Karlsruhe Institute of Technology

Visualization of Topic Models for Patent Documents

BACHELOR THESIS

of

Paolo Ramella-Ratin

Student-ID: 1958685

Study programme: Industrial Engineering and Management

Institute of Economics (ECON)

Chair in Economic Policy

Examiner: Prof. Dr. Ingrid Ott

Supervisor: M. Sc. David Bälz

Time period: 01.05.2019 – 31.10.2019

Contents

1 Exposé

1.1 Motivation

With the digitalization of our societies and in particular the economy, we have seen a rise in new technologies. Those technologies have fuelled patent applications by corporations seeking to protect their intellectual property. Patents which are listed and organized by international organisations like the World Intellectual Property Organization therefore represent a colossal document corpus and could serve as a window giving new insights into industries and technologies. Thanks to the developments in the fields of machine learning and natural language processing we are now able to systematically analyse aforementioned documents. The popular text mining methods called topic models are well suited for understanding a corpus of documents by discovering which topics are represented in it. Unfortunately simply applying existing algorithms to a set of data is not enough. It is essential to present and visualize results in a way that is tailored to the field of application. In this context visualization serves as a way to communicate, compare and understand the work done by data scientists and researchers. This thesis is about the visualization of topic models applied to patent documents.

1.2 Research Questions

In the context of this bachelor thesis we seek to explore the different but related following questions:

- What existing individual methods (i.e word-clouds etc.) and implemented tools are referenced in literature?
- What should a visualization of topic models consist of? Are there specific points which cater to topic models of patent documents? What (meta-)information about patents or topic models should be included to improve the visualization? Using what metrics can we compare different topic models?

- Finally we want to answer the question of implementation. What is the best visualization tool we can implement? How can we make use of techniques like dashboards and agile visualization?

1.3 Literature

The Book "Applications of topic models" (Boyd-Graber et al. 2017) serves as a foundation in understanding topic models and obtaining an overview of visualization concepts. In addition, a few chapters of the book address the visualization of specific types of documents which resemble patents. We reference "Concurrent visualization of relationships between words and topics in topic models" (Smith et al. 2014) when implementing a visualization allowing users to better understand the relationship between topic models and the words they are made out of.

As scientific papers on existing visualization tools for topic models we use "The topic browser: An interactive tool for browsing topic models" (Gardner et al. 2010), "TopicViz: interactive topic exploration in document collections" (Eisenstein et al. 2012) and "Visualizing topic models" (Chaney and Blei 2012) as guidance on the final implementation.

In regards to evaluation of topic models we reference following literature "External evaluation of topic models" (Newman, Karimi, et al. 2009), "Automatic Evaluation of Topic Coherence" (Newman, Lau, et al. 2010) and "Optimizing Semantic Coherence in Topic Models" (Mimno et al. 2011). All three research papers provide ample work on measuring topic models against each other and individually.

1.4 Implementation

As of the implementation of the tool we limit ourselves to using python. This approach allows to avoid complications related to learning a new programming language. We are nonetheless using libraries which are based on javascript since they allow for sophisticated visualizations. Our main and most general goal is to create an agile visualization tool which allows user interaction. The tool takes on the form of a dashboard.

With these parameters in mind two python libraries seem appropriate for the task. The Pyvis[1] and Plotly[2] Python packages.

[1] `https://github.com/WestHealth/pyvis`
[2] `https://github.com/plotly`

The Pyvis library is useful for implementing word clouds since it is build for visualizing interactive networks. This allows the creation of word clouds which go beyond the basic definition and display additional information and relationships between topics, words and patents.

The dashboard is build using the Plotly package since it has a library solely dedicated to creating dashboards. The standard Plotly library is used for interactive graphing in python.

Both libraries are well documented and provide user-friendly examples.

1.5 Preliminary Structure

The structure of the thesis loosely follows the structure of this exposé. After a concise motivation and the presentation of our research questions, readers are lead through an introduction in the fields of topic models as well as the possibilities of visualization. Readers also learn about available metrics used for comparing multiple topic models. Afterwards we give an overview on the composition of patent documents as well as metadata and relational data which is of interest for the implementation. Finally we give insights into the implemented visualization tool in addition to examining if the research goals have adequately been met.

1.6 Timetable

In the first two weeks we will review aforementioned literature. During this phase we will extract and synthesize ideas for the implementation.

During the implementation phase we will follow the spiral model. This approach will allow setting our goals, implementing and reviewing them periodically. We chose this model in order to quickly develop a working prototype and improving it incrementally. This way we are able to improve the tool given our time constraints.

Both during the literature and the implementation phase we will work on intermediate presentations and writing the thesis. The bulk of this work will be done after the implementation phase has been completed.

References

Boyd-Graber, J.; Hu, Y.; Mimno, D., et al. (2017): Applications of topic models. In: *Foundations and Trends® in Information Retrieval* 11(2-3), pp. 143–296.

Chaney, A. J.-B.; Blei, D. M. (2012): "Visualizing topic models". In: *Sixth international AAAI conference on weblogs and social media*.

Eisenstein, J.; Chau, D. H.; Kittur, A.; Xing, E. (2012): "TopicViz: interactive topic exploration in document collections". In: *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, pp. 2177–2182.

Gardner, M. J.; Lutes, J.; Lund, J.; Hansen, J.; Walker, D.; Ringger, E.; Seppi, K. (2010): "The topic browser: An interactive tool for browsing topic models". In: *NIPS Workshop on Challenges of Data Visualization*. Vol. 2. Whistler Canada.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; McCallum, A. (2011): "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 262–272. URL: http://dl.acm.org/citation.cfm?id=2145432.2145462.

Newman, D.; Karimi, S.; Cavedon, L. (2009): "External evaluation of topic models". In: *in Australasian Doc. Comp. Symp., 2009*. Citeseer.

Newman, D.; Lau, J. H.; Grieser, K.; Baldwin, T. (2010): "Automatic Evaluation of Topic Coherence". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Los Angeles, California: Association for Computational Linguistics, pp. 100–108. URL: http://dl.acm.org/citation.cfm?id=1857999.1858011.

Smith, A.; Chuang, J.; Hu, Y.; Boyd-Graber, J.; Findlater, L. (2014): "Concurrent visualization of relationships between words and topics in topic models". In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 79–82.