector is considered as a topic and the most interesting topics are associated with the first singular values.

A lot of studies on topic modeling have focused on the use of probabilistic methods, where the topic is defined as a discrete distribution on a set of terms, and each document as a discrete distribution on a set of topics [3]. The most widely used method of topic modeling is the Latent Dirichlet Allocation (LDA), proposed in [4]. LDA greatly influenced the field of natural language processing and statistical machine learning and inspired a number of research papers in this direction. In particular, authors consider using word associations [5] and topic correlations [6]. The composition of probabalistic models and LSI is considered in [7], where probabilistic latent semantic analysis is introduced. Ambiguity is a distinctive feature of probabilistic models and researchers propose various methods of regularization to reduce this effect [8, 9].

Algorithms based on matrix decompositions, such as NMF, ICA and PCA were also effective in detecting the main topics in a corpus of texts [10, 11].

There are different approaches to modeling the evolution of topics over time. Here, two basic ideas should be distinguished: (a) models that assume topics to be static semantic concepts that are used uniquely over the period of analysis and (b) models that allow for a dynamic change of topics by modeling changes in the word sets describing the topic over time. We are more interested in models that fall in the second category.

The Dynamic Topic Model (DTM) introduced in [12] is the work to which we compare our research. Here, topics cease to generalize semantic concepts that undergo some change. Other related work that is based on this assumption was done by [13, 14]. The approach proposed by [15] uses a parametric model, which allows us to find topics by linking them to timestamps. Here each topic is associated with a continuous distribution over timestamps, and for each generated document, the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp.

It was noted that there were few publications dedicated to dynamic topic models based on simpler methods such as LSI, PCA, ICA, NMF. The authors of [16] propose two-layer NMF methodology for identifying topics in large political speech corpora over time and apply it to a corpus of speeches of Members of the European Parliament. The obtained results proved to be semantically more coherent when compared with LDA. The approach proposed in this paper allows to track the evolution of topics over time in a sequentially organized set of documents. This approach was introduced in our work [17] but the results were not compared to appropriate baselines and such matrix decompositions as PCA and ICA were not considered as well.

2 Approach

2.1 Methods based on matrix decompositions

In the paper we consider four methods LSI, PCA, ICA and NMF. All methods take as input a bag of words matrix (i.e. each document represented as row, with each column containing the word frequency in the collection). TF-IDF term weighting and document length normalization is applied to the bag of words matrix to filter frequent words that provide the most information about the document The mentioned decomposition methods are applied to document collections from non-overlapping time windows.

PCA is based on second-order statistics and ICA exploits inherently non-Gaussian features of the data and employs higher moments. PCA minimizes the covariance of the data, while ICA minimizes higher-order statistics such as fourth-order cummulant, thus minimizing the mutual information of the output. LSI is very similar to PCA, but differs in that it works on sample matrices directly instead of their covariance matrices. NMF is an alternative approach to decomposition that assumes that the data and the components are non-negative. Unlike PCA, the representation of a vector is obtained in an additive fashion, by superimposing the components, without subtracting.

2.2 Two-pass algorithm

First of all, for the purposes of the present paper the following definitions are used. The entire time interval is a sequence of disjoint time windows. Each window contains related documents. Each document may reflect one or more topics. Each topic is represented by its top-terms. Top-terms are terms that have the highest frequency (on average) in those documents that contain the topic. The number of top-terms for all topics, regardless of the time window, is the same and is assigned by the user (for example, 10, 20, 30, etc.). When applying matrix decompositions to each time window the user must specify the number of topics. One of the quality measures that allows us to choose the best number of topics is the so-called coherence measure.

The main hypothesis of the approach is that topics from different time windows, which share the same general topic, will have similar sets of top-terms describing them. So by reapplying matrix decompositions to an aggregated Topic-Term matrix from all time windows we obtain dynamic topics, which are related to a set of window topics.

The approach is represented by the following algorithm:

First pass. One of the indicated methods (LSI, PCA, ICA or NMF) is applied to each time window. As a result, for each window a set of k topics is obtained, where k is defined by the user. Topics are described by a user-specified number of top-terms t and a set of all related documents.

Data Transformation. Using the topic models obtained after the first pass we construct a new compressed representation, looking through the rows of each Topic-Term matrix of each window topic model. Each row contains weights of all the terms of a particular topic of the time window under consideration. We construct the new Topic-Term matrix with two subsequent procedures:

(1) In each topic from each window topic model, the top-t terms are taken from the appropriate topic-term matrix, all weights for the remaining terms are set to 0.
(2) The obtained vectors for all window topic models from all time windows are combined into one matrix.

Second pass. The considered topic modeling methods are re-applied to the transformed data, outputting a set of dynamic topics, each of which has a set of window topics associated with it. Applying matrix decompositions in this step, we identify k' dynamic topics that potentially span several time windows. The number of dynamic topics k' to be found in this step is specified by the user.

The matrix has the size $m \times n$, where m is the total number of topics in all time windows, and n is the subset of the terms remaining after the data transformation. By using only the top-t terms in each topic we include only the terms that were important in any time window and exclude the terms that never figured in any window topic. This reduces computational costs.

2.3 Quality of modeling

Coherence measures evaluate the interpretability of the automatically generated topics and find the best number of topics. The higher the coherence score, the better the topic model. In the paper we have applied three most widely used coherence measures to determine the optimal number of topics in each time window and the optimal number of dynamic topics, such as UCI [18], $NPMI$ [19], C_v [20].

In [18] the authors propose the UCI coherence measure based on pointwise mutual information (PMI). It is based on the assumption that the co-occurrence of words within documents in the corpus can indicate semantic relatedness.

Th authors of [20] introduced topic coherence based on word co-occurrence counts determined using context windows that contain all words located 5 tokens around the ocurrences of a specific word. Additionally, [20] showed that the UCI coherence performs better if the PMI is replaced by the normalized PMI (NPMI) [19].

3 Data

The proposed algorithm was applied to the data collected from the Russian news resource RosBusinessConsulting from the "Political News" section [1]. The key statistics for this corpus are presented in Table 1. The data was divided into a set of consecutive disjoint sections or time windows - in particular, 12 monthly windows from January 2016 to December 2016. We chose 1 month as the length of the time window to ensure that there is enough data in each time window for topic modeling. Every article from the dataset went through the following preprocessing procedures:

- removal of stopwords
- removal of short words (less than 3 characters)
- lemmatization.

Also TF-IDF term weighting and document length normalization is applied to the Document-Term matrices for each time window.

Table 1: Key statistics of the RBC corpus

Dataset size	
Num. of articles	14725
Average num. of articles per month	1500
Article size	
Average num. of words per article	331

4 Experiments, the first pass

In this section we consider the results of the first pass of the presented approach obtained for each time window. We compare the algorithms by varying the following settings:

- base method (LSI, ICA, PCA, NMF)
- quality measure (UCI, $NPMI$, C_v)
- preprocessing (number of top-terms: 10, 15, 20)

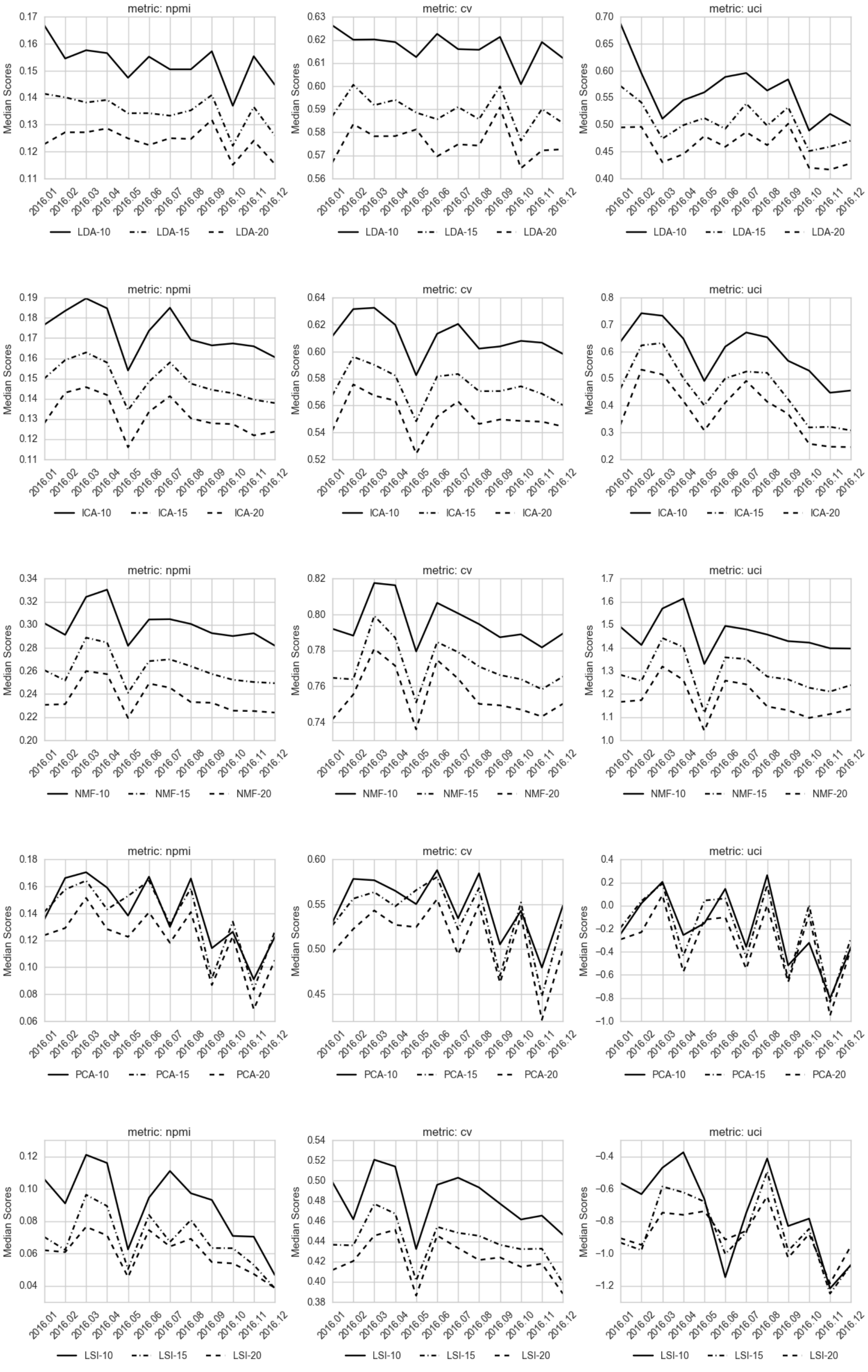

Fig. 1: Changes in the coherence scores depending on modeling method and number of top terms for different coherence measures

4.1 Evaluation of static topic coherence

Figure 1 shows the median coherence scores of topics for all 12 time windows for topic models created by LSI, ICA, PCA, NMF and LDA depending on the number of top terms. We compare median scores to provide a more robust evaluation. Regardless of the topic modeling method, the highest coherence score is achieved with the number of terms $t = 10$.

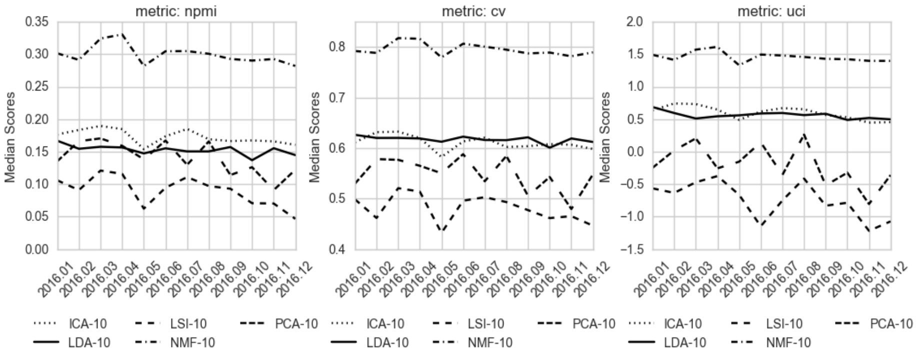

Fig. 2: Coherence scores of static topics

Figure 2 allows comparison of the mentioned topic models. The results show that the NMF achieves higher topic coherence scores in each of the time windows examined for any coherence measure. The results obtained by using ICA are comparable to those found by LDA in terms of coherence.

4.2 Evaluation of static topics descriptors

The results of the first pass of topic modeling are shown in Tables A1-A5 of Appendix A. All terms were translated to English and all topics were manually labeled by experts for better interpretation of results. In these tables, the month and topic number are indicated in the first column. The second column contains the topic label and the top-10 terms that represent the topic.

Tables A3 and A4 show the top-10 terms of topics from the month of January that were obtained with LSI and LDA, respectively. It can be observed, that LSI is less sensitive to more narrow topics and is able to distinguish only broader general topics. For example, it can be seen, that LSI distinguished the following topics: international politics, Iran-Saudi relations, rallies in Chechnya and Kadyrov, events in Syria, the Litvinenko case and nuclear tests in North Korea.

Comparing the mentioned topics with the results obtained by the LDA, it can be noted that the LDA has a wider range of topics, in particular, one should mention the diversity of topics on international politics, for example, topic 6 on Russian-Ukrainian relations, topic 7 on US-Russian relations. Comparing the topics and descriptors obtained with ICA, PCA, NMF and LDA (Tables A1, A2, A3, A5, it is clear that the sets of topics and their descriptive terms overlap and are similar to each other.

5 Experiments, the second pass

In this section we consider the results of the second pass of the presented approach. We compare the algorithms by varying the following settings:

- base method (LSI, ICA, PCA, NMF)
- quality measure (UCI, $NPMI$, C_v)
- preprocessing (number of top-terms: 10, 15, 20)

5.1 Evaluation of dynamic topic coherence

Analyzing the influence of the number of top terms on the interpretability of dynamic topics (Figure 3), it is noticeable that regardless of the model under consideration, the highest coherence score is achieved with the number of terms $t = 10$. In Figure 4 it is shown that the two-pass NMF achieves higher topic coherence scores for any coherence measure. In Table 2 the optimal numbers of dynamic topics sorted by coherence scores are presented. It can be seen that PCA, ICA and NMF find more dynamic topics, this is because the methods are more sensitive to narrower topics.

Table 2: Number of Topics with highest coherence scores for top-10 terms

Coherence measure	**LSI**	**PCA**	**ICA**	**NMF**	**LDA**
UCI	10, 11, 12	48, 47, 29	27, 12, 26	39, 38, 42	11, 13, 10
$NPMI$	10, 12, 11	35, 30, 45	10, 31, 14	48, 49, 26	13, 15, 11
C_v	10, 11, 12	33, 44, 22	10, 31, 42	27, 33, 29	14, 20, 18

5.2 Evaluation of dynamic topics descriptors

Analyzing descriptor sets, you can track dynamic theme changes in time and get a brief overview of events. It should be noted that in one month, within a single dynamic topic, several topics can be found (in this case they are assigned sequential numbers). Examples of obtained dynamic topics are shown in Tables B1-A4 of Appendix B.

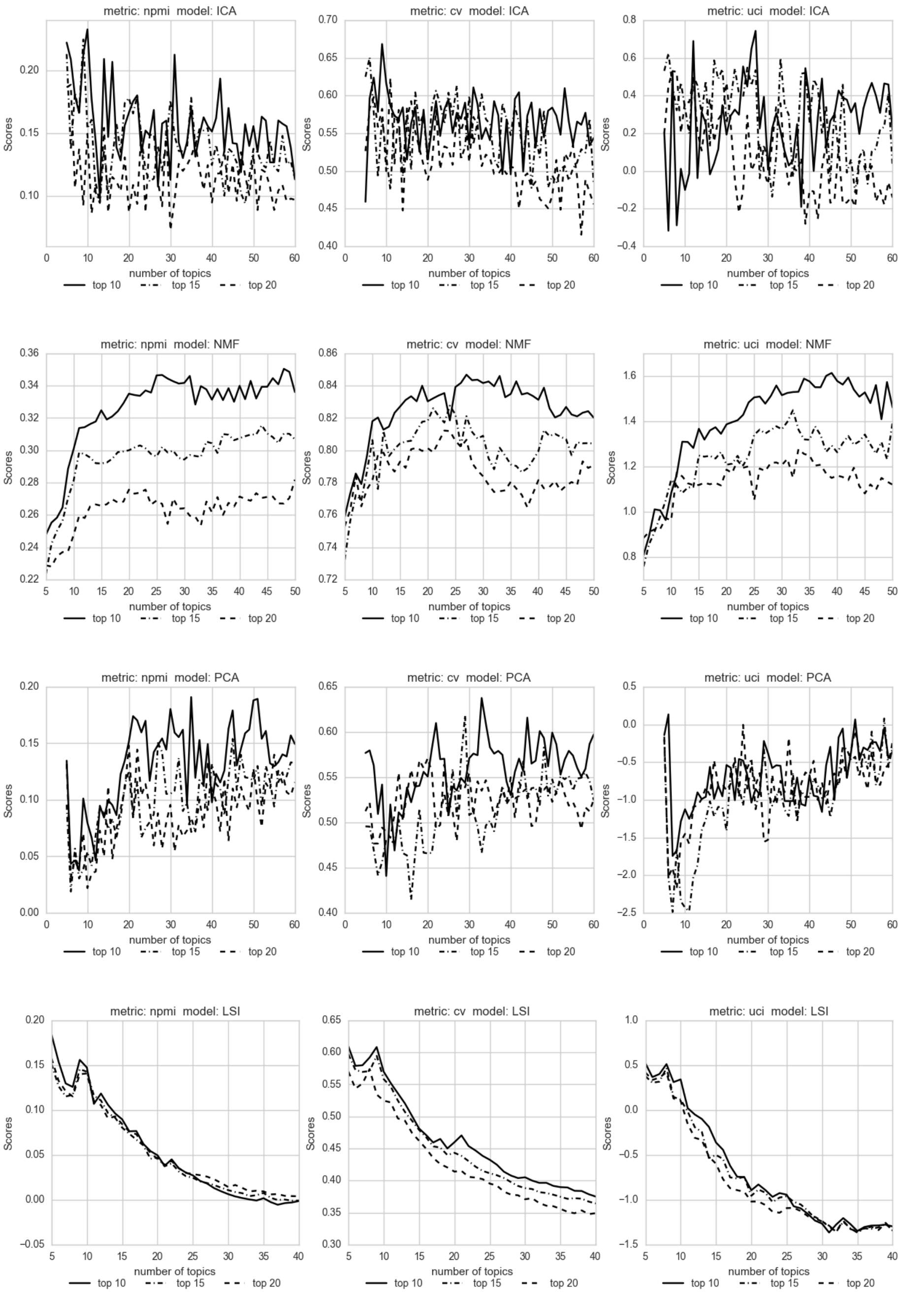

Fig. 3: Coherence scores depending on modeling method and number of top terms for different coherence measures

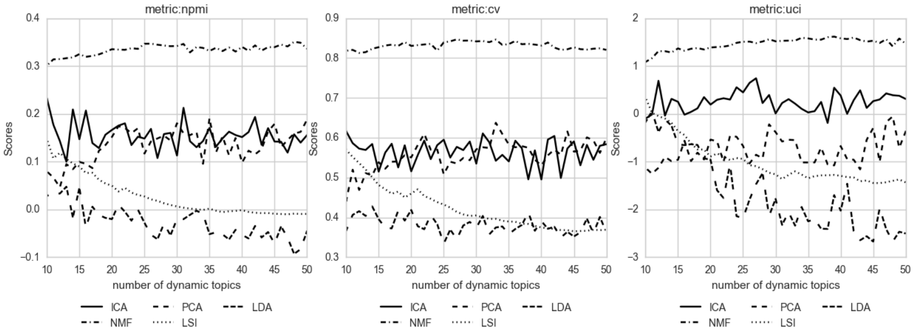

Fig. 4: Coherence scores of dynamic topics

Tables B1, B2, B3 show examples of the evolution of the dynamic topic "War in Syria" obtained using the two-pass model based on NMF, ICA, PCA. All terms were translated to English. It can be seen that the NMF method within one general topic is able to identify narrower subtopics, for example, in Table B3 in February 2016, two topics related to Syria are found, one of which describes news articles related to the ceasefire, and the second - airstrikes on hospitals.

5.3 Comparison to baseline

As a baseline for evaluating the obtained results, we considered the probabilistic algorithm for dynamic topic modeling (DTM), which is based on LDA. We use the C ++ implementation of model proposed by [12] and apply the algorithm with parameters recommended by the authors by default. For comparison, we used the best model obtained in our study - it is the model based on NMF. When comparing NMF and DTM, we fixed the number of dynamic topics. Naturally, the division into time windows is the same for NMF and for DTM. In order to quantitatively compare the results, we evaluated the interpretability of the dynamic topics using the C_v coherence measure for the top-10 terms describing each topic. This measure gives the results closest to human judgments [20]. The two-pass algorithm based on NMF method a higher coherence score of **0.78** versus **0.67** obtained by DTM.

Despite the closeness of the coherence scores, the terms describing window topics created by each model, are very different. Since the dynamic topics generated by DTM are built sequentially and the results obtained for the new time windows depend on the previous ones, the top-terms in each time window are relatively stable. In the approach based on NMF, each model of the time window theme is created independently, based only on the data present in the given window. As a result, the top terms for each topic are much more focused on the trends associated with this topic at the given time.

Tables B3, B4 show examples of the evolution of the dynamic topic "War in Syria" obtained using DTM and the two-pass model based on NMF. It should be noted that in one month, within a single dynamic topic, several topics can be found (in this case they are assigned sequential numbers).

We see that the first 10 terms for topics based on NMF are much more diverse, reflecting the changing nature of news on topics related to Syria. Namely, there is information about the ceasefire, airstrikes on hospitals, the armistice. In the dynamic topics received by DTM, these descriptors did not appear, and the lexical diversity, in general, is much lower.

6 Conclusions

In the paper we compare the performance of four methods of topic modeling (LSI, PCA, ICA and NMF) in the framework of the two-pass algorithm, which was recently developed for dynamic topic modeling. The comparison has been made using different coherence measures and different numbers of topic descriptors. The results showed that that the proposed method based on NMF obtains topics with higher cohesion scores and the median scores for topics with the number of descriptors $t = 10$ is higher regardless of the method.

Comparing the proposed method to DTM, the two-pass algorithm based on NMF outperforms DTM. Although the considered dynamic topic models are relatively similar in terms of their cohesion scores, the two-pass approach based on NMF provides a greater lexical variety/diversity than DTM. Obviously, this improves the interpretability of topics and thereby enhances the quality of the analysis of the evolution of topics in time.

References

1. RosBusinessConsulting. (http://www.rbc.ru/) Accessed: 2017-05-01.
2. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science **41** (1990) 391–407
3. Steyvers, M., Griffiths, T.L.: Probabilistic topic models. Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum. Tang, Z. and MacLennan (2006) 1–6
4. Blei, D.M., Edu, B.B., Ng, A.Y., Edu, A.S., Jordan, M.I., Edu, J.B.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3** (2003) 993–1022
5. Wei, X., Croft, W.B.: Modeling term associations for ad-hoc retrieval performance within language modeling framework. In: Lecture Notes in Computer Science. Volume 4425 LNCS. (2007) 52–63
6. Blei, D.M., Lafferty, J.D.: Correlated Topic Models. Advances in Neural Information Processing Systems 18 (2006) 147–154
7. Daud, A., Li, J., Zhou, L., Muhammad, F.: Knowledge discovery through directed probabilistic topic models: A survey (2010)
8. Vorontsov, K., Potapenko, A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. Communications in Computer and Information Science **436** (2014) 29–46
9. Vorontsov, K., Potapenko, A.: Additive regularization of topic models. Machine Learning **101** (2014) 303–323

10. Wang, Q., Cao, Z., Xu, J., Li, H.: Group matrix factorization for scalable topic modeling. In Proc. 35th SIGIR Conf. on research and Developement in Information Retrieval (2012) 375–384
11. Grant, S., Skillicorn, D., Cordy, J.R.: Topic Detection Using Independent Component Analysis. Proceedings of the Workshop on Link Analysis, Counterterrorism and Security (LACTS'08) (2008) 23–28
12. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning - ICML '06. (2006) 113–120
13. Caron, F., Davy, M., Doucet, A.: Generalized Polya Urn for Time-varying Dirichlet Process Mixtures. 23rd Conference on Uncertainty in Artificial Intelligence UAI'2007 Vancouver Canada (2007) 33–40
14. Wang, C., Blei, D., Heckerman, D.: Continuous Time Dynamic Topic Models. Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08) (2008) 579–586
15. 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. (2006) 424–433
16. Greene, D., Cross, J.P.: Exploring the Political Agenda of the EuropeanParliament Using a Dynamic TopicModeling Approach. Political Analysis **25** (2017) 77–94
17. Skitalinskaya, G.: Analysis of news dynamics using two-pass algorithms of dynamic topic modeling. Mathematical methods and informatics of social processes, Publ. House KIAM-RAS **19** (2017) 13
18. Newman, D., Lau, J., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. … Language Technologies: The … (2010) 100–108
19. Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. Proceedings of German Society for Computational Linguistics (GSCL 2009) (2009) 31–40
20. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers (2013) 13–22

APPENDIX

A Static topics with top-10 descriptors

Table A1: Window topics found in January 2016 using NMF

Time Window	**Descriptors**
2016.1-01:	*Government of Russia*: putin, president, year, russia, party, head, vladimir, choice, government, country
2016.1-02:	*Iran-Saudi relations*: arabia, saudi, nimir, iran, saudian, al, penalty, embassy, shiite, tehran
2016.1-03:	*Chechnya*: kadyrov, chechnya, opposition, people, enemy, ramsan, rally, statement, head, senchenko
2016.1-04:	*Terrorist act in Istanbul*: explosion, terrorist attack, istanbul, happen, terrorist, victim, suicide bomber, perish, terrorist, police

2016.1-05:	*Nemtsov murder case*: court, case, year, ruble, criminal, attorney, investigation, million, bulk, nemtsov
2016.1-06:	*Nuclear weapon in North Korea*: test, dprk, korea, nuclear, bomb, hydrogen, pyongyang, northern, northern, rocket
2016.1-07:	*Litvinenko murder case*: litvinenko, meadow, koktun, judge, fsb, owen, murder, london, case, report
2016.1-08:	*Russian-Ukrainian relations*: ukraine, crimea, ukrainian, kiev, poroshenko, gryzlov, donbass, minsk, russia, negotiations
2016.1-09:	*War in Syria*: syria, negotiations, extremist, military, syrian, us, russia, islamic, operation, isis
2016.1-10:	*Nuclear program of Iran*: iran, usa, sanction, iranian, tehran, american, nuclear, sailor, magate, program
2016.1-11:	*Downed aircraft in Turkey*: turkey, turkish, sukhoi, air, space, airplane, ankara, russia, border, russia

Table A2: Window topics found in January 2016 using PCA

Time Window	**Descriptors**
2016.1-01:	*Nemtsov murder case*: nemtsov, murder, dadaev, gubashev, ruslan, boris, goremeis, anzor, business, bastrykin
2016.1-02:	*Chechnya*: kadyrov, chechnya, opposition, people, enemy, ramzan, statement, meeting, extra-systemic, senchenko
2016.1-03:	*Litvinenko murder case*: litvinenko, meadow, kovtun, kadyrov, judge, fsb, owen, murder, london, case
2016.1-04:	*Iran-Saudi relations*: arabia, saudi, iran, nimir, saudian, penalty, al, shiite, kadyrov, tehran
2016.1-05:	*Nuclear weapon in North Korea*: test, dprk, korea, bomb, nuclear, hydrogen, pyongyang, northern, northern, rocket
2016.1-06:	*Nuclear weapon in North Korea*: commodity, test, dprk, import, korea, embargo, product, ukraine, nuclear, bomb
2016.1-07:	*Russia, corruption*: putin, million, shubin, president, ministry of finance, film, usa, wealth, corruption, thousand
2016.1-08:	*War in Syria*: bbc, action, syria, turkey, shubin, united states, explosion, statement, military, syrian
2016.1-09:	*Russian-Ukrainian relations*: negotiations, syria, ukraine, united states, russia, minsk, sanction, marmots, gryzlov, meeting
2016.1-10:	*War in Syria*: syria, state, strike, islamic, country, turkey, bill, sirian, isis, party
2016.1-11:	*Problems of migrants in Germany*: cologne, police, woman, germany, assault, migrant, harassment, refugee, merkel, sexual
2016.1-12:	*Turkey, downed aircraft*:iran, turkey, turkish, space, sukhoi, air, airplane, sailor, anchor, sanction

Table A3: Window topics found in January 2016 using LDA

Time Window	**Descriptors**
2016.1-01:	*Iran-Saudi relations*: iran, arabia, saudi, nimir, saudian, tehran, al, penalty, embassy, iranian
2016.1-02:	*Elections in Russia*: party, choice, deputy, parliament, rbk, state duma, talk, candidate, elections, question
2016.1-03:	*Russian-Ukrainian relations*: ukraine, ukrainian, russia, year, crimea, kiev, january, poroshenko, donbass, president
2016.1-04:	*Chechnya*: kadyrov, russia, litvinenko, chechnya, name, head, statement, word, opposition, call
2016.1-05:	*Government of the Russian Federation*: putin, president, vladimir, russia, declare, sand, head, press, call, kremlin
2016.1-06:	*Russian-Turkish relations*: russia, country, united states, year, declare, president, turkey, own, sanction, russian
2016.1-07:	*War in Syria*: military, russia, russian, airplane, syria, ministry of defense, strike, sukhoi, air, force
2016.1-08:	*Corruption in Russia*: year, million, ruble, russia, thousand, court, head, law, decision, billion
2016.1-09:	*Problems of migrants in Germany*: germany, refugee, migrant, eu, cologne, woman, country, merkel, border, attack
2016.1-10:	*Terrorist acts in the world*: police, january, action, detain, report, action, report, employee, terrorist act, information
2016.1-11:	*Terrorist act in Istanbul*: explosion, terrorist attack, victim, perish, happen, russian, among, istanbul, reuters, embassy
2016.1-12:	*War in Syria*: syria, negotiations, test, islamic, dprk, military, united states, un, iraq, january
2016.1-13:	*Nemtsov murder case*: case, court, criminal, investigation, lawyer, murder, crime, nemtsov, investigation, year
2016.1-14:	*USA, elections*: president, usa, trump, post, state, candidate, party, donald, billionaire, presidential

Table A4: Window topics found in January 2016 using LSI

Time Window	**Descriptors**
2016.1-01:	*International relations*: russia, year, president, united states, ukraine, country, putin, state, iran, syria
2016.1-02:	*Iran-Saudi relations*: iran, arabia, saudi, nimir, saudian, al, tehran, execution, embassy, shiite
2016.1-03:	*Chechnya, Islamic world*: kadyrov, chechnya, iran, opposition, people, enemy, arabia, ramzan, rally, saudi
2016.1-04:	*Chechnya, war in Syria*: kadyrov, syria, turkey, terrorist, explosion, islamic, terrorist attack, chechnya, isis, military

2016.1-05:	*Litvinenko murder case*: litvinenko, case, court, murder, meadow, fsb, judge, koltun, investigation, criminal
2016.1-06:	*Nuclear weapon in North Korea, Litvinenko murder case*: test, dprk, korea, nuclear, bomb, litvinenko, hydrogen, missile, pyongyang, united states

Table A5: Window topics found in January 2016 using ICA

Time Window	**Descriptors**
2016.1-01:	*War in Syria*: nimir, al, arabia, saudi, penalty, crime, party, putin, shiite, protest
2016.1-02:	*War in Syria*: negotiations, syria, geneva, en, delegation, rebel, opposition, syrian, carrie, meeting
2016.1-03:	*Nuclear weapon test in North Korea*: test, dprk, korea, nuclear, bomb, hydrogen, pyongyang, northern, northern, rocket
2016.1-04:	*Litvinenko murder case*: litvinenko, meadow, koltun, judge, killing, fsb, owen, london, case, putin
2016.1-05:	*Undefined*:terrorist, town, bbc, al, assault, united states, dagestan, police, hostage, woman
2016.1-06:	*Undefined*: putin, year, ruble, syria, court, russia, president, business, million, russian
2016.1-07:	*Russian-Ukrainian relations*: ukraine, crimea, ukrainian, poroshenko, kiev, gryzlov, donbas, goods, minsk, contact
2016.1-08:	*War in Syria*: syria, action, military, isis, islamic, operation, al, strike, iraq, zorah
2016.1-09:	*Iran-Saudi relations*: arabia, saudi, nimir, saudian, iran, al, penalty, shiite, embassy, raqqa
2016.1-10:	*Nemtsov murder case*: court, case, nemtsov, criminal, attorney, investigation, bulk, murder, crime, detain
2016.1-11:	*Undefined*: year, ruble, government, refugee, minister, prime minister, billion, choice, candidate, court

B Dynamic topics with top-10 descriptors

Table B1: ”War in Syria” Topic Evolution - ICA

	Feb 2016	**May 2016**	**Jun 2016**	**Aug 2016**
1	aleppo	syria	syria	syria
2	army	aleppo	aleppo	operation
3	town	mode	terrorist	aleppo
4	offensive	ceasefire	strike	turkey
5	strike	fire	an	turkish
6	asad	nusra	nusra	syrian
7	hospital	an	syrian	town
8	syria	silence	ministry of defense	strike
9	aviation	province	gunning	jarabulus
10	province	organization	fire	kurdish

Table B2: "War in Syria" Topic Evolution - PCA

	Jan 2016	**Feb 2016**	**Mar 2016**	**Apr 2016**
1	syria	admiral	aleppo	aleppo
2	aleppo	strike	terrorist	syria
3	mode	kuznetsov	syria	terrorist
4	fire	aleppo	town	town
5	ceasefire	syria	ministry of defense	syrian
6	military	cruiser	army	strike
7	province	plane	plane	humanitarian
8	an	ministry of defense	eastern	army
9	terrorist	military	palmira	un
10	plane	ukraine	military	military

Table B3: "War in Syria" Topic Evolution - NMF

	Jan 2016	**Feb(1) 2016**	**Feb(2) 2016**	**Mar 2016**
1	syria	syria	aleppo	syria
2	aleppo	ceasefire	syria	palmira
3	military	armistice	army	russia
4	isis	fire	town	military
5	islamic	usa	strike	russian
6	iraq	russia	offensive	military
7	operation	un	assad	aircraft
8	town	agreement	hospital	syrian
9	state	mode	aviation	assad
10	beat	assad	russia	operation

Table B4: "War in Syria" Topic Evolution - DTM

	Jan 2016	**Feb 2016**	**Mar 2016**	**Apr 2016**
1	operation	operation	town	town
2	town	town	operation	terrorist
3	mosul	mosul	terrorist	operation
4	iraq	extremist	isis	syria
5	isis	isis	syria	army
6	iraq	iraq	islamic	isis
7	islamic	islamic	mosul	islamic
8	state	syria	army	state
9	syria	state	state	palmira
10	army	army	iraq	iraq