The Weight of Words When You Are Judged

Picascia Sergio
Data Science and Economics
University of Milan

Abstract

One of the main application of information retrieval is in the field of law, where legal professionals need effective systems in order to access the required knowledge. Classic boolean search method have been found not able to achieve high performance in terms or precision and recall and, thus, techniques based on artificial intelligence and statistics have been employed. In this research, the focus will be on identifying and retrieving relevant terms for each of the document in the dataset. The data consists in court decisions of the state of Illinois and, in particular, opinions will the subject of analysis. A Word2Vec model is implemented in order to better allocate words from a legal jargon in the vector space, while the TF-IDF scores will be employed to define which are the most relevant term for a specific opinion. Finally, it will be possible to perform inference based on the time trends of these terms and the correlation coefficients computed among them.

Introduction

Legal informatics is a fast-pacing growing field, thanks to the increase in the amount of legal documents available in digital formats. Legal information retrieval is a part of it which focuses on providing access to the law to people working in this sector. Legal texts can assume different forms and, in particular, it is possible to identify two main types: the law, as the set of legislative rules which a particular country or community recognizes as regulating the actions of its members, and the court decisions, the decision of a court regarding the rights and liabilities of parties in a legal action or proceeding.

Due to the complexity of these documents, the lexicon used, the structure of phrases, tasks like text analysis or information retrieval result being very challenging. In literature, two broad approaches have been developed in order to solve this issue: those based on knowledge engineering, which focuses on artificial intelligence and case-based reasoning, and others based on natural language processing, which instead follows a statistical approach. The focus of this study will be on retrieving specific terms for each court decision, estimating their relevance and frequency, analyzing their behavior over the temporal dimension and computing the correlation between them.

Research Question and Methodology

The project consists in building a model which is able to retrieve relevant terms for each document, implementing techniques for embedding the words and using measurements of frequency in order to estimate the degree of relevance and occurrence of each term; the final goal is to make inference about the terminological trends, analyzing the temporal dimension, and compute correlations among terms. In particular, the focus will be on a selection of items, regarding narcotics, weapons, different types of crime and demographic aspects.

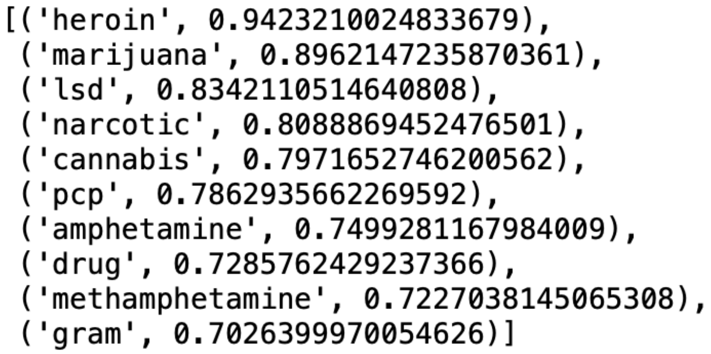
Firstly, a word embedding model has been estimated and used as a tool for document retrieval. In fact, thanks to the implementation of Word2Vec, mapping terms onto vectors allows to quickly find which are the words that are most similar to each other and, therefore, expand the query when performing a search over the whole corpus. Thus, the retrieval of relevant documents does not rely solely on the keywords used as inquiry, but on a whole set of terms related to the same context.

However, the task of document retrieval can not be based only on the occurrences or not of words inside the text, but it needs another kind of measurements in order to understand if a specific term can be considered relevant to the document, rather than simply appearing in it. The technique used to accomplish this goal is the TF-IDF, which allows to compute the relevance of each word in the text, based on its occurrences in the document taken into consideration, but also on the occurrences in all the other documents forming the corpus. This method, combined with the Word2Vec model defined above, provides the ability of searching for documents for which the query, and also terms similar to it, are truly relevant. Therefore, the results will be a set of court decisions, regarding a certain topic, that can be quickly used for further analysis.

At this point, performing inference is not a demanding task, even though some clarifications must be done before proceeding. First of all, it is important to understand the distribution of the observations over time, since an increase in court decisions about a certain topic may only be due to an overall increase; then, considering the relative frequency of documents rather than the absolute one may help in evaluating the temporal behavior of terms. Secondly, correlation among terms can be computed on data where the terms themselves are variables: a document-term matrix may be a clear example of that. It happens that we already have data in this format, after having used the TF-IDF method for terms relevance. Correlations computed on those values can indicate if words are likely to be more or less relevant for the same documents, even though they refer to different context and, therefore, not found as similar in the word embeddings.

Experimental Results

The analysis has been conducted on the Caselaw Access Project (CAP) dataset, provided by the Harvard Law School. The data is divided according to the dif-



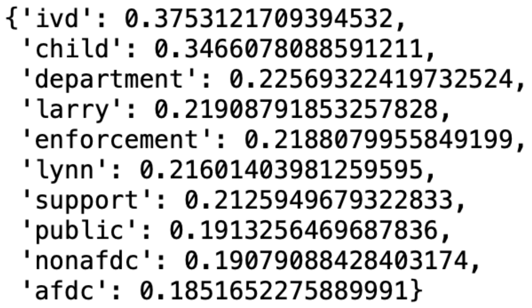
```
[('heroin', 0.9423210024833679),
 ('marijuana', 0.8962147235870361),
 ('lsd', 0.8342110514640808),
 ('narcotic', 0.8088869452476501),
 ('cannabis', 0.7971652746200562),
 ('pcp', 0.7862935662269592),
 ('amphetamine', 0.7499281167984009),
 ('drug', 0.7285762429237366),
 ('methamphetamine', 0.7227038145065308),
 ('gram', 0.7026399970054626)]
```

Figure 1: Top 10 most similar words to 'cocaine' according to the Word2Vec model.

ferent jurisdictions and, for the scope of this research, only the portion of the dataset regarding the state of Illinois has been taken into consideration. It is composed by 183.146 court decisions and, since each decision may have more than one opinion, the final corpus consists in a total of 194.366 opinions. Each court decision presents different metadata: case ID, attorneys and judges taking part to the trial, date, opinions and many others. Since this project focuses mainly on the task of document retrieval, the data useful at this cause will be opinions and, for the analysis of the temporal dimension, the date.

After having imported the data and retrieved useful information, the following step regards the preprocessing of the text: removal of special characters, digits and punctuation, fixing of contractions, characters lowering and lemmatization. This last task has been accomplished using the WordNet lemmatizer provided by the nltk library. The result of this process is a list of tokens for each opinion, that can be used as input for the following steps.

As explained in the previous section, a Word2Vec model has been used in order to build a vector representation of the words. Instead of using a pre-trained model, a new one has been built over the available data; this choice was lead by the fact that, on one hand, there was enough data available to let the model properly catch the similarities among terms and, on the other hand, already existing models were usually trained on general corpus, that may use a different jargon than the one used in a legal context. The model was trained over 10 epochs with the continuous bag of words (CBOW) method, a vector size of 150 and a minimum word count of 2; the final vocabulary counts 242.779 terms. To understand the potential of this approach, it is sufficient to choose a word and find which are the most similar ones according to the cosine similarity; as showed in the example below (Figure 1), looking for terms similar to 'cocaine', the results are all relevant with an high similarity score. For each of the area of interest defined before (narcotics, weapons, . . .), it has been chosen a selection of pre-defined terms, to which it has been added a set of most similar ones, ac-



```
{'ivd': 0.3753121709394532,
 'child': 0.3466078088591211,
 'department': 0.22569322419732524,
 'larry': 0.21908791853257828,
 'enforcement': 0.2188079955849199,
 'lynn': 0.21601403981259595,
 'support': 0.2125949679322833,
 'public': 0.1913256469687836,
 'nonafdc': 0.19079088428403174,
 'afdc': 0.1851652275889991}
```

Figure 2: The 10 most relevant terms for the first opinion based on the TF-IDF score.

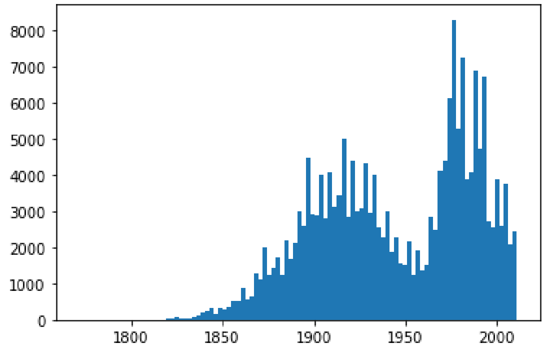


Figure 3: Distribution of opinions over years.

cording to a certain threshold, forming a list of words which will be used further in the analysis.

The implementation of the TF-IDF vectorizer from the 'scikit-learn' library has been used in order to estimate the relevance of each term per document. After having fitted the vectorizer over the corpus, excluding only stop words, the ten most relevant terms for each document have been retrieved and identified as representative for the document itself. Again, in the below example (Figure 2), it is possible to see how well the model performs in identifying the subject of the opinion (motion for child support) and the names of the people involved. Combining the TF-IDF scores with the Word2Vec model estimated before, allows to find which are the terms similar to a certain search query and retrieve the documents having one of these words as relevant.

The distribution of court decisions is not constant over time, as expected; there have been a substantial increase, especially at the end of the 19th century and in the 70's (Figure 3). Therefore, when retrieving document about a certain topic, instead of looking at the absolute count of court decisions per year, it would be a better to compute the relative frequency with respect to the total

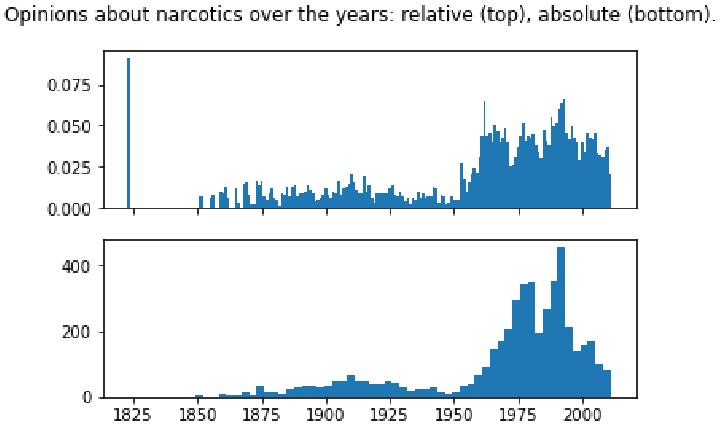


Figure 4: Distribution of opinions about narcotics over years.

number of observations. Following the previous steps, it is possible to define a set of terms regarding a certain context and retrieve documents for which those words are relevant; at this point, it would be sufficient to compute the relative frequency of these documents per year. Examples below show: the temporal trend of opinions regarding narcotics, with a comparison between absolute and relative frequency (Figure 4); a focus on the juxtaposition between opinions about marijuana and synthetic drugs (Figure 5); a view of the considerable change in court decisions regarding homosexuality (Figure 6).

Finally, it is possible to retrieve the TF-IDF scores for a set of selected terms and compute the Persons' correlation coefficient. The following examples show the correlation scores computed for the words 'gun', 'weapon', 'murder' and 'male' (Figure 7). Even though the example is trivial and the positive values are predictable, the coefficients are not large enough to make particular conclusion. An implementation of canonical correlation analysis has been also enforced, applying this method on two set of terms, weapons and crimes, without significant results.

Concluding Remarks

The performance of both Word2Vec and TF-IDF were satisfying since, as showed in the previous examples, the models are able to accurately catch the similarity between words and the relevance of the terms for each document respectively. The choice of building a new embedding model, based on the corpus used in the analysis, has turned out to the right one, being able to understand the semantics of the words in a context, the legal one, which is significantly different from a generic one. In the same manner, the TF-IDF vectorizer was capable of identifying the terms that were more relevant for each document, ensuring an easy and quick way to retrieve texts according to a certain query. In this case,

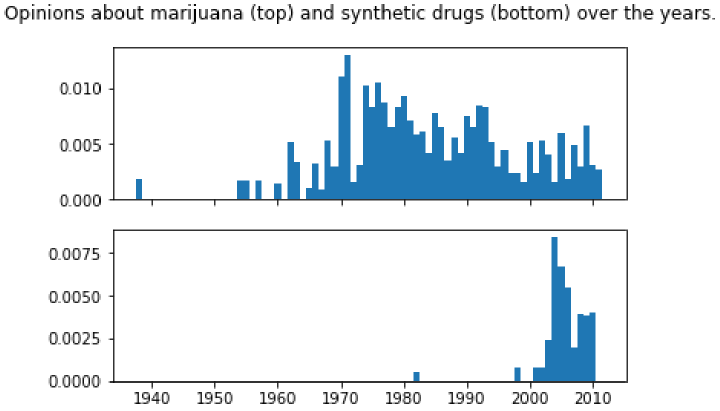


Figure 5: Distribution of opinions about marijuana and synthetic drugs over years.

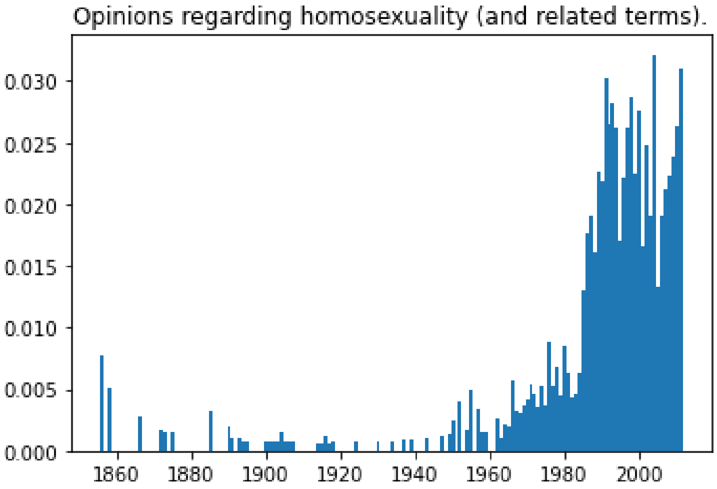
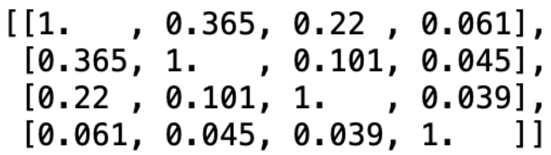


Figure 6: Distribution of opinions about homosexuality over years.



```
[[1.   , 0.365, 0.22 , 0.061],
 [0.365, 1.   , 0.101, 0.045],
 [0.22 , 0.101, 1.   , 0.039],
 [0.061, 0.045, 0.039, 1.   ]]
```

Figure 7: Correlation coefficients for the words ‘gun’, ‘weapon’, ‘murder’ and ‘male’.

however, it was necessary to exclude stop words that were indeed considered as relevant in many documents by a first version of the vectorizer.

Further development in performances can be expected increasing the vector size of the Word2Vec model, from 150 to 300, and including n-grams rather than only single words. If, on one side, this would increase a lot the need for computational resources, on the other hand it would ensure an even better representation of the embedding space, comprising also expressions made by more than one word, which are common in the legal jargon (serial killer, sexual assault, ...). For the remaining part of the analysis, trend and correlation evaluation, a consult with an expert in the legal field would benefit in order to extract more knowledge from the model at disposal. As always, when dealing with data, it is extremely important to pose the right question or, as in this case, the right queries, the ones that can truly help people working in the field gaining the information they need.