THE IMPORTANCE OF WORDS IN THE LEGAL CONTEXT

MSC. DATA SCIENCE AND ECONOMICS

Tommaso Pessina*
Department of Economics, Management and Quantitative Methods
Department of Computer Science
University of Milan
Milan, Italy
tommaso.pessina@studenti.unimi.it

May 5, 2022

ABSTRACT

In these difficult days we know more than ever the importance and the weight of words, especially in the legal context.
This project is aimed to analyse a collection of court decision in order to better understand the importance of words in the legal context. We will focus on words frequency, relevance, correlation and historical trend.

Keywords Topic Model · Term Frequency · Trend · More

1 Introduction

First of all, we should spend few word about the dataset, the Illinois Bulk Dataset which is a collection of many court decision over the year.
Particularly, for our purpose, we are interested only in two fields:

- casebody: the actual body content of the court decision (CD)
- decision_date: date of resolution of the court decision

We decide to divide this research into three different fields:

- Terms frequency: we will analyse the more relevant term (also by subject of interest);
- Trend: here we will discuss, and try to correlate, the relevance of term with their historical trend;
- Topic model: we will run a topic model analysis to better understand this kind of topic, in order also to find any particular evidence regarding certain words.

In each step we will use three categories of words as subject of interest:

- narcotics: cannabis, cocaine, methamphetamine, smart drugs, marijuana, MDMA, LSD, KETAMINA, heroin, fentanyl;
- weapons: gun, knife, weapon, firearm, rifle, carabine, shotgun, assaults rifle, sword, blunt objects;
- investigation: gang, mafia, serial kiler, rape, thefts, recidivism, arrest, ethnicity, caucasian, afroamerican, native american, hispanic, gender, male, female, man, woman, girl, boy, robbery, cybercrime;

*MSc. Data Science and Economics student & Bsc. Computer Science & System Engineer @ Eltek Group

2 Data cleaning and preparation

Firstly, we read our dataset as a Json list by selecting only the casebody and the decision date. Then we apply a custom function that will clean each casebody.
Particularly, we convert everything to lower case, remove any special character, number and format symbol. Then we will remove all the stopwords (according to the English dictionary) and remove all the punctuation.
In the case of the decision date, we saw that older court decision has only the year or not more than year and month. So we decided to limit the information of the decision date only to the year.

3 Term frequency

This is the first analysis that we conduced on our dataset, because we wanted to discover the more relevant, but also more used, word and try to understand why. After understanding why, we may also ask ourself if there are also some more important factors.
The very first analysis that we may run, in this sense, is the TF-IDF function (short for term frequency–inverse document frequency).
We should start by saying that the TF-IDF's purpose is finding importance of words within a series of documents, but here we analyse all words as a unique big document. So, actually, we know that the TF-IDF is not the right tool here, but since we want to find the importance of words we think that, starting from the TF-IDF, we can arrange a formula for our problem.
As matter of fact instead of using the TF-IDF formula:

$$w_{ij} = tf_{ij} * log(\frac{N}{df_i})$$

In which:

- $tf_{i,j}$: term frequency of i in j;
- df_i: total document that contains i;
- N: overall number of documents.

Starting from here we came to the following formula:

$$w_i = tf_i * log(\frac{N}{tf_i})$$

Where:

- w_i is the score of the i-th word;
- tf_i is the frequency of the word i;
- N since in any case we should take a "reference", we downloaded the brown dataset and take its length as overall number of words;

Finally, if we select the 10 most common terms we will obtain:

Score	Word
427139.55	whether
427016.78	error
427009.97	illinois
426959.04	appeal
426918.08	motion
426793.69	could
426644.58	first
426496.84	counsel
426434.15	question
426334.80	appellant

Table 1: TF-IDF: Top 10 words

Score	Word	Delta
78790.93	cocaine	348348.62
38950.94	cannabis	388188.61
37253.07	marijuana	389886.48
30791.07	heroin	396348.47
83.20	fentanyl	427056.35
0.00	methamphetamine	427139.55
0.00	smart drugs	427139.55
0.00	MDMA	427139.55
0.00	LSD	427139.55
0.00	KETAMINA	427139.55

Table 2: TF-IDF: narcotics subject

Score	Word	Delta
159020.97	gun	268118.58
85831.46	weapon	341308.08
65316.64	knife	361822.90
42583.54	firearm	384556.01
32449.91	shotgun	394689.63
16244.95	rifle	410894.59
3479.15	sword	423660.39
0.00	carabine	427139.55
0.00	assaults rifle	427139.55
0.00	blunt objects	427139.55

Table 3: TF-IDF: weapon subject

Despite the fact that we may expect "Illinois" as frequent word since we are talking about the state of Illinois, seems that others are word often used in court.
By going further with our analysis we may search the word's score organised by our three subjects of interest. Moreover, we include for each word the delta from the score of the most important word.
In Table 2 you can see the result for the narcotics subject. From this table seems that "Cocaine" is the most important narcotic and also it outrank the other by more or less the double.
Instead, in Table 3 we can see the result for the weapon subject and we can notice that in the overall scores the word "gun" outrank the word "cocaine".

Finally, in Table 4 we can see the result of the investigation subject.

Moving on, there exist some other analysis that might be done, basing on the previous result. We may analyse a Word Embedding model in order to get any correlation between terms and see if there is some implication about frequent words.
We decided to use a pre-trained model because:

- train a model turns out be a time consuming job, as matter of fact after 48 hours of continuously training we decided to kill the process, also because of the following point;
- by using a pre-trained model we think that actually we can get some more interesting outcomes (since should not be biased).

So, we decide to use the Google's news vector model and separately analyse the frequent word and the three topic.
Let's discuss firstly about the frequent word group. As we can see from Figure 1, there are different groups of words and in particular:

- "Illinois" seems to be similar to other state's names. As matter of fact we should expect that this word is not relevant in the legal context;
- the central part regards more the legal context, as matter of fact for example the word "appeal" seems to be near to the word "motion" but also to "counsel" and "appellant". Moreover seems that "appellant" imply "Movant", "Defedants Motion" and more general something about defendats. If we pass to the word "Counsel" seems

Score	Word	Delta
203715.38	arrest	223424.17
198729.30	man	228410.25
166304.48	robbery	260835.06
73472.22	woman	353667.33
68102.24	rape	359037.31
57529.83	boy	369609.72
53142.12	gang	373997.43
47206.85	girl	379932.70
35444.56	male	391694.99
29467.96	female	397671.59
6004.89	hispanic	421134.66
5118.08	gender	422021.47
4807.74	thefts	422331.81
4129.28	recidivism	423010.27
3013.26	caucasian	424126.28
1092.12	mafia	426047.43
591.25	ethnicity	426548.30
0.00	serial kiler	427139.55
0.00	afroamerican	427139.55
0.00	native american	427139.55
0.00	cybercrime	427139.55

Table 4: TF-IDF: investigation subject

that lawyer and attorney are similar. Moreover seems that the words "motion" and "appeal" partially overlap, and actually motions are addressed by appeal.

- Another interesting fact is that the purple and light green groups, by going further away from the center, leads to more informal words.

But, as general outcomes, we say that in the legal context:

- formal word are preferred;
- motion imply an appeal;
- question seems to be little bit negatively correlated to the appellant
- the appellant is negatively correlated to word like "could" (e.g. should, can, could);

Finally we may discuss a little about our three topics, from Figure 2 As we can see:

- luckily words like child and adolescent are negatively correlated with drugs and guns in general;
- seems that identity thief might be correlated to mafia and/or abuser;
- abuser might imply teen pregnancy;

Generally speaking we cannot find any "very" relevant correlation form this graphical representation.

4 Terminological trend

In this section of our project we will analyze historic terminological trend, again by subject of interest. Firstly, and only as guideline, in Figure 3 we may analyse the complete distribution of the Illinois CD. We can seen that the major part of court decision was deliberated in the near past, note that our dataset start from 1828. Moreover, seems that we can consider a big peak any value bigger that 300, so we select only the year with more than 300 CD and the result are as shown in Table 5.

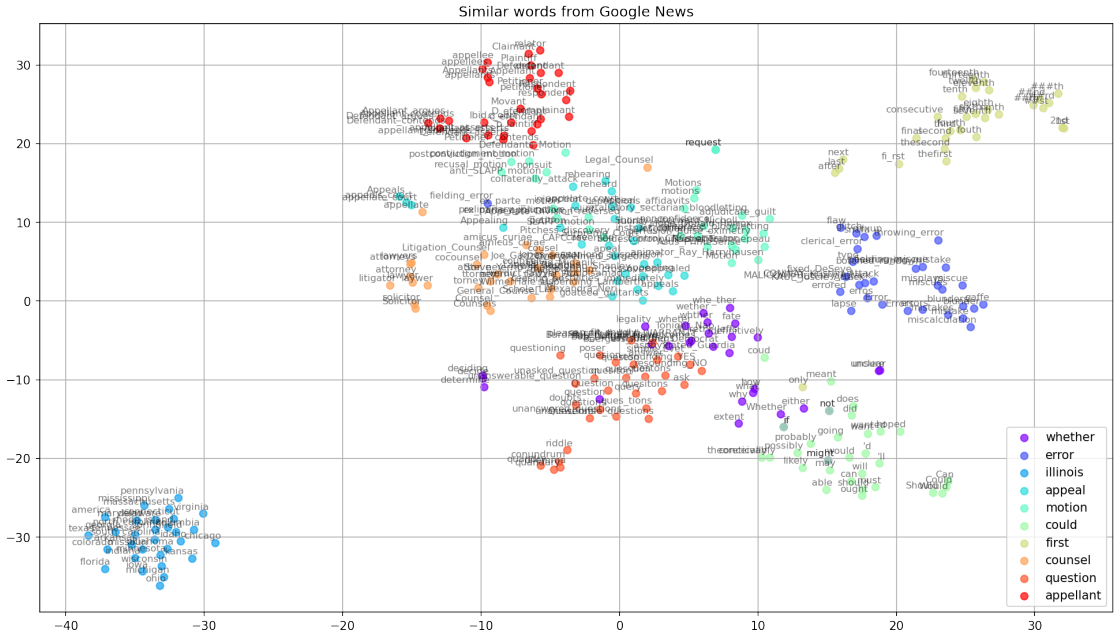

Figure 1: Similar frequent words

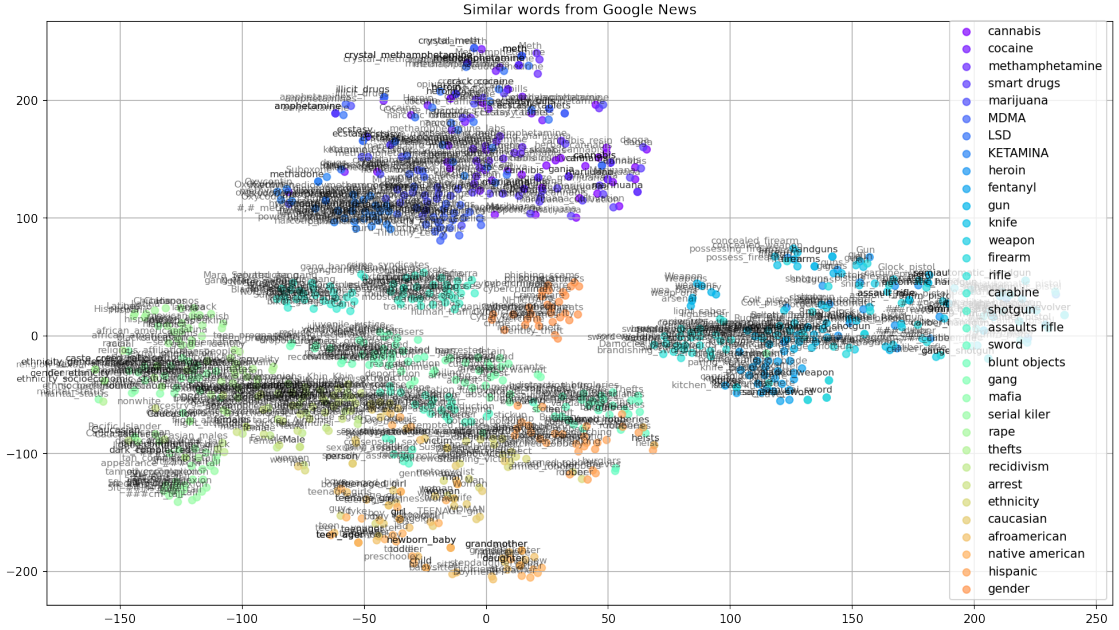

Figure 2: Similar frequent words according to our three sample topic.

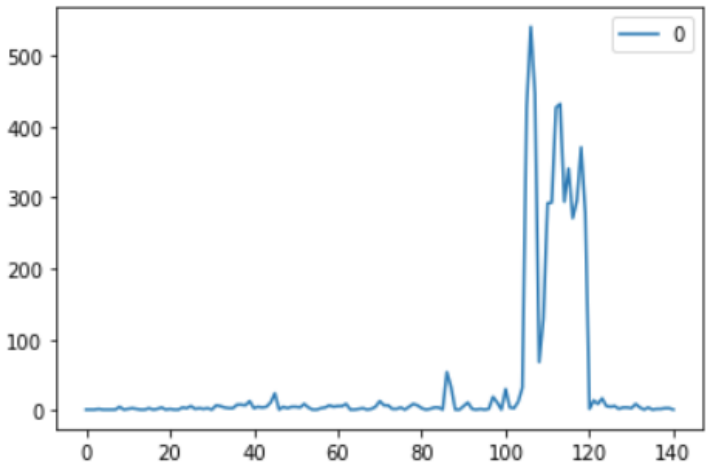

Figure 3: Complete historic terminological trend.

Year	Count
1973	427
1974	541
1975	446
1980	427
1981	432
1983	341
1986	371

Table 5: Year with bigger number of Court Decision

Now we became curios about why this years contains such a huge number of Court Decision and the answers are:

- In 1973 was found the Drug Enforcement Administrator, aka DEA;
- Turns out the in the years 1980-1985 there was a lot of activity about the DEA;
- The Federal Comprehensive Drug Abuse Prevention and Control Act of 1970, more commonly known as the Controlled Substances Act, became effective on May 1, 1971. The goal of the Controlled Substances Act is to improve the manufacturing, importation and exportation, distribution, and dispensing of controlled substances [1].

5 Topic model

Finally, we came to our last analysis of our project, that Topic Modelling. We choose to run two type of algorithm:

- Latent Semantic Indexing (LSI): implements fast truncated SVD (Singular Value Decomposition) and it's usually fast but not too much accurate;
- Latent Dirichlet Allocation (LDA): actually one of the most popular topic modelling method.

Firstly we create a dictionary representation of the documents (i.e. the collection of CD) and than we need to convert it to a BagOfWord format. Now we can choose some hyper-parameters, but we will discuss about this later on. Let's start from the LSI, since in the end it does not lead to a consistent result. In order to evaluate both methods we used the coherence method. In this case we use the U_MASS coherence method and we obtain -1.16028219230003 (with 30 topics) which is not too bad (the closer to zero, the better) and indeed it was fast. So, we will discuss about hyper-parameters and result only with the LDA method.
We started by tune our hyper-parameters by starting with a coherence, using the c_v method, of 0.32 but with end to 0.55 (although it could probably be improved further). For consistency purpose we calculate also the coherence using

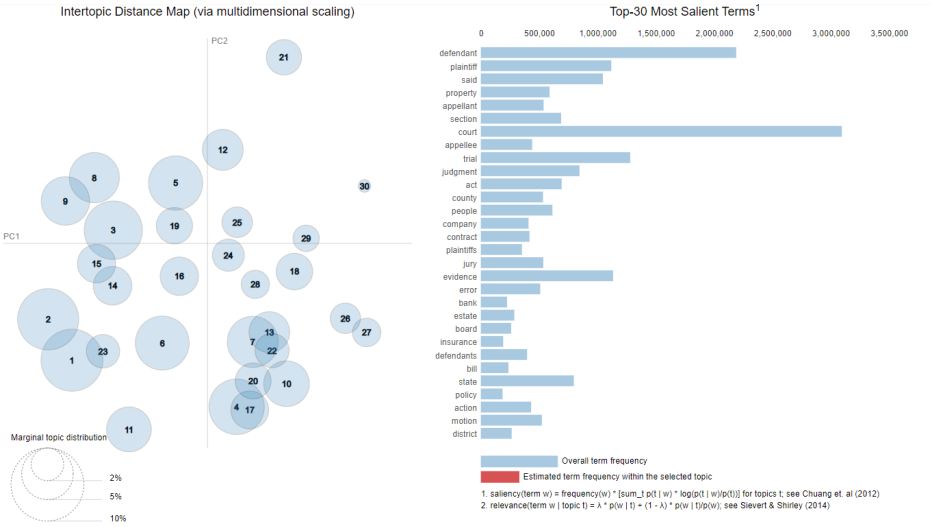

Figure 4: Topic distribution

the U_MASS method that give us -1.0193124596850156 which is slightly better than the previous method. In this final situation we choose the following configuration:

- 30 requested latent topics to be extracted from the training corpus;
- 2000 documents to be used in each training chunk;
- 35 passes through the corpus during training;
- 600 maximum iterations through the corpus when inferring the topic distribution of a corpus;
- we don't evaluate model perplexity because it takes too much time;
- we choose online learning.

Now, since 30 topic are a lot, we will analyse only the ones that we consider very relevant for our scope. As matter of fact:

- From topic 21 we may see the word "appellant" is in absolute contrast with Topic 1 and 2 which contains the words "defendant" and "evidence" (which, by the way, are correlated);
- The defendant, from Topic 9, is negatively correlated to everything that regard banks, as we can see from Topic 26 and 27;
- From Topic 1 and 2 we can see perhaps the most obvious thing: "evidence", "testimony" and "defendant" are positively correlated; But less obvious is the, again positive, correlation between "evidence", "testimony", "trial" and medical-related word (e.g. hospital, treatment, care, etc.) that can be seen from Topic 1-2 and 23;

Finally, from Figure 4 we can see the topic graphical distribution.

6 Conclusion

Finally we can sum up everything in order to understand the importance of word in the legal context.
The more relevant result are the ones coming out from the word embedding model and the topic model. As matter of fact we learn that:

- there exist some correlation between word that might imply the legal decision (like, for example, in the case of a trial for hospital-related words);
- the appellant does not pair well with questions;
- we may expect that a formal language is preferred;

- the appellant should be careful about evidence. As matter of fact, from Court of Appeal, BC we see: "in general, you cannot introduce new or additional evidence at your appeal. You must rely on the evidence that you submitted in the previous proceedings."[2].
- the defendant should pay attention about the financial word (e.g. banks)

This is only a little analysis, that of course can be improved and enlarged, but might prove that this kind of method may give us some relevant and useful results, regarding the importance of word in the legal context.

References

[1] Michael Gabay, PharmD, JD, BCPS (2013) *The Federal Controlled Substances Act: Schedules and Pharmacy Registration*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3839489/

[2] British Columbia Court of Appeal. Legally reviewed: 2020/ Apr *BC Court of Appeal: Guidebook for Appellants*, https://www.courtofappealbc.ca/appellant-guidebook/3.5-introducing-new-evidence