PROBABILISTIC GRAPHICAL MODELS, SS 2022 PORTFOLIO EXAM REPORT

Saida Yusupova
Department of Computer Science
Technical University of Kaiserslautern
Kaiserslautern, 67663, Germany
yusupova@rhrk.uni-kl.de

ABSTRACT

Topic modeling analyzes a collection of text documents and determines which topics each document in the collection belongs to by learning interpretable list of words. Latent Dirichlet Allocation (LDA) is commonly used parametric model, which requires pre-defined number of topics to be generated. However, we cannot always know how many topics the collection of documents have or to expect. In this case, we consider non-parametric models. To this end, we train parametric and non-parametric models on two versions of the same dataset (original & modified version) and analyze the topics discrepancy. Our results show that the parametric model outperforms the non-parametric. Moreover, the coherence score of the original dataset is higher that the score of the modified dataset. Finally, we propose a method to improve the quality of the poor model.

1 INTRODUCTION

Topic modeling is a set of unsupervised techniques used to analyze text in a collection of documents and identify meaningful group of words (topics) (Jordan Boyd-Graber & Mimno (2017)). The topic model construction algorithm receives a collection of text documents as input. At the output, for each document, a numerical vector is produced, made up of estimates of the degree of belonging of this document to each of the topics. The dimension of this vector, equal to the number of topics, can either be set at the input or determined automatically by the model.

Given a collection of text documents D. Each document d from collection D is a sequence of words $w_d = (w_1, \ldots, w_{n_d})$ from dictionary W, where n_d is the length of document d. It is assumed that each document may relate to one or more topics. Topics differ from each other by varying the frequency of the use of words. The purpose of building a topic model can be both directly identifying a variety of latent topics, and solving various additional tasks such as ranking documents according to the degree of relevance to a given topic, determining how the topics have changed over time and etc.

2 APPROACH

The whole work consist of three steps. First, we preprocess the dataset (\$2.1). Second, we deploy parametric (LDA) or non-parametrric (HDP) training (\$2.2). Lastly, we evaluate the coherence measures and analyze the topics difference for two versions of the dataset().

2.1 DATASETS AND PREPROCESSING

We use '20NewsGroups' dataset in two versions: original (19997 documents) and slightly modified versions (18397 documents). The exact modifications are unknown. One of the tasks is to roughly identify the modifications in the second dataset.

I use nltk's RegexpTokenizer to tokenize and to identify entities in the textual file weuse gensim package. See Appendix A.1 forfull preprocessing pipeline.

2.2 MODELS

I evaluate one parametric and one non-parametric model:

LDA: To train Latent Dirichlet Allocation (LDA) model, I use gensim package with the following parameters: the number of topic, I assume, is 20, number of interations is 1000. The hyperparameters α, which is the topic density parameter, I initially assume as 'auto' and η, which is the word density parameter is 'auto' as well.

HDP: To train non-parametric model, I use Hierarchical Dirichlet Process (HDP) model provided by gensim package. Here, we dont need specify number of topics, the model will learn it automatically. However, for better performance, I specify some priors, such as initial number of topics K=10 and maximum possible topics as T=50. In case these priors are not assumed, the model will generate 150 topics. The topic Dirichlet η, I assume, is 0.01, while the α and γ parameters are set to 1.

2.3 EVALUATION

In case of LDA model, as evaluation criterias, I use topic's log-perplexity score and C_V coherence score. During the evaluation, I found out that calculation of perplexity score for HDP model is not available, as a result the topics of this model are evaluated only by their coherence measures.

3 EXPERIMENTS

3.1 FIRST EXPERIMENT

I train LDA and calculate metrics for original and modified datasets with number of topics 20. Similarly, I train HDP for both versions of dataset and calculate coherence score for 20 topics.The Table 1 and Figure 1 show the results:

	LDA_original	LDA_modified	HDP_original	HDP_modified
Perplexity	-10.42	-10.76	-	-
Coherence	0.51	0.48	0.46	0.44

Table 1: Metrics for the four topic models, where LDA_original and LDA_modified shows relults of LDA tarining for original and modified datasets. Results for HDP respectfully.

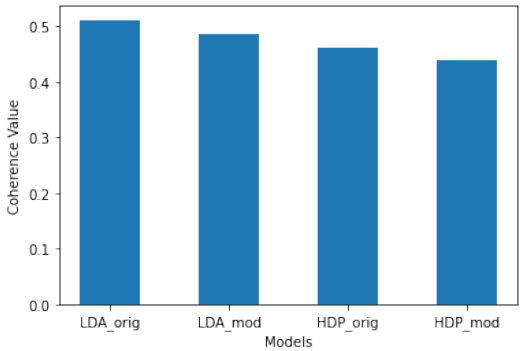

Figure 1: Comparison of coherence scores

As we can see from the results, the LDA model trained on original dataset outperformed the modified version. Same scenario in training HDP model. Moreover, during implementation[1], I noted some differences in two datasets. First, the modified version of dataset is smaller than the original. The

[1] https://gitlab.rhrk.uni-kl.de/yusupova/pgm_portfolio/-/tree/main/

number of documents in original dataset is 19997, while in modified version is 18397. Moreover, in modified dataset does not contain topics about autos, sport, games, while the original dataset does.

3.2 SECOND EXPERIMENT

In the second experiment, I analyze three dataset length, namely I sample 60%, 80% and 100% of the dataset. I train LDA and HDP models, and analyze metrics for each dataset. As shown in Figure 2, the perplexity score decreases while the number of topics increases. Same situation can be observed for all dataset sizes, the drastic decrease for number of topics greater that 12. We can conclude, that the dataset size do not affect the log-perplexity score.

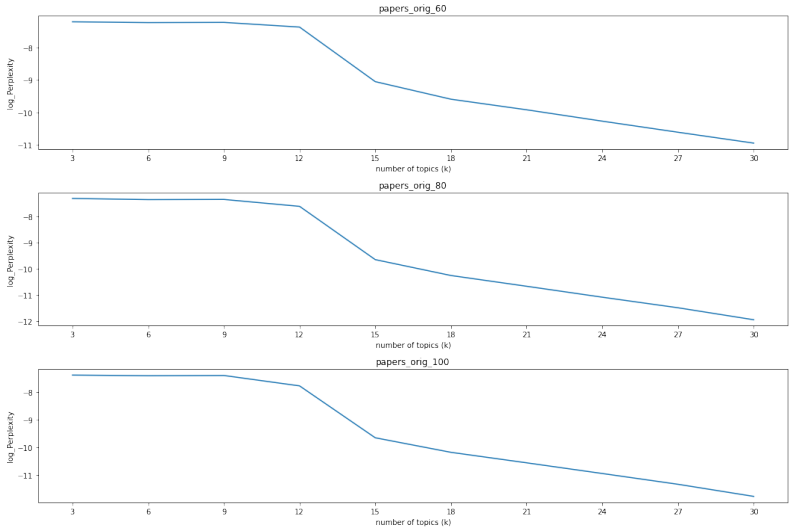

Figure 2: Comparison of log-perplexity scores (LDA) for original dataset with different length.

In case coherence score, see Figure 3, we observe fluctuation in scores. For smaller dataset, the coherence decreases, while for other two datasets we can see that the higher coherence score for 15 topics.

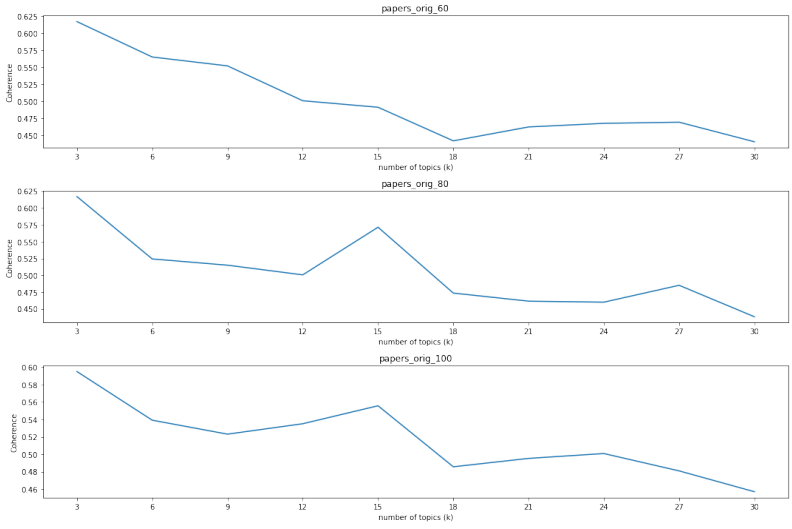

Figure 3: Comparison of C_V coherence scores (LDA) for original dataset with different length.

See Appendix 2 for comparison of metrics for other models.

To conclude, for LDA model the log-perplexity score decreases with the increase the number of topics for both datasets and for all dataset sizes. The coherence score for smaller dataset in original and modified datasets significantly decreases with the increase of topic number, while for 80% and 100% datasets of original and modified datasets, the coherence score higher for topic numbers of 15 and 18, respectively.

For HDP model, we observe the decrease in C_V coherence scores after 6 topics for both datasets with different length.

3.3 THIRD EXPERIMENT

In the third experiment, I sample sentence-wise, namely 500 000, 1 500 000 and 2 500 000 sentences. I train LDA and HDP models, and analyze metrics for each dataset. As shown in Figure 4, the perplexity score decreases while the number of topics increases. Same situation can be observed for all dataset sizes, the drastic decrease for number of topics greater that 12. We can conclude, that the dataset size also do not affect the log-perplexity score.

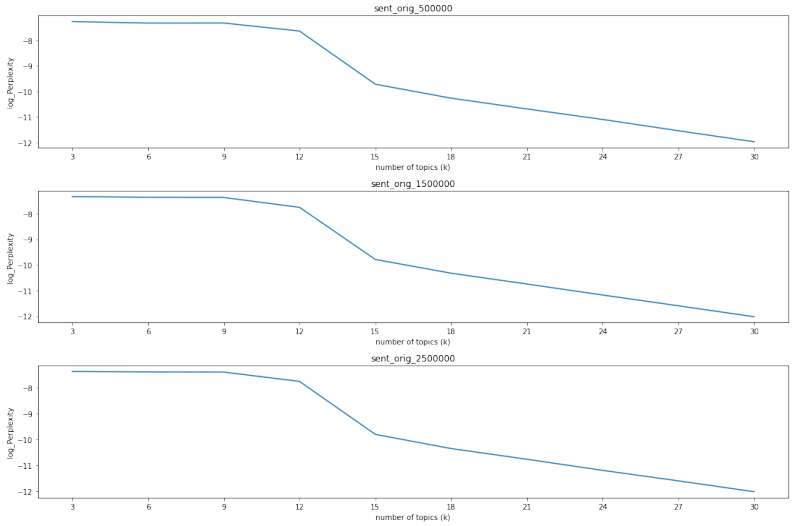

Figure 4: Comparison of log-perplexity scores (LDA) for original sentence-wise dataset.

In case coherence score, see Figure 5, we observe decrease in scores and the optimal topic number is not clear.

See implementstion for comparison of metrics for other models.

To conclude, for LDA model the log-perplexity score decreases with the increase the number of topics for both sentence-wise datasets, the coherence score fluctuates between 15 and 21 topics. For HDP model, we also observe the decrease in C_V coherence scores after 6 topics for both sentence-wise datasets.

4 CONCLUSION

In conclusion, as you can see in implementation the proper choice of parameters can increase the coherence value of the topic, this the the method I propose. I select LDA training of modified dataset and I train this model for different values of α and η: $\alpha \in 0.01, 0.05, 0.1, 0.25, 1.0, 5.0$ and $\eta \in 0.01, 0.05, 0.1$. According to results, the optimal coherence score is 0.6190 for $\alpha = 1.0$ and $\eta = 0.01$

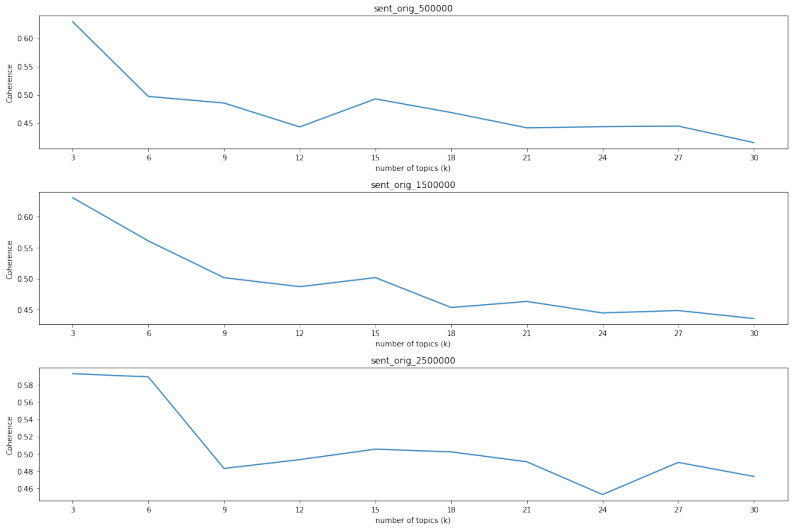

Figure 5: Comparison of C_V coherence scores (LDA) for original sentence-wise dataset.

REFERENCES

Matthew J Denny and Arthur Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. in political analysis.

Yuening Hu Jordan Boyd-Graber and David Mimno. Applications of topic models. 2017.

A APPENDIX

A.1 DETAILS ON PREPROCESSING STEPS

- I process further text files with more than 25 tokens
- I remove emails from the text files
- I tokenize using RegexpTokenizer with the r'[a-zA-Z]{3,}' condition, namely I do not process numbers and strings that contain numeric character. Additionally, strings with length greater that two is processed further.
- I lowercase the tokens.
- I remove stopwords.
- I identify entities and keep only those, whose NPMI score is greater than 0.8.
- I lemmatize with pos_tag in [NOUN='n', VERB='v'].
- To create vocabulary, I filter out terms that appear in more that 80% of the documents and in less that 50 documents (Denny & Spirling)

A.2 COMPARISON OF METRICS FOR MODELS WITH DIFFERENC DATASET SIZES

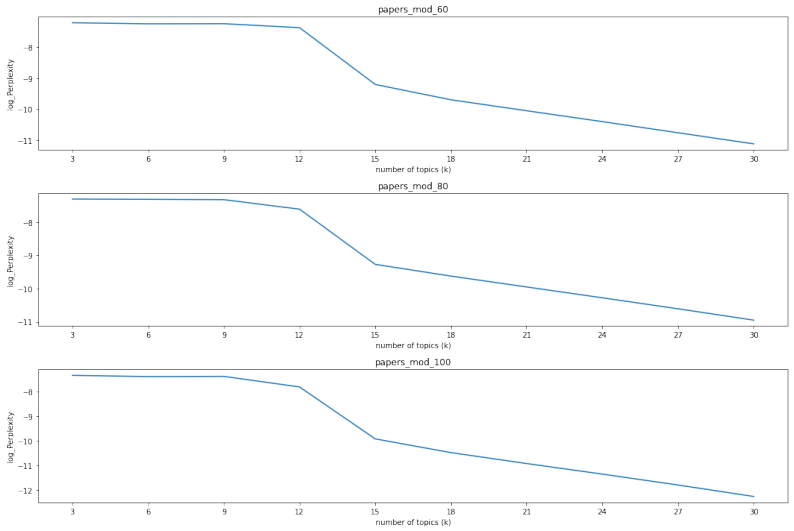

Figure 6: Comparison of log-perplexity scores (LDA) for modified dataset with different length. The perplexity score decreases while the number of topics increases. Same situation can be observed for all dataset sizes, the drastic decrease for number of topics greater that 12.

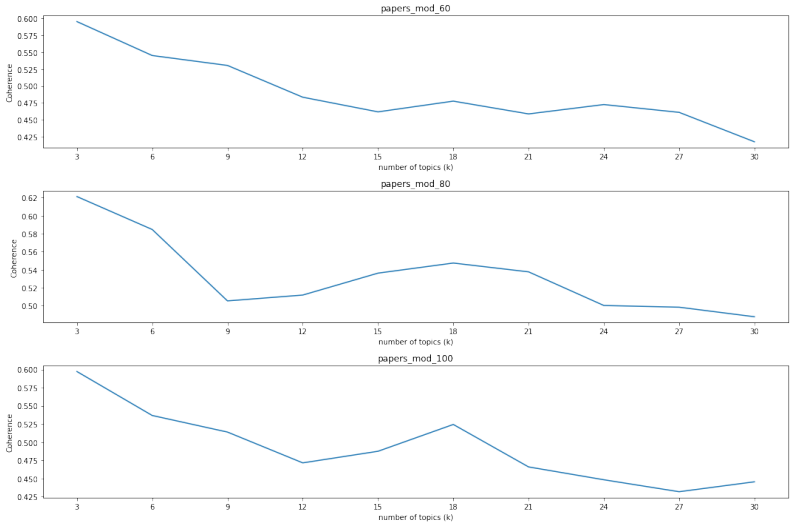

Figure 7: Comparison of C_V coherence scores for modified dataset with different length. We observe fluctuation in scores. For smaller dataset, the coherence decreases, while for other two datasets we can see that the higher coherence score for roughly 18 topics.

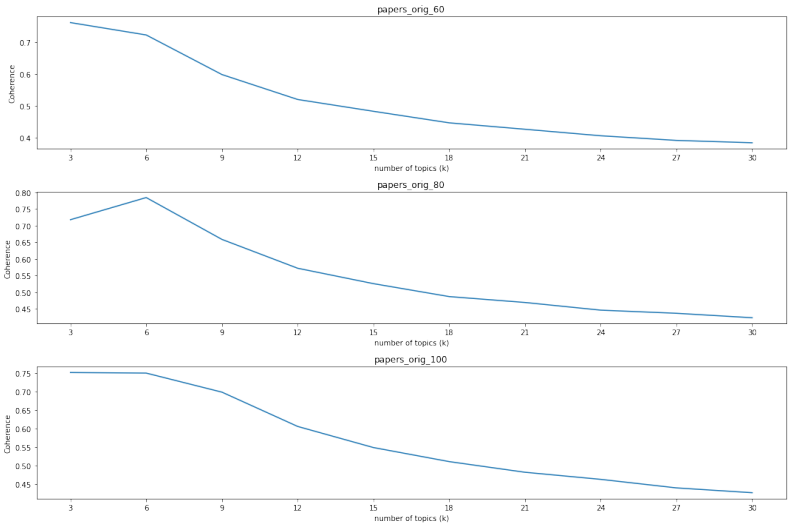

Figure 8: Comparison of C_V coherence scores of HDP model for original dataset with different length. We observe the decrease in scores after 6 topics for all dataset sizes.

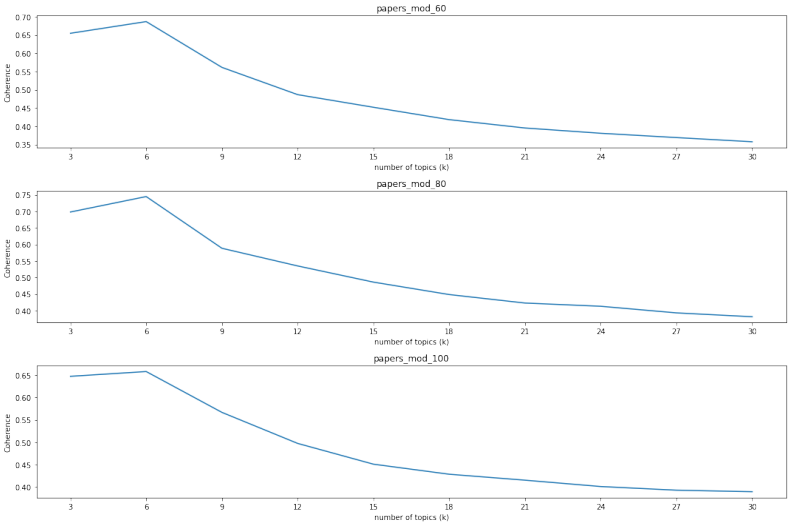

Figure 9: Comparison of C_V coherence scores of HDP model for modified dataset with different length. We observe the decrease in scores after 6 topics for all dataset sizes.