UNIVERSITY COLLEGE LONDON

MASTERS THESIS

Dynamic Topic Modeling of PATSTAT Patents Using LDA

Author:
Christopher MARTIN

Supervisor:
Dr. John Shawe TAYLOR

A thesis submitted in fulfillment of the requirements for the degree of Masters of Science

in the

Research Group Name
UCL Computer Science

August 9, 2016

Declaration of Authorship

I, Christopher MARTIN, declare that this thesis titled, "Dynamic Topic Modeling of PATSTAT Patents Using LDA" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITY COLLEGE LONDON

Abstract

Faculty Name
UCL Computer Science

Masters of Science

Dynamic Topic Modeling of PATSTAT Patents Using LDA

by Christopher MARTIN

In this paper we evaluate the performance of a time varying family of LDA based topic models meant to capture both the underlying semantic structure of a document collection and the evolution of that structure in time. Such models are useful for illustrating changes in the use of language regarding specialized subject matter and provide a window into the progression of that change. We compare these models to traditional topic models such as LDA as a benchmark and explore the efficacy of such models in a range of applications including document classification, clustering, and influence prediction. Finally, we present results on over 18 years of patent data from the PATSTAT database across X classes of patents demonstrating interpretable trends, better document classification and clustering, and improved topic coherence.

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor…

Contents

List of Figures

List of Tables

Chapter 1

Introduction to the Thesis Topic

1.1 Motivation and Background

1.1.1 layout

- Chapter 1: Introduction to the thesis topic
- Chapter 2: Background information and theory
- Chapter 3: (Laboratory) experimental setup
- Chapter 4: Details of experiment 1
- Chapter 5: Details of experiment 2
- Chapter 6: Discussion of the experimental results
- Chapter 7: Conclusion and future directions

This chapter layout is specialised for the experimental sciences.

Figures – this folder contains all figures for the thesis. These are the final images that will go into the thesis document.

1.1.2 References

Blei and Lafferty, 2006 (Chang et al., 2009) (Rosner et al., 2014) (Wang, Blei, and Heckerman, 2012) (Hall, Jurafsky, and Manning, 2008) (Wang and McCallum, 2006) **conf/icdm/AlSumaitBD08** (Gerrish and Blei, 2010)

Multiple references are separated by semicolons (e.g. (Blei and Lafferty, 2006; Wang and McCallum, 2006)) and

Scientific references should come *before* the punctuation mark if there is one (such as a comma or period). The same goes for footnotes[1].states: "Footnote numbers should be superscripted, [...], following any punctuation mark except a dash." The Chicago manual of style states: "A note number should be placed at the end of a sentence or clause. The number follows any punctuation mark except the dash, which it precedes. It follows a closing parenthesis."

The bibliography is typeset with references listed in alphabetical order by the first author's last name. This is similar to the APA referencing style. To see how LaTeX typesets the bibliography, have a look at the very end of this document (or just click on the reference number links in in-text citations).

[1] Such as this footnote, here down at the bottom of the page.

TABLE 1.1: The effects of treatments X and Y on the four groups studied.

Groups	Treatment X	Treatment Y
1	0.2	0.8
2	0.17	0.7
3	0.24	0.75
4	0.68	0.3

A Note on bibtex

The bibtex backend used in the template by default does not correctly handle unicode character encoding (i.e. "international" characters). You may see a warning about this in the compilation log and, if your references contain unicode characters, they may not show up correctly or at all. The solution to this is to use the biber backend instead of the outdated bibtex backend. This is done by finding this in **`main.tex`**: `backend=bibtex` and changing it to `backend=biber`. You will then need to delete all auxiliary BibTeX files and navigate to the template directory in your terminal (command prompt). Once there, simply type `biber main` and biber will compile your bibliography. You can then compile **`main.tex`** as normal and your bibliography will be updated. An alternative is to set up your LaTeX editor to compile with biber instead of bibtex, see here for how to do this for various editors.

1.1.3 Tables

Tables are an important way of displaying your results, below is an example table which was generated with this code:

```
\begin{table}
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\tabhead{Groups} & \tabhead{Treatment X} & \tabhead{Treatment Y} \\
\midrule
1 & 0.2 & 0.8\\
2 & 0.17 & 0.7\\
3 & 0.24 & 0.75\\
4 & 0.68 & 0.3\\
\bottomrule\\
\end{tabular}
\end{table}
```

You can reference tables with `\ref{<label>}` where the label is defined within the table environment. See **`Chapter1.tex`** for an example of the label and citation (e.g. Table 1.1).

1.1.4 Figures

$$E = mc^2 \tag{1.1}$$

FIGURE 1.1: An electron (artist's impression).

1.2 In Closing

Guide written by —
Sunil Patel: www.sunilpatel.co.uk
Vel: LaTeXTemplates.com

Chapter 2

Background Information and Theory

2.1 Literature Review

2.2 DTM model

2.2.1 Issues

addresses several latent structures in the document collection such as topic evolution and prevalence, however does not address the birth and death of topics, like models such as Ahmed and Xing, 2012.

2.3 DIM model

Chapter 3

Experimental Set Up

3.1 Data Prep and Considerations

Chapter 4

DTM Results and Insights

4.1 Topics Through Time

4.1.1 validating topic histories in technology

Chapter 5

DIM Results and Insights

5.1 Influence Metric

5.1.1 validating influential patents

5.1.2 correlation with forward citations

5.1.3 correlation with page-rank

Chapter 6

Performance Evaluation

6.0.1 what we're not doing

predictive perplexity, as it is not always correlated with human opinion.

6.1 Classification

6.2 Clustering

Chapter 7

Usefulness in Other Models

7.1 Economic Model

influence can be used as a proxy for forward citations when citations are not available. used as a gamma in model for likelihood of innovation

Chapter 8

Conclusions and Future Work

8.1 Conclusions

8.2 Future Work

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

Ahmed, Amr and Eric P. Xing (2012). "Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream". In: *CoRR* abs/1203.3463. URL: http://arxiv.org/abs/1203.3463.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: ACM, pp. 113–120. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143859. URL: http://doi.acm.org/10.1145/1143844.1143859.

Chang, Jonathan et al. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Neural Information Processing Systems*. Vancouver, BC. URL: docs/nips2009-rtl.pdf.

Gerrish, Sean and David M. Blei (2010). "A Language-based Approach to Measuring Scholarly Impact". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Ed. by Johannes Fürnkranz and Thorsten Joachims. Omnipress, pp. 375–382. URL: http://www.icml2010.org/papers/384.pdf.

Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008). "Studying the History of Ideas Using Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Honolulu, Hawaii: Association for Computational Linguistics, pp. 363–371. URL: http://dl.acm.org/citation.cfm?id=1613715.1613763.

Rosner, Frank et al. (2014). "Evaluating topic coherence measures". In: *CoRR* abs/1403.6397. URL: http://arxiv.org/abs/1403.6397.

Wang, Chong, David M. Blei, and David Heckerman (2012). "Continuous Time Dynamic Topic Models". In: *CoRR* abs/1206.3298. URL: http://arxiv.org/abs/1206.3298.

Wang, Xuerui and Andrew McCallum (2006). "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*. KDD '06. Philadelphia, PA, USA: ACM, pp. 424–433. ISBN: 1-59593-339-5. DOI: 10.1145/1150402.1150450. URL: http://doi.acm.org/10.1145/1150402.1150450.