Dynamic topic modelling

Kirill Yakunin, Sanzhar Murzakhmetov*

Institute of Information and Computational Technologies, Kazakhstan

**Corresponding author's e-mail: sanzharmrz@gmail.com*

Abstract

Nowadays topic modelling is actively developing field in many industries. For instance In fintech companies, topic modelling is used for clustering transactions and detect latent interests of holders, in recruiting industry topic models can be used in ranking candidates CV, etc. Many of these applications can create impressive value for companies. One of the interesting research areas is dynamic approach to topic modelling, which can be applied to objects related to some point in time (for example news publications). The key idea is to perform separate topic modelling on different time intervals in order to find which topics continue to exist in time and to what extent and how the content of the topics change.

Keywords: topic modelling, bigartm, nlp

1 Introduction

Today, topic modelling is an important part in modern natural language processing tools, this algorithm uses a matrix of unique words and documents, called corpus, and formally, decomposes current matrix, with many specific restrictions on received matrices. New dimension of resulting matrices is searched hidden topic space.

In the variety of modern implementations of topic modelling BigARTM is considered to be state-of-the art model [1], implemented as a cross-platform library with Python API and parallel processing, main advantage of BigARTM, is a stack of regularization techniques.

Topic modelling on documents with timestamps can be performed in order to analyze how topics persist and transform in time. The main issue here is to create an algorithm to map topics from different topic-modelling, considering each topic as a weighted bag of vectors of words. This issue the main scientific problem of the research, since currently existing algorithms for comparing two weighted bags of words have significant flaws: for example, WMD is a very low-performance algorithm, while Jaccard distance is low quality, since it doesn't take vector semantic representation into account.

2 Decision

One of the ways to detect topic transformations, on a timeline is building nested topic models, and trying to connect topics in every iteration. There are a number of parameters which need to be optimized, such as how big should be the overlay for the selected volume of topic modelling, how we should find connections between topics, etc. In these cases, there are logical heuristics, for example in big overlapping value, links between topics will be more symbolic and less semantic, while large volume of topic models and low number of topics will lead to links between topics will be too predictable.

3 Conclusion

Dynamic topic modelling is a nontrivial problem in the natural language processing ecosystem, and can have different applications in industry, for example media analytics [2], fakes detection [3] or smart news aggregation algorithms.

The main scientific interest of this problem is represented by finding an optimal topic-mapping algorithm, optimizing time intervals and interval steps for different practical problems. Visualization of the results is also an interesting engineering problem.

Acknowledgments

The work was funded by grant No. BR05236839 of the Ministry of Education and Science of the Republic of Kazakhstan.

References

[1] Vorontsov K V, Potapenko A A 2014 Additive regularization of topic models *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization*

[2] Mukhamedyev R, etc. 2019 Assessment of the dynamics of publication activity in the field of natural language processing and deep learning. *Proceedings of the International Conference on Digital Transformation and Global Society (St. Petersburg, JUNE 19–21, 2019). Springer, Cham. 130–5*

[3] Barakhnin V B, Kuchin Ia I, Muhamedyev R I 2018 On the problem of identification of fake news and of the algorithms for monitoring them *Proceedings of the III International Conference on Informatics and Applied Mathematics (Almaty, Kazakhstan, The Institute of Information and Computational Technologies, September 26-29, 2018) 113–8*