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. [**?**]

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.

The resulting topics are promising but the words of interest, namely a collection of narcotics, investigation and weapons terms, are merged together in a few topics. To solve the issue we decide to run topic modelling on a subset of the previous topics with the same technique as before. The result is similar, we find again 14 topics, but this time they are much more specific.

3.3 Temporal word embeddings

One of the objective of the project is to find correlations among words and one tool that can be used is word embeddings. The technique assigns a real vector of a given dimensions to each word in the document collection, creating a way to directly compare the context similarity between two terms.

For this task we used Gensim's Word2Vec implementation and trained different models in three ways:

1. *Global model*: trained on the entire document collection, with 100 components vectors;
2. *One year models*: they are trained on subsets of the dataset divided by years;
3. *Ten years models*: trained on epochs of 10 years each.

The first model gives information about the whole dataset, it can be used to compute similar words queries, while the others can be exploited to find context and semantic shifts among the temporal axis. Taking inspiration from the HistWords [1] work on semantic shift, we start by aligning the models, and then find, given a word and a base year, the similarity of that word in time with respect to the base year. This approach can reveal if a word changed semantic or context, and when, with respect to a given year, an example can be seen in 1 on the facing page.

Similarly, the ten year models are aligned, but this time, given a term, we compute the difference among two consecutive epochs. The idea is similar to the previous one but slightly different, this time a drop in the similarity sequence reveals a change of meaning from an epoch to the other, the resulting sequence can be seen in 2 on the next page.

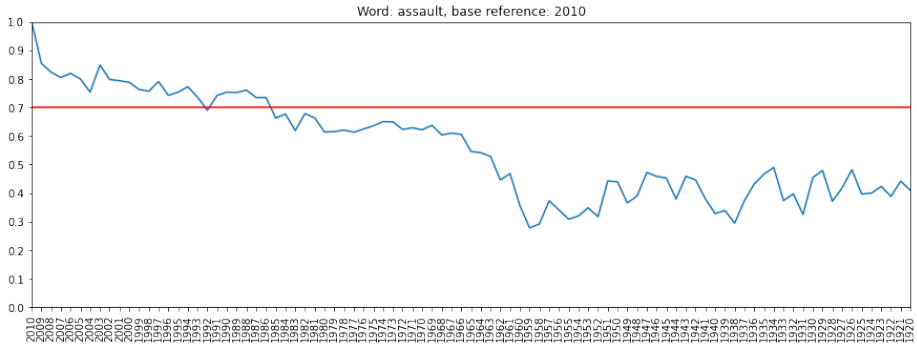

Fig. 1. Semantic shift of the word assault with respect to the year 2010, we can see a drop if similarity around 1960.

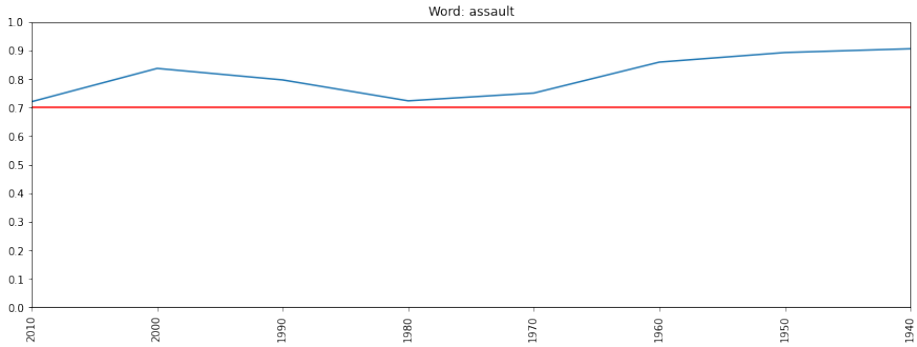

Fig. 2. Similarity between consecutive epochs of the term assault, we can see a drop between 1980-1970 and 1970-1960.

3.4 Webapp

4 Concluding remarks

References

1. William L. Hamilton, Jure Leskovec, and Dan Jurafsky. ACL 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.