

Topic Models Application on e-discovery

Tiago Zortea¹, James Colee¹

¹ Computer & Information Science & Engineering Department – University of Florida



INTRODUCTION

With the amount of digital information stored growing every year, it is becoming more and more costly for attorneys to weed out relevant information from this massive digital data. In this context, e-discovery refers to the process of using computer algorithms to aid attorneys on “discovering” relevant information from large electronic datasets which may include emails, voicemails, files from computers and databases. In this study we evaluate the usage of topic modeling in e-discovery with the objective of aiding the task of finding relevant documents among large datasets of digital data.

OBJECTIVES

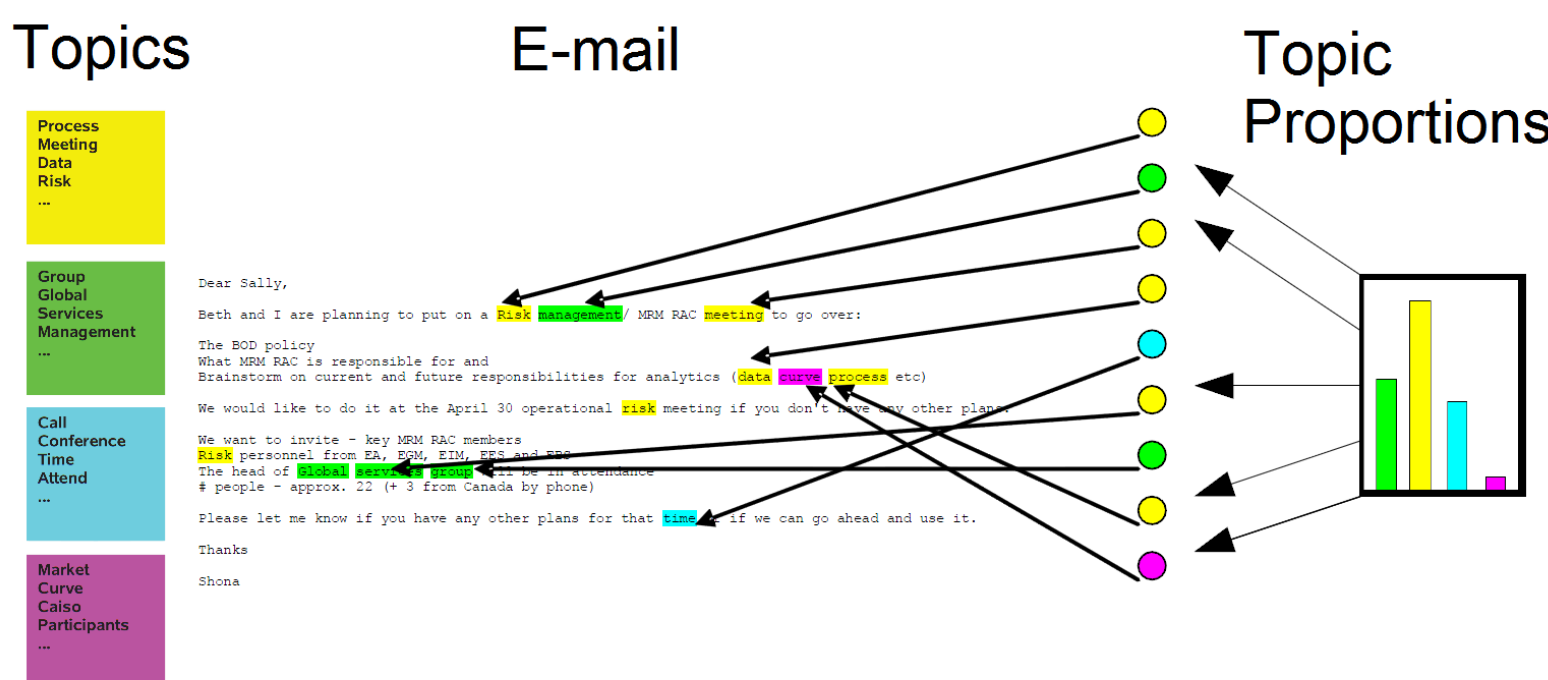
The main objective of this study is to improve the understanding of the behavior of topic models for e-discovery. With this knowledge, it will be possible to better plan its interface and know how to use topic models as a feature for smart ranking in ediscovery.

The specific objectives include:

- Execute topic models over the Enron dataset using LDA
- Create a Python UI for topics visualization
- Improve the responsiveness of this UI over such a large dataset
- Create a metric of reliability of topics so they can be ordered by it
- Analyse the usefulness of topic modeling in e-discovery

MATERIALS & METHODS

Topic modeling is a statistical model which aims to cluster text documents with similar topics in an unsupervised procedure. The topics are formed by a probability of every word in a dictionary co-occurring with others, so words that occur more frequently in the same documents induce topics. The topics also have a distribution over documents, so each document has a probability of containing few topics. For this work, latent Dirichlet allocation (LDA) was used as the algorithm.



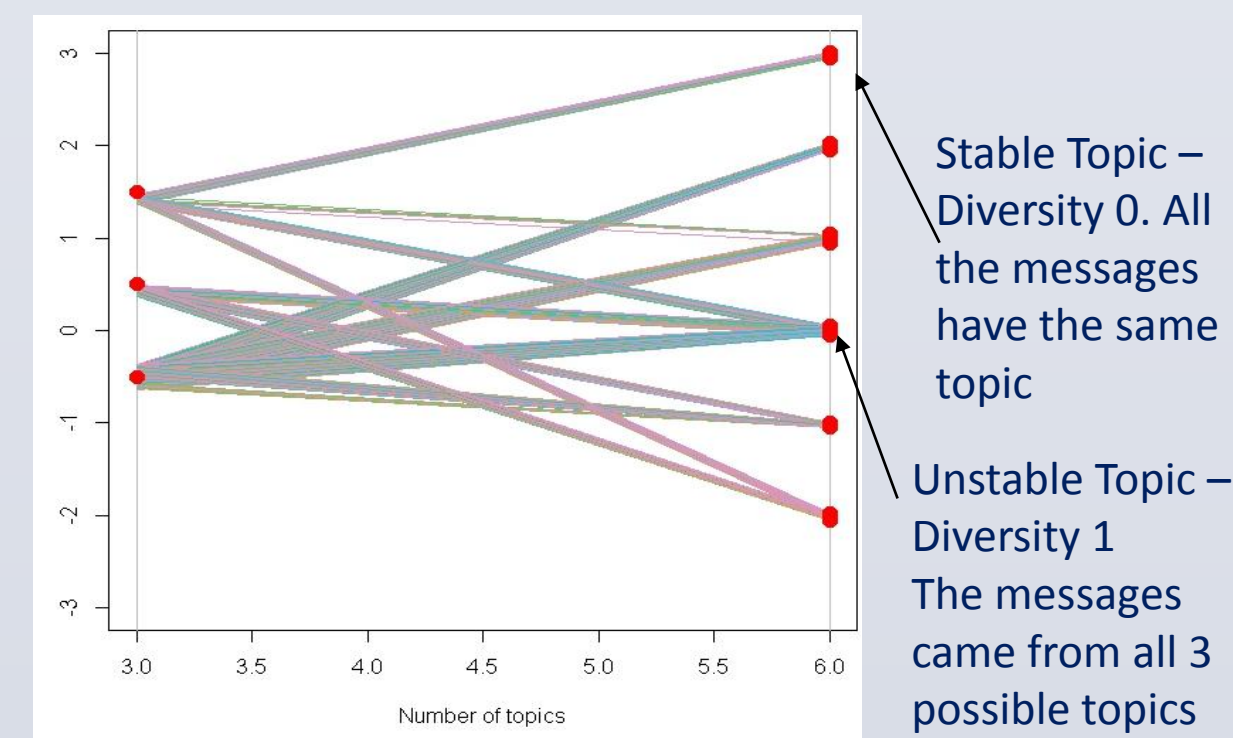
THE ENRON CORPUS

As the information used in litigations are often sensitive, it was needed for this study an already widely available dataset which is large enough to provide a realistic test environment. For this purpose it the Enron corpus was used which was extracted from the Enron servers after the Eron scandal in 2001. This corpus is composed of more than 500,000 emails exchanged between 158 Enron employees summing to about 6Gb of text data.

To perform the topic modeling calculation on the Enron corpus, the Gensim python package was used. This package has a scalable distributed implementation of LDA.

TOPICS DIVERSITY

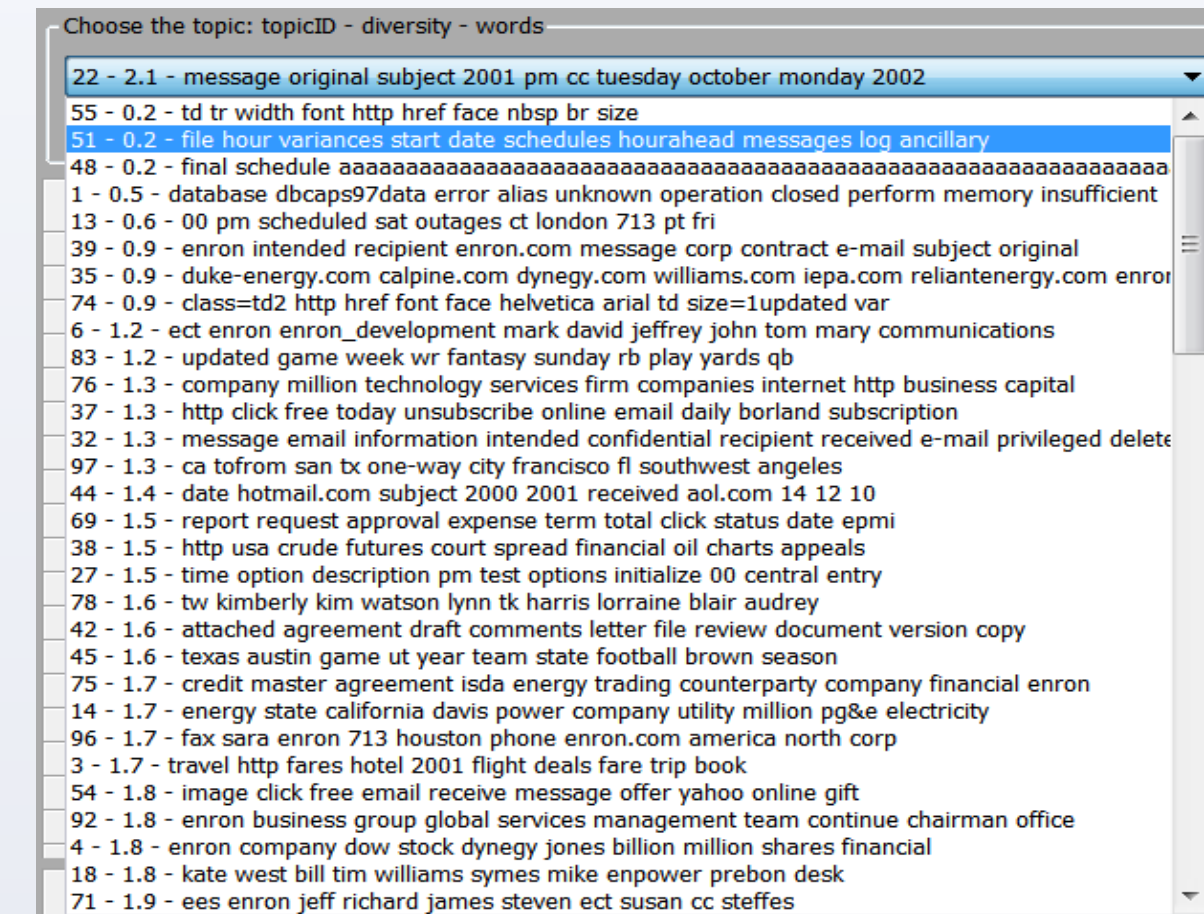
As topics are formed out of words typically occurring together, it is expected that some topics end up clustering “background” words and having little meaning for the human observer. For that reason we created a topic diversity measurement to be able to separate the stable topics from the background topics. The approach is to generate two different numbers of topics from the same corpus, then observe how the documents map between these 2 different clusterings. If a topic contains almost the same documents in both numbers or clusterings, then it is a solid topic which probably is based in very specific words. But, if the documents from a topic are formed from topics scattered all over the place, then it is probably unreliable.



At first, the results of topic modeling over the Enron dataset looked very sparse and lacking on meaning, but with the produced visualization tool and with correct ordering of topics, documents and probabilities of topics over topics, the inner structure of topics became evident.

RESULTS

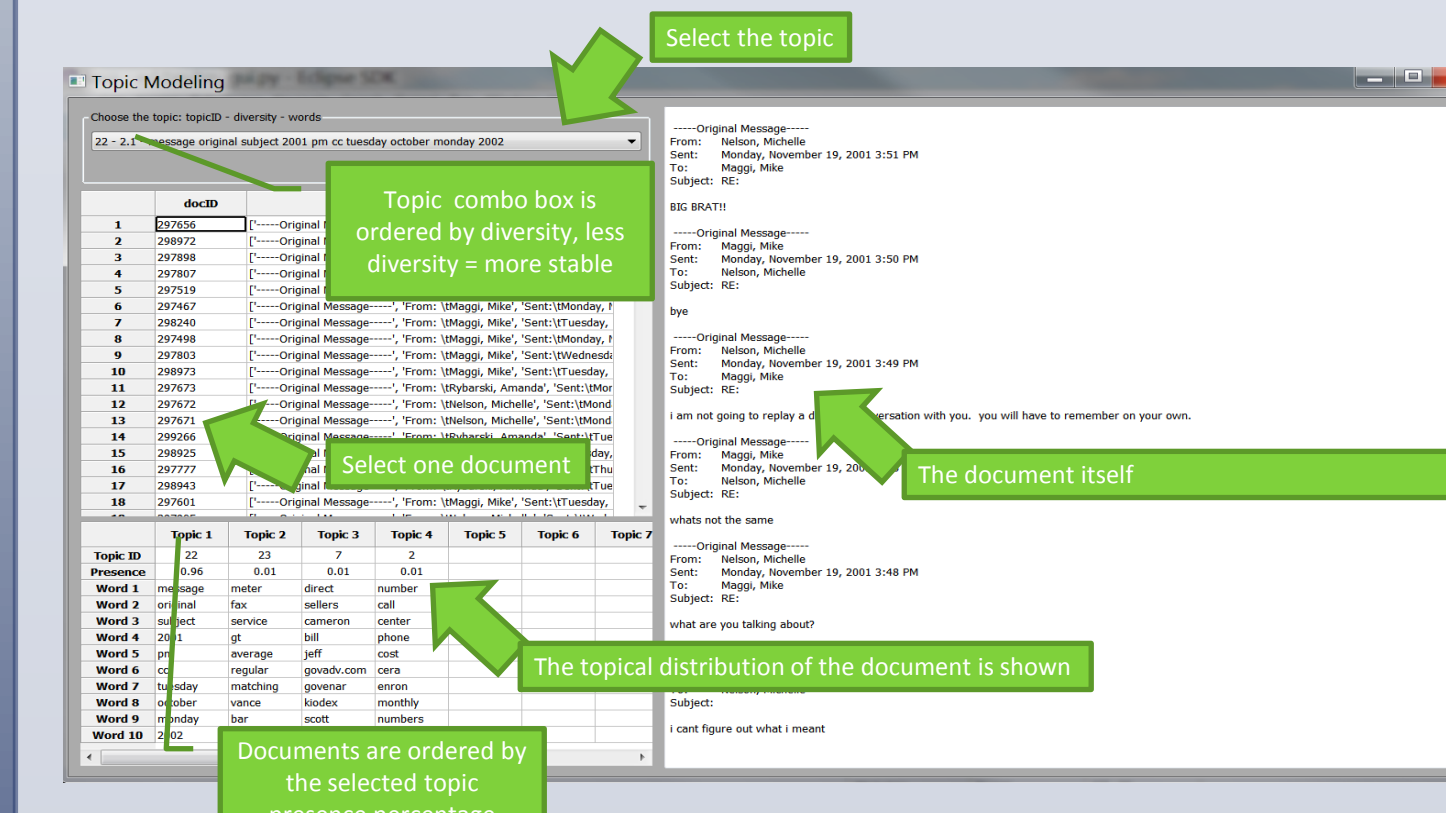
The LDA algorithm was successfully run for the Enron dataset and a python UI was created to allow better exploration of the topics generated..



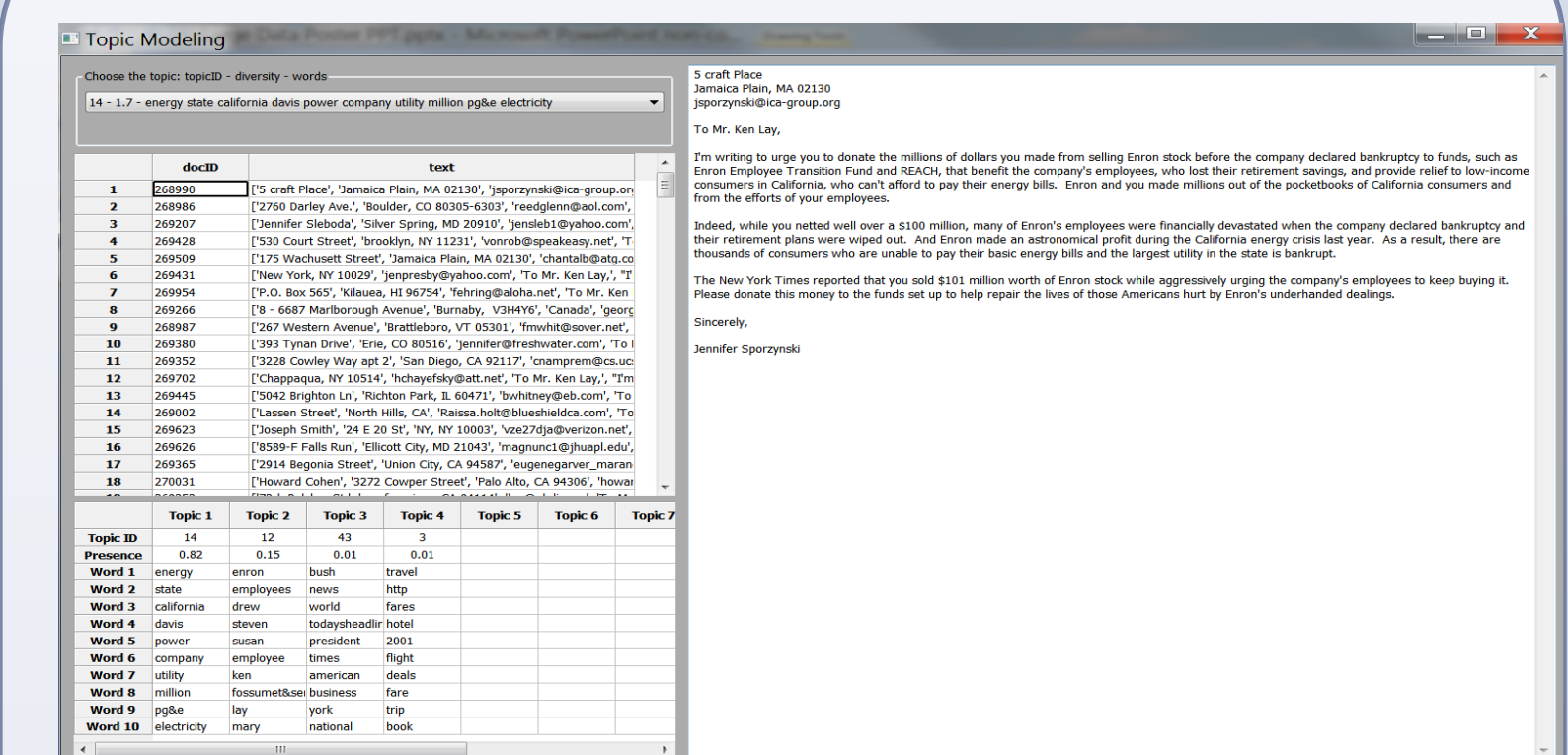
Python UI combobox of the topics generated, when a topic is selected the list of documents is ordered by biggest topical presence of the selected topic.

This is a list of some meaning topics generated.

- Topic 55 - Groups most of the emails that contain HTML
- Topic 51 and 48 - Groups many system logs
- Topic 16 - Groups seems to group up many "motivational" messages
- Topic 22 - Groups many short personal messages with many small replies (chitchat)
- Topic 39 and 32 - Groups short messages with a long confidentiality signature
- Topic 83 - Groups updates on a internal fantasy game
- Topic 42 - Groups short emails with attachments
- Topic 25 - Groups emails about the performance feedback of employees
- Topic 2 - Groups conference call notifications
- Topic 3 - Groups travel/hotel scheduling and fares
- Topic 11 - Groups emails about hiring processes
- Topic 69 - Groups internal reports request



The topic 14 is an interesting example of topic which could be useful for in a legal case, it groups emails from hundreds of employees to Mr. Ken Lay (one of the convicted high executives) containing a desperate plea for him to donate the millions he made with his stock transactions to help out with the employee retirement funds.



Topic 14 visualization on the python system.

CONCLUSION

It was possible to observe that topic modeling has great potential of usage in the e-discovery field. It was able to create clusters of similar emails without any human intervention. These clusters will be useful in the future to help having a smart raking of documents in the software.

The contributions of this work are:

- A python UI for visualization of topic modeling over large datasets of emails
- Insights about the usage of topic modeling for e-discovery
- A new approach for measuring topic stability
- The results of topic modeling over the Enron corpus

However, more research is needed to allow automatic tuning of the parameters for any dataset. In the future this software is supposed to be able to run in the user's computer with any dataset.

REFERENCES

D. Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

Loulwah Alsumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic Significance Ranking of LDA Generative Models. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I

Shannon, C. E. (1948) A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423 and 623-656.

ACKNOWLEDGEMENTS

The authors acknowledge the PhD student Clint P. George for providing a base python code for the LDA calculation as well as useful advice.

CONTACT

Tiago Zortea (zortea@ufl.edu)
James Colee (colee@ufl.edu)