The weight of words when you are judged

Davide Carletti and Francesco Tomaselli

University of Milan

Abstract. The goal of the project …

Keywords: Legal texts · Topic modelling · Word embeddings

1 Introduction

The goal of this project is to study words among the temporal axis in court decisions. More precisely, we aim to find information about relevance and frequency of single or multiple terms, and also top discover correlations between them.

2 Research question and methodology

The goal of the project is to analyze, from a collection of court decisions, the relevance of terms in time while also gathering information about possible correlations between them. The reasons to perform such a study are various, firstly, one could find interesting and unexpected correlations between terms, also, studying them in time could reveal a change of context of some words.

3 Experimental results

3.1 Dataset preprocessing

The dataset in use is the Illinois Bulk Dataset, that contains 183146 cases with 194366 judges opinions.

The first step of the preprocessing is to merge the opinions about a case into one, obtaining a single document for each dataset entry. Then, each document goes through a text cleaning and tokenization phase, where the first part is done with the help of regular expressions, while the second uses Spacy to obtain a list of terms.

3.2 Topic modelling

To have an overview of the topics discussed on the dataset a Latent Dirichlet Allocation model is trained on the tokenized texts. One of the key parameters of such a model is the number of topics, and, given the fact that cases could potentially talk about anything, an *Halving search* is performed to find a good value. [1]

We opt for an halving search since the number of topics could be anything, we fixed a range between ten and thirty and training each model to then evaluate results would take a huge amount of time. Halving search mitigates the problem, as it trains firstly on smaller datasets, select the best models, and retrain on bigger slices of data until a final model is found. This methodology can be ten times faster then grid search. The selection criterium for the search is the Log Likelihood of each model. The search revealed that the optimal number of topics is 14.

3.3 Temporal word embeddings

3.4 Webapp

4 Concluding remarks

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)