A topic modeling analysis of V. Putin’s addresses transcripts

Vasyukova Viktorija

Introduction

Russian President's speech is a thing that is commonly imitated in a number of russian jokes. Topics like taxes and GDP growth are usually considered to be an essential part of Putin's image. In this work we will track how Vladimir Putin's speech has been changing over time since we are able to obtain hundreds of texts. To achieve that we will use topic modelling techniques. Determining main themes of presidential speech will help us get a better understanding of what his speech, in essence, was and is.

Related work

Villadsen (2014) describes a topic modelling text analysis on a corpus of 622 key U.S. presidential speeches using the gensim toolkit, which returned 25 topics out to 20 words for each topic. Tracing the appearance of the topics over time, visualising the topic distribution and finding a trend are the main outcomes of the work. One of the conclusions the author comes to is that Latent Dirichlet Allocation (LDA) is indeed an effective technique for identifying underlining themes of various texts, including political speeches, enabling us to get more insights. Moreover, he leaves us advice for successful topic modelling research, notably he marks the importance of corpus tokenization and stop words removal; running several iterations of the model; looking at more words to correctly interpret the theme and etc.

Analogous usage of content analysis was already made for «oil and gas» topic in Sergeeva (2013) conducted in a form of a qualitative study which might be used for explaining our results later.

Another paper related to our field of interest by Tong, Zhou & Zhang, Haiyi (2016) includes two experiments based on LDA usage — Wikipedia articles and Twitter Data analysis. After the model is built for Wikipedia, the Jensen-Shannon divergence method is used for calculating the distance between articles.

Finally, an interesting post called «Interpreting Putin: a machine learning approach» focuses on word embeddings to better understand the topics used in Putin's speeches. It provides a list of the most common terms and highest in tf-idf value words within Putin speech corpus by year. For instance, in 2014 they are 'Ukraine' and 'Sevastopol' as a result of annexation of Crimea. The author uses the same data we are intended to use but in English rather than in Russian. One of the findings of this work is that FastText which splits individual words into n-gram chunks performs better than more popular GloVe and word2vec.

Method

In order to topic model some set of documents we first must choose a suitable tool. There are a few methods to choose from. Latent Dirichlet allocation (LDA) and Non-Negative Matrix Factorization (NMF) are the two most popular ones. The first one is a probabilistic approach based on the distributional hypothesis, the second one is a matrix factorisation technique used for multivariate data decomposition. LDA is more common among the articles we have researched for our task. Thus, we will focus on this one.

Keeping in mind the advice from Villadsen (2014) we first have to obtain large amount of data which in our case should be downloaded from the Official Internet Resources of the President of Russia website (http://kremlin.ru/events/president/transcripts/speeches). We focus on addresses that are speeches addressed to public. There are 740 of them starting from 2012 when Putin came back as a leader after Dmitry Medvedev. We collect them using urllib python packages.

Then the documents must be preprocessed: transformation to lower case, punctuation, stop words and blanks removal. Additionally, we apply lemmatization. We extend the list of stop words for our dataset to avoid processing words that are always present in presidential speech (like 'ladies', 'gentlemen', 'hello'). We also drop insignificant parts of speech - numbers, pronouns, adverbs and etc. The only remaining problem is named entities.

When we have a dictionary and document-term matrix at our disposal we may continue with the algorithm for determining topics — *LDA* (using gensim).

Results

To check whether topics are consistent, and we've chosen the right quantity, we can apply topic coherence measures (the degree of semantic similarity between high scoring words in the topics — topic is considered to be coherent if most of the top-N words in this topic are related). We will use c_v measure provided by CoherenceModel in gensim to calculate coherence score for multiple LDA models with the same hyperparameters but different number of topics. C_v measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information and the cosine similarity. It is believed to be a measure that correlates the most with human topic ranking data. More on c_v calculation and history in Syed, S., & Spruit, M. (2017). By looking at the graph of number of topics and coherence score we can determine how many topics is enough — with the end of a rapid growth of the score. Here we assume that the optimal number of topics is around 6.

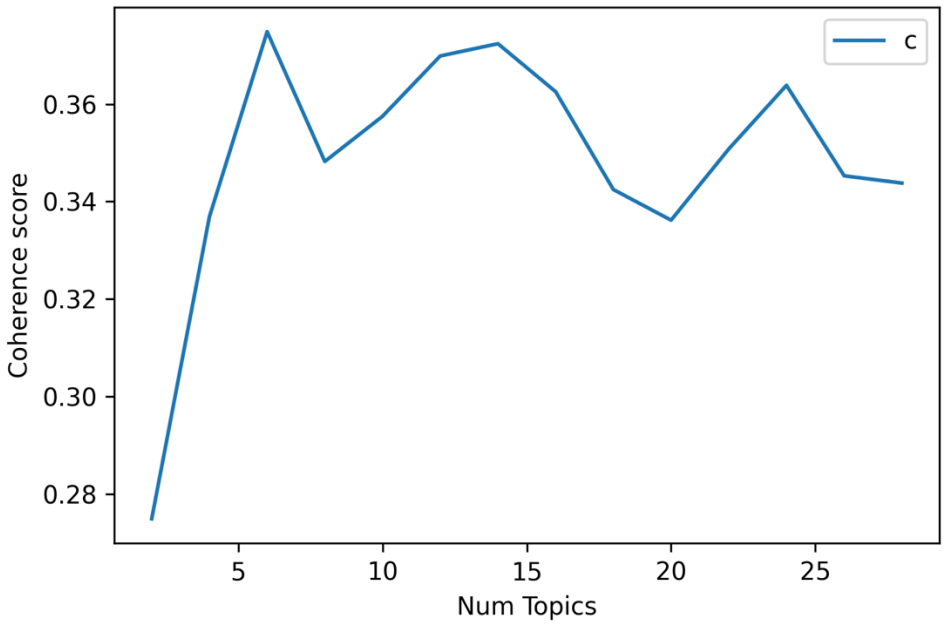

Figure 1. Coherence scores for different number of topics

After dozens of attempts we finally obtain interpretable topic model that consists of 6 topics. Previous models showed themes like 'sports' and 'victory day', and with tf-df plugged in we got topics that are focused just on named entities.

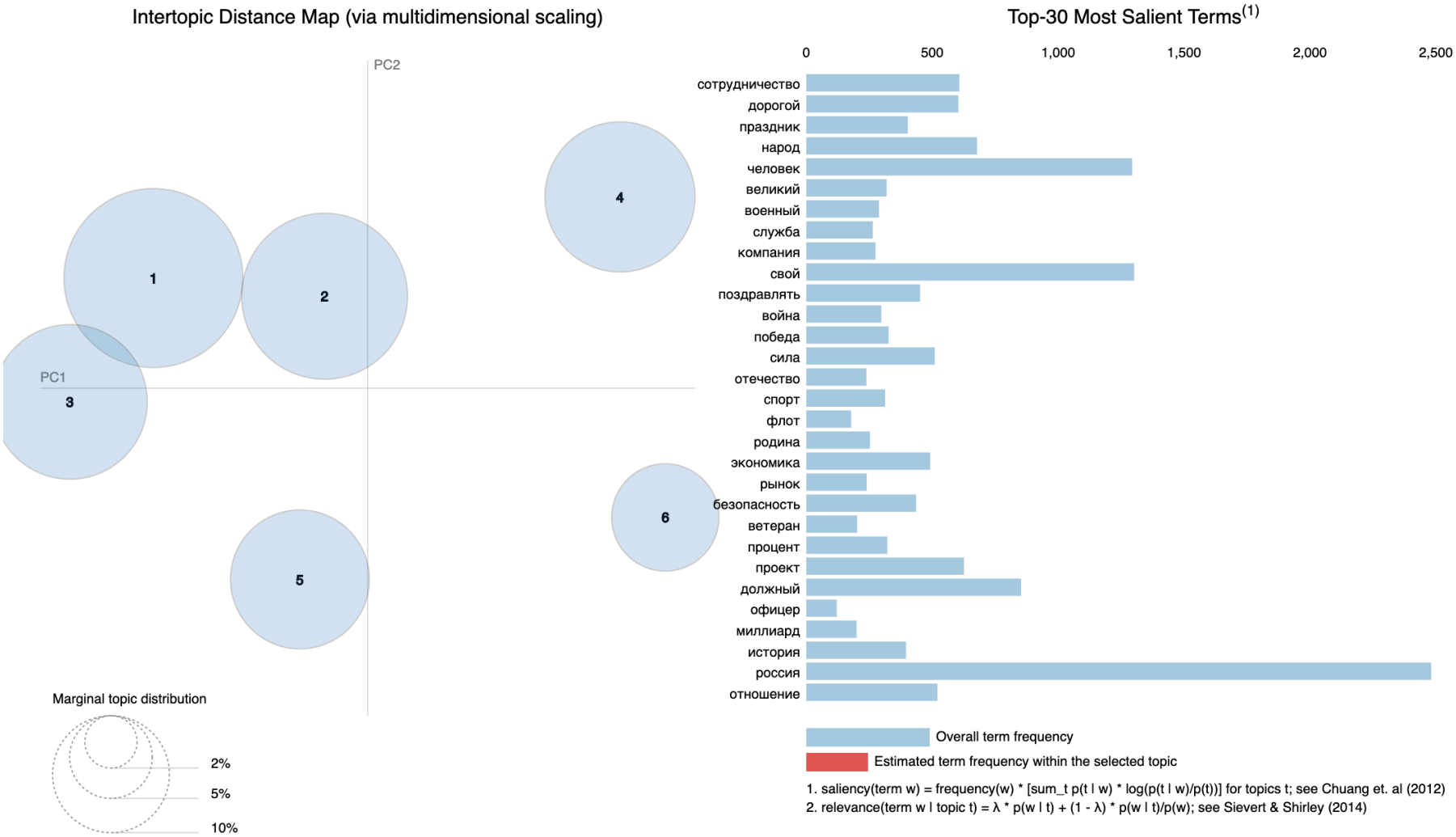

Figure 2. Intertopic Distance Map and Top-30 Most Salient Terms (pyLDAvis visualization)

These are 6 topics created when using initial corpus. As described in Sievert, C., & Shirley, K. (2014), left part of the pyLDAvis topic representation is based on multidimensional scaling, namely Principal Component Analysis. Scaling helps to plot the topics in the two-dimentsional space by projecting the distances between topics. The default for computing inter-topic distances is Jensen-Shannon divergence mentioned earlier.

As we see on inter topic distance map, there is almost no overlap. The sixth topic is the smallest in size. Suggested topics, based on the top-30 most relevant terms (belonging primarily to the corresponding topic with relevance metric lambda = 0.2) are shown in the following table. Although, they all contain frequent words like Russia ('россия' is the most frequent term in our corpus used over 2,500 times), they do not focus on names and titles as tf-idf models did.

Table 1. Topic names with 10 words for each (for a translated version see Appendix, Table 2)

Political partnership, foreign affairs	**Church, Ukraine, patriotism**	**Economy**	**Sports**	**National security**	**Victory Day (the Great Patriotic War), Military**
1	**2**	**3**	**4**	**5**	**6**
сотрудничество	церковь	компания	спорт	сотрудник	офицер
взаимодействие	партия	рынок	олимпийский	служба	флот
президент	православный	миллиард	хоккей	орган	моряк
отношение	украина	инвестиция	победа	спецслужба	солдат
республика	крым	бизнес	тренер	оперативный	доблесть
переговоры	власть	отрасль	награда	правоохранительный	воинский
брикс	пенсия	экономика	праздник	коррупция	военный
международный	выбор	доллар	команда	защита	война
саммит	должный	рост	игра	нелегальный	великий
дипломатический	конституция	нефть	выдающийся	наказание	герой

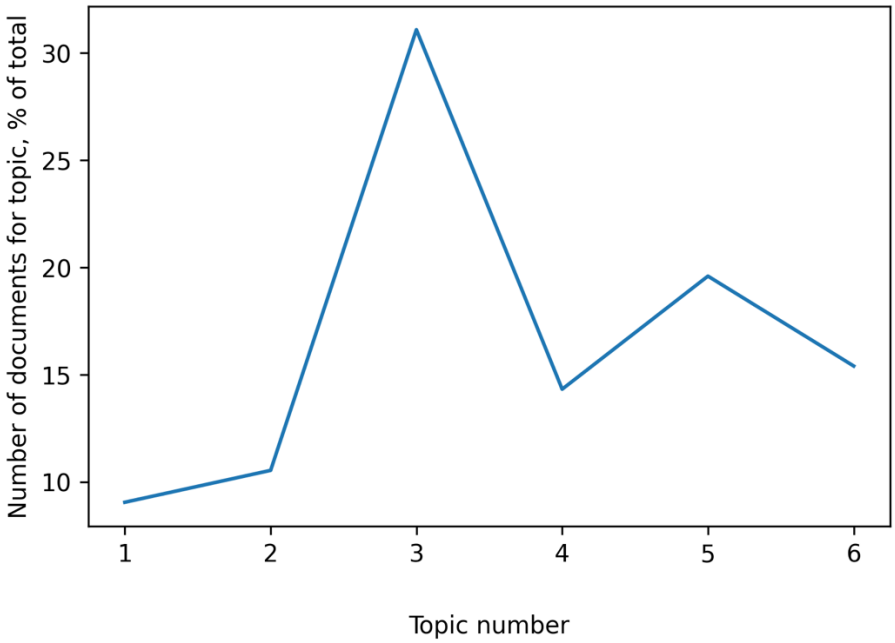

Figure 3. Percentage of documents per each topic

This figure shows that the most common topic in the corpus is the third one, which is Economy. Next in size is the 5th one — National security. The smallest number of documents were assigned topic 1, which is Foreign affairs. Surprisingly, the number of patriotic or religion related speeches is very low, too.

On the figure below we can see that the third topic is the most common each year with a peak at 2014. It is probably because of the Russian financial crisis that took place in 2014–2015. So, the focus over time indeed has remained relatively constant, as mentioned in Chestnutt (2018). In 2012 and 2020 the topic distribution is a lot different due to small number of documents available. Oddly, topic 1 is leading in 2020, maybe due to worldwide coronavirus situation.

Since Olympics is the main sporting event, it was expected that number of documents designated to sport will increase in even years and decrease in odd ones. It is true for all the years except 2018 when Russia was banned from participation in the 2018 Winter Olympics in Pyeongchang, South Korea after the doping scandal.

Topic distribution over month doesn't give us any more insights. Interestingly, national holidays don't affect how topics are situated on a graph. For example, it would be logical if the number of texts about war rose in May (Victory Day is on the 9th of May), but it doesn't happen.

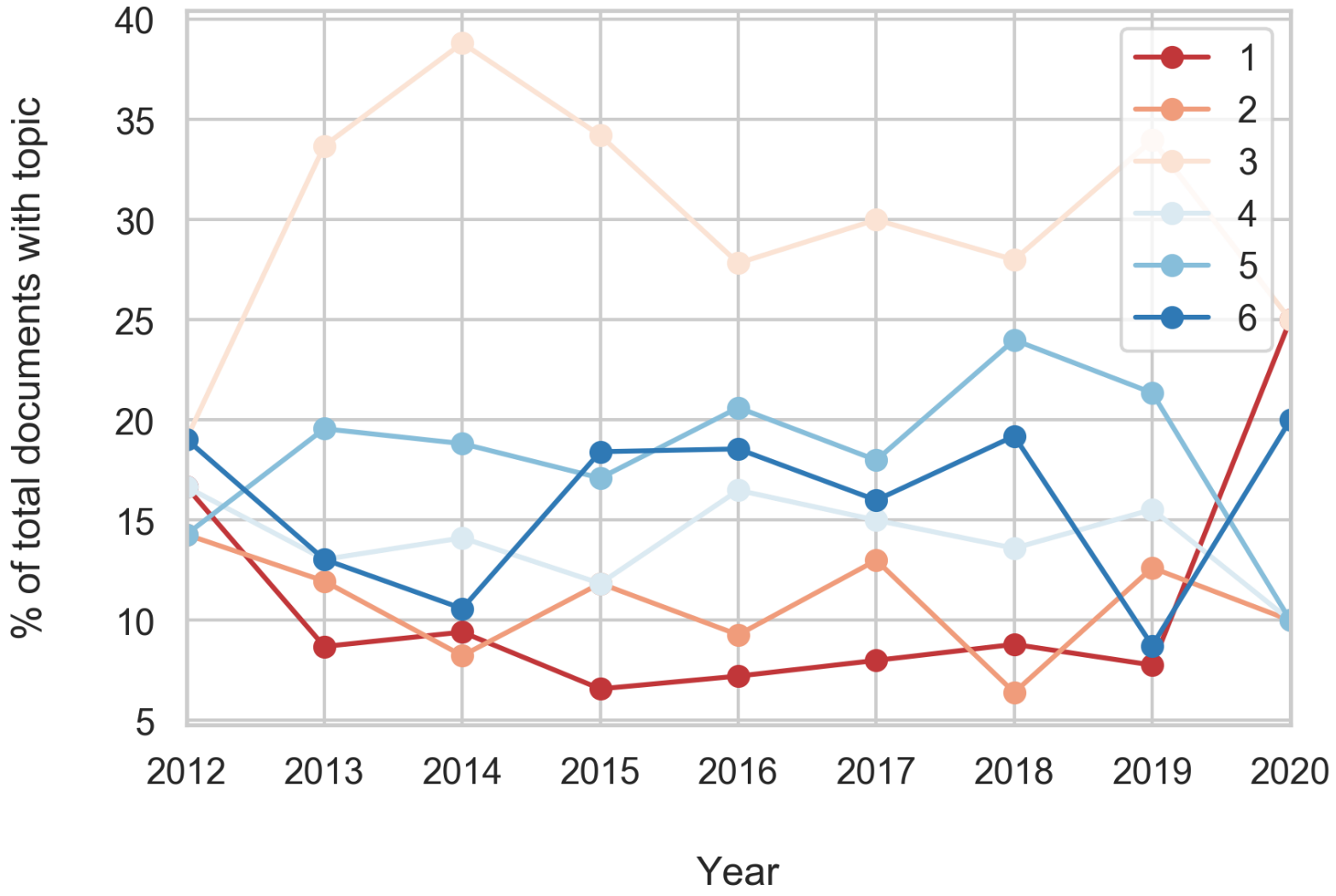

Figure 4. Topic distribution over years

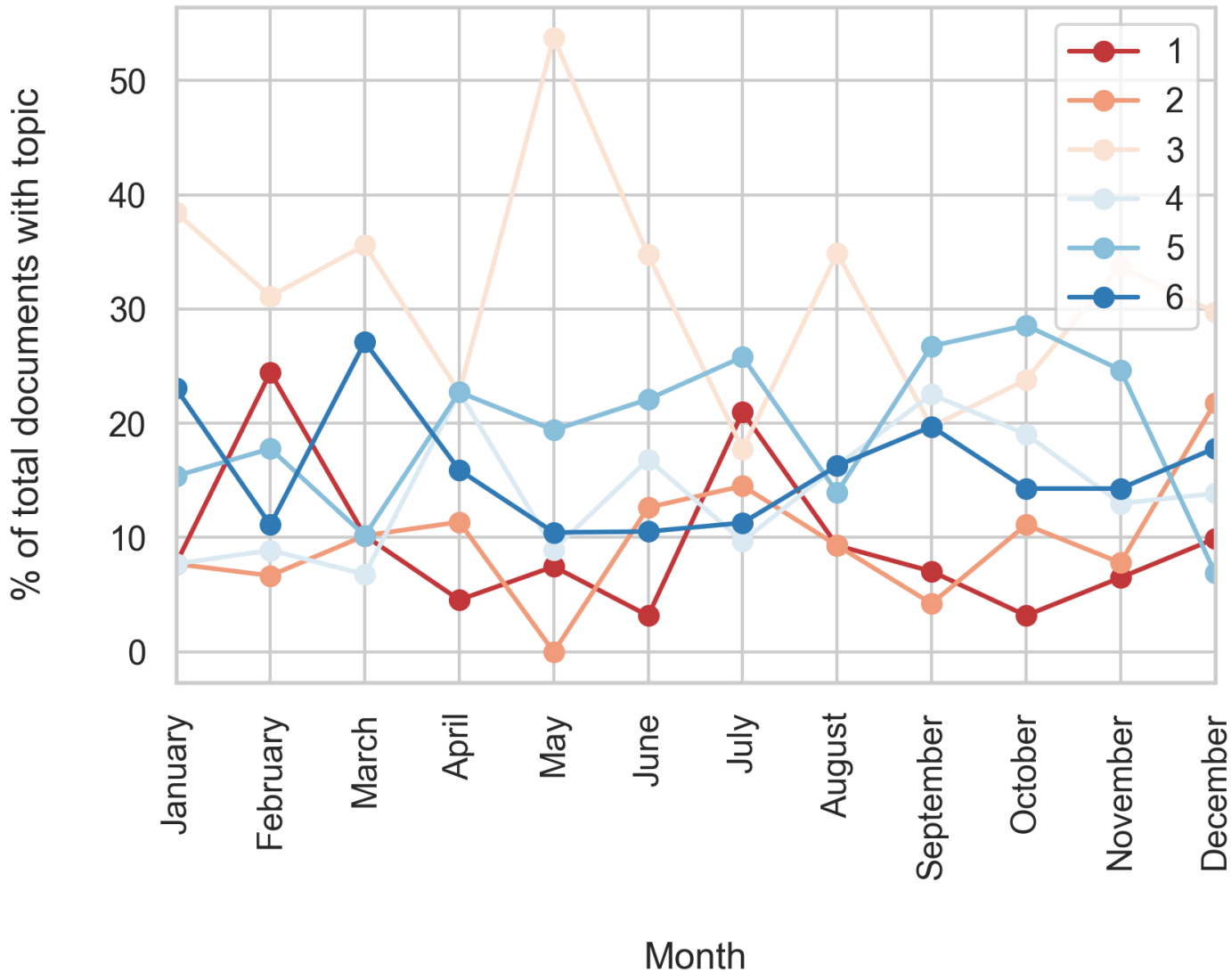

Figure 5. Topic distribution over months

Conclusion

In this paper we've conducted a research on texts documenting Vladimir Putin's speech. Using Latent Dirichlet Allocation method, 6 topics were extracted. The number of topics was chosen by coherence calculation. Overall, the results showed that it is possible to determine some topics in presidential speech that will make sense (based on top-30 terms), and their distribution over time (only when considering year) proves that the method worked.

As a possible improvement of the research there could have been applied the second mentioned method — NMF. Also, the documents could have been assigned more than 1 topic each. This way it would be possible to see how often do the themes coincide in one speech.

References

1. Chestnutt, Robert F. (2018) Interpreting Putin: a machine learning approach. URL: http://caspianet.eu/2018/11/28/interpreting-putin-a-machine-learning-approach/#fn1
2. Sergeeva Z. H. (2013) Neft' i gaz v rossijskom politicheskom diskurse (po rezul'tatam kontent-analiza poslanij Prezidenta RF Federal'nomu Sobraniju RF za 2002-2012 gg.) *Vestnik Kazanskogo tehnologicheskogo universiteta*. 4.
3. Tong, Zhou & Zhang, Haiyi. (2016). A Text Mining Research Based on LDA Topic Modelling. *Computer Science & Information Technology.* 6. (pp. 201-210)
4. Villadsen, Ole R. (2014). Analyzing Presidential Speeches with Topic Modeling
5. Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In 2017 IEEE International conference on data science and advanced analytics (DSAA) (pp. 165-174)
6. Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70)

Appendix

Table 2. Topic names with 10 words for each (translated)

Political partnership, foreign affairs	Church, Ukraine, Patriotism	Economics	Sports	National security	Victory Day (the Great Patriotic War), Military
1	2	3	4	5	6
cooperation	church	company	sports	officer	army officer
collaboration	party	market	olympic	service	navy
president	orthodox	billion	hokkey	agency	marine
relationship	ukraine	investment	victory	intelligence agency	soldier
republic	crimea	business	coach	оперативный	valour
negotiation	power	industry	prize	law enforcement	wartime
brics	pension	economics	celebration	corruption	military
international	choice	dollar	team	protection	war
summit	due	growth	game	illegal	great
diplomatic	constitution	oil	outstanding	punishment	hero