The importance of word in the legal context

PESSINA TOMMASO

961739

Dataset

- It is the Illinois Bulk Dataset, a collection of many court decision over the years.
- The two relevant filed are:
  - Casebody: content of the court decision
  - Decision date: date of resolution of the court decision
- For our analysis we define three subject of interest:
  - narcotics
  - weapons
  - investigation

Steps

- Data cleaning:
  - convert to lower case, remove special character, number and format symbol. Then remove all the stopwords and punctuation
- Analysis:
  - Term frequency
  - Word similarity
  - Terminological trend over the years
  - Topic model

What is the aim?

Find a method to exploit any correlation/correspondence between word in the legal context.

Term frequency

- We read everything as a unique big document
  - TF-IDF is not intended for this porpuse
  - We use a modified formula:

$$occurences\_word_{\_i} * \log\left(\frac{tot\_num\_words}{occurences\_word_{\_i}}\right)$$

tot_num_words according to the brown dataset

- To the right we can see the ten most frequent word

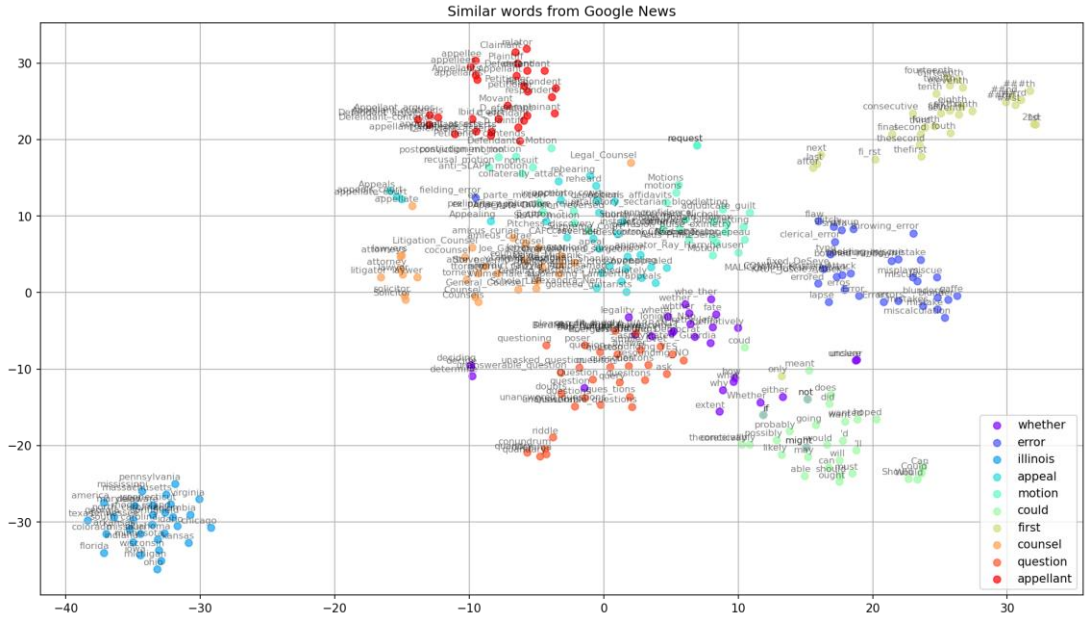
- We analyse also the score for our subjects of interest and:
  - «Cocaine» is the narcotic with highest score
  - «gun» is the weapon with highest score
  - «arrest» is the investigation-related word with highest score

- As expected, this analysis does not leads to useful result

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

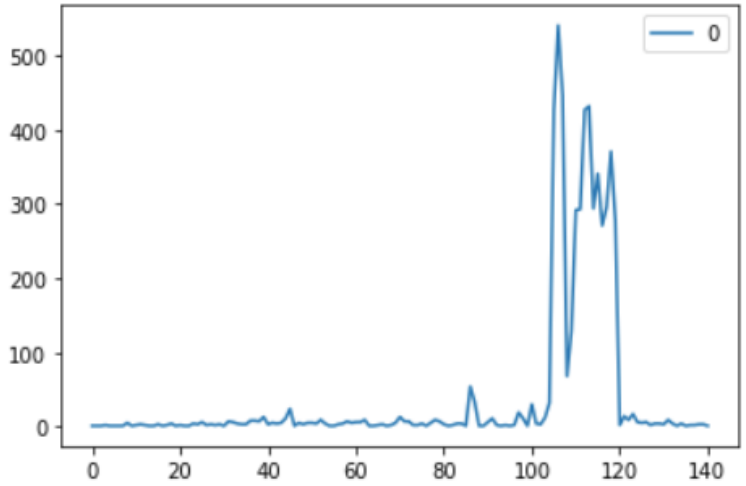
Word similarity

- We use the Google news pre-trained word embedding model
- Key results:
  - «Illinois» is similar to other states name
  - Similar word are:
    - «appeal» and «motion»
    - «counsel» and «appellant»
    - «appellant» imply «movant» and «Defedants Motion»
    - «lawyer» and «attorney»
    - «motion» and «appeal» (and motions are addressed by 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)
  - seems that identity thief might be correlated to mafia and/or abuser



Terminological trend

- In 1973 was found the Drug Enforcement Administrator, aka DEA;
- 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.. The goal of the Controlled Substances Act is to improve the manufacturing, importation and exportation, distribution, and dispensing of controlled substances (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3839489/)



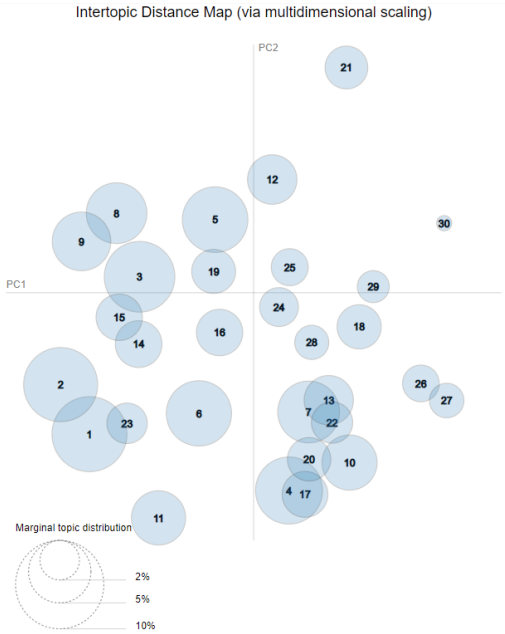
Year	Count
1973	427
1974	541
1975	446
1980	427
1981	432
1983	341
1986	371

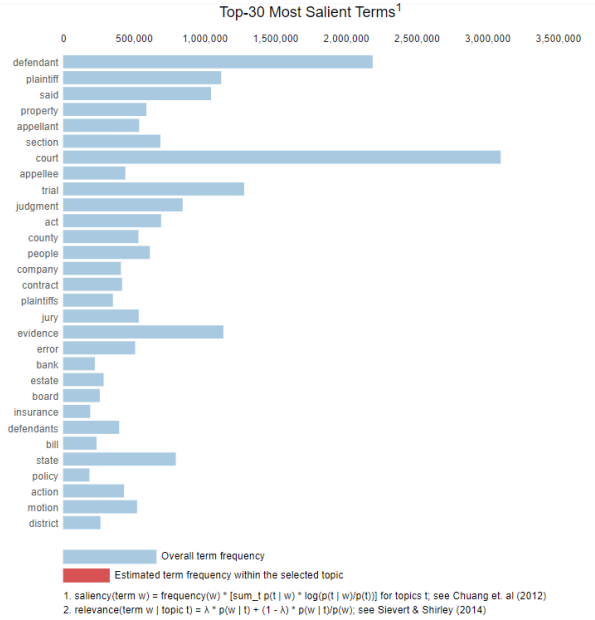
Topic model

- Methods:
  - Latent Semantic Indexing (LSI): U_MASS coherence method -1.16028219230003
  - Latent Dirichlet Allocation (LDA): U_MASS method give us -1.0193124596850156 ⇦
- Steps:
  - create a dictionary representation of the documents
  - convert it to a BagOfWord format.
  - tune hyperparameters:
    - • 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.

Topic model

- 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 Topic26 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





Conclusion

- 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.“ (https://www.courtofappealbc.ca/appellant-guidebook/3.5-introducing-new-evidence).
- the defendant should pay attention about the financial word (e.g., banks)