WORDS IN JUDGEMENT

Relevant Terms in Court Decisions about Narcotics

Martina Viggiano (954603)

MSc Data Science and Economics - Università degli Studi di Milano

Abstract. This project presents an analysis of textual data of United States case law dataset provided by Caselaw Access Project. The object is to study relevant terms within the text of court decisions, to measure frequency and correlation between words.

Keywords: Case-law analysis; Knowledge extraction; Information retrieval; Tf-Idf; Word2Vec; Pointwise Mutual Information.

1 Introduction

Between the subjects of interest we selected the lists of terms composed of various names of narcotics.

Schedule I	Schedule II	Schedule III	Schedule IV
– cannabis – marijuana – lsd – heroin – methaqualone – ecstasy – peyote – mescaline – mda – mdma	– cocaine – methamphetamine – hydromorphone – dilaudid – meperidine – demerol – oxycodone – dexedrine – fentanyl – ritalin – methadone – amphetamine – phencyclidine – pseudoephedrine – ephedrine – opium – dilaudid – preludin	– ketamine – anabolic – steroids – testosterone	– modafinil – provigil – adderall – methylphenidate – memantine – axura – soma – xanax – darvon – darvocet – valium – ativan – talwin – ambien – tramadol – ethclorvynol

Table 1: Lists of narcotics per schedule.

We created subgroups of narcotics, according to the classification found on DEA[1]

[1] Drug Enforcement Administration of USA.

website [2]. The differentiation was based on the drug's acceptable medical use and the drug's abuse or dependency potential. In particular, we obtained 4 schedules:

- Schedule I substances are defined as drugs with no currently accepted medical use and a high potential for abuse;
- Schedule II have a high potential for abuse, with use potentially leading to severe psychological or physical dependence, thus are also considered dangerous;
- Schedule III substances are defined as drugs with a moderate to low potential for physical and psychological dependence, which potential abuse is less than Schedule I and Schedule II drugs but more than Schedule IV.;
- Schedule IV drugs, substances, or chemicals are defined as drugs with a low potential for abuse and low risk of dependence.

We collected the substances of each schedule found inside the dataset.

1.1 Data Cleaning

Before moving to the exploratory part of the project, we needed to deeply clean textual data. In particular, we re-build the structure of the dataset, extracting judges' opinions, together with CDs' metadata: for each opinion, we attached related information, such as the id of the opinion, the author, the decision date and judicial opinion type.
After filtering by subject of interest, i.e. narcotics, we corrected typos and removed some expressions; then we lemmatized the words in order to obtain a less sparse term matrix in the followinf steps.
Then, we repeated the operations above on opinions split into sentences: thereby, we created a new set of data, that will be used for defining co-occurrences of terms.

2 Research question and methodology

In the following chapters we are going to address the research questions: the goal of the project is to retrieve specific terms within the text of each schedule of narcotics and to estimate the relevance and frequency of each word along the temporal dimension.
To accomplish this, we detected opinions containing at least one of the drugs belonging to only one of the schedules listed in Table 1. Since one of our goals was to compare specific terminology between different schedules, we erased opinions in which we found narcotics from two or more distinct classes - i.e. schedules - in order not to get intersections between sets, which may have biased our results. We labeled data in four classes, one for each schedule, and we merged opinions belonging to the same schedule, building four pseudo-documents.
Unfortunately, the numerousness of opinions forming the pseudo-documents were rather different: so we selected only two of them, Schedule I and Schedule II, which represented the largest sets - 1421 and 2287 opinions respectively - while the other two contained less than a dozen of opinions.

2.1 Frequent terms and Tf-Idf

After pre-processing data, we looked for most frequent terms computing relative, logarithmic and augmented term frequency.
Then, to study the differences between the two classes, we computed the Tf-Idf, which is a measure that combines term frequency and inverse document frequency, to take into account the specific relevance of a term with respect to a document, not only considering its frequency.

$$TfIdf(t,d) = tf_{t,d} idf_t = tf_{t,d} log \frac{N}{df_t}$$

Then, we worked on terminological trends along temporal dimension by partitioning by decades the opinions - from 1950 to 2000 - in order to investigate possible lexicon variations between years.

2.2 Word2Vec - Cosine similarity

To meet *Step 2* of *research questions* we decided to also employ another method in order to find the most relevant terms by subject of interest: to this end *gensim* word embedding model Word2Vec was enforced.
Through this word embedding model, we represented words as vectors in the multidimensional space defined by the other words, to extract their meaning by words' contexts: we start from the idea that the context can be seen as the neighbourhood of the single word, so the meaning of a word is determined by the terms that come with it.
Based on the assumption that in each of the opinions that we collected we have at least one of the narcotics listed in Table 1 columns, we can reasonably say that those drugs may represent the "topics" of our textual data, since they can summarize the subject addressed in the sentences.
Thus, if we seek terms that are close to those words - i.e. "cannabis" , "lsd", "cocaine" etcetera - we would find the "relevant terminology" we are looking for.
In Word2Vec, similar words have similar word embeddings; this means that they are close to each other in terms of cosine distance.
There are two main algorithms to obtain a Word2Vec implementation: Continuous Bag of Words (CBOW) and Skip-Gram. In this case we implemented the default one - the CBOW - which is based on the process that, starting from the context words, predicts the main word. The input layer of the neural network consists of the context words in one-hot encoding form with size 1*V.[2] For every context word, we get the hidden layer resulting from the weight matrix W_{VxE}.
Then we average them into a single hidden layer, which is passed onto the output layer.
Once the training is completed, the weight matrix W_{VxE} is used to generate the word embeddings from the one-hot encodings. [8]
In our case, we trained *gensim* word embedding model on data taken from the

[2] Where V is the size of the vocabulary.

two pseudo-documents we built previously, setting the window size equal to 5: we obtained two models trained on different sets of data. The objective was to find potential differences between them.
After training one model for each of the two schedules - Schedule I and Schedule II - we looked for the most similar words to each element in the list of narcotics - i.e. for Schedule I model we looked for terms with highest similarity to "cannabis", "marijuana", "lsd" and so on.
To achieve it, we computed the *gensim Word2Vec* function for cosine similarity, which measures the distance between two vectors by considering the cosine of the angle between two word vectors projected in a multi-dimensional space.
We collected the top 10 *Word2Vec most_similar* words for the 5 most frequent narcotics[3] of each schedule.
After that, we repeated the process by looking for most similar terms to common words: instead of using specific classes of narcotics, which are not shared by the two schedules, we set a list of words that are present in both our subsets and which are typical in narcotics field: "drug", "narcotic", "substance", "crime", "addiction".
Then, to meet *Step 3* of *research questions* we split opinions of both schedules by decades - from 1950 to 2000 - and we replicated the previous steps for each of the periods: the objective was to look for differences in lexicon between decades.

2.3 Pointwise Mutual Information

The next research question was about finding a measure of correlation between wordsboth single words and group of words.
To answer this question we deployed the computation of positive Pointwise Mutual Information (PMI).
PMI evaluates the relation between the words joint probability and their marginal probability: it is a statistical measurement that permits to calculate the probability of observing word i and word j together with the probabilities of observing words i and j independently.[6]

$$[e_{ij}] = PMI(w_i, w_j) = log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} =$$
$$log \left[\left(\sum_{w'_i, w'_j \in D} count(w'_i, w'_j) \right) \frac{count(w_i, w_j)}{count(w_i)count(w_j)} \right]$$

3 Experimental results

3.1 TfIdf results

In each case, the top most frequent words are very common terms for legal field texts. Some examples are: "defendant", "trial", "evidence", "testify", "officer", "police" and so on.

[3] Cannabis, marijuana, lsd, heroin and mdma for Schedule I; cocaine, methamphetamine, amphetamine, pseudoephedrine and phencyclidine for Schedule II.

Also in this case the most relevant terms belong to specific legal vocabulary and they are common to both schedules. Thus, it was not possible to find significant differences between the two sets of relevant terminology.

Similarly to what we have experienced before, we did not notice any pertinent difference between top relevant terms for each set of years.

3.2 Word embedding results

cannabis		**marijuana**		**lsd**		**heroin**		**mdma**	
sativa	0.58	pot	0.57	diacetyl	0.64	narcotic	0.73	homogeneous	0.64
heroin	0.57	cannabis	0.56	gram	0.62	drug	0.64	tablet	0.60
marijuana	0.56	junk	0.55	barbituric	0.61	cannabis	0.57	twig	0.59
gram	0.55	reefer	0.54	tablet	0.61	tinfoil	0.56	vile	0.58
substance	0.53	underage	0.54	controlled	0.60	dope	0.54	pentazocine	0.57
contraband	0.53	cider	0.54	mda	0.60	marijuana	0.53	resin	0.57
drug	0.53	heroin	0.53	derivative	0.60	gram	0.50	diacetyl	0.55
plant	0.51	dope	0.52	twig	0.59	pill	0.50	grain	0.55
hashish	0.50	cigarette	0.49	pentazocine	0.59	tablet	0.50	lsd	0.54
manufacture	0.48	whiskey	0.49	indole	0.59	junk	0.50	medicinal	0.54

Table 2: Lists of top 10 similar terms for Schedule I.

cocaine		**methamph.**[4]		**amphetamine**		**pseudoephedrine**		**phencyclidine**	
drug	0.72	pseudoephedrine	0.70	caffeine	0.72	ephedrine	0.72	mixture	0.71
narcotic	0.64	phencyclidine	0.62	pill	0.68	methamph.	0.70	salt	0.68
sixteenth	0.56	manufacture	0.58	capsule	0.67	capsule	0.70	acid	0.68
gram	0.54	ephedrine	0.54	tablet	0.66	filter	0.67	ephedrine	0.66
prepackage	0.53	manufacturing	0.54	barbiturate	0.66	pill	0.65	codeine	0.66
amphetamine	0.53	cocaine	0.51	dilaudid	0.65	barbituric	0.65	barbituric	0.66
powder	0.52	simulation	0.50	phencyclidine	0.64	tablet	0.65	caffeine	0.65
kilogram	0.51	mixture	0.50	milligram	0.63	milligram	0.64	amphetamine	0.64
methamph.	0.51	inhaler	0.48	acid	0.61	salt	0.64	hydrochloric	0.64
substance	0.51	narcotic	0.48	nonnarcotic	0.61	mixture	0.63	isomer	0.63

Table 3: Lists of top 10 similar terms for Schedule II.

3.3 PMI results

??

drug		**narcotic**		**substance**		**addiction**	
I	II	I	II	I	II	I	II
narcotic	narcotic	heroin	drug	lsd	phencyclidine	addict	craving
heroin	cocaine	drug	cocaine	tablet	cocaine	habit	drunkenness
dope	meperdine	dope	contraband	cannabis	gram	usage	addict
junk	clandestinely	buyer	clandestinely	gram	narcotic	pusher	abuser
cannabis	dope	junk	substance	heroin	profiteer	dependency	overdose
organizer	amphetamine	marijuana	meperdine	capsule	chemical	ingest	usage
transportation	profit	contraband	methamph.	quantity	amphetamine	user	alcoholism
petty	substance	output	seller	pill	drug	chronic	anticonvulsive
embalming	consummate	pusher	amphetamine	manufacturing	powder	alcoholism	relapse
dispense	methamph.	illicit	ongoing	material	classifie	overdose	bronchodilator

Table 4: Lists of top 10 similar terms for shared words.

4 Concluding remarks

??

References

1. Chang, Felix and McCabe, Erin and Lee, James: Mining the Harvard Caselaw Access Project (September 29, 2020). Available at SSRN: https://ssrn.com/abstract=3529257.
2. Drug Scheduling on US Drug Enforcement Administration https://www.dea.gov/drug-information/drug-scheduling.
3. Shravan Kuchkula: Document Similarity (2019). https://shravan-kuchkula.github.io/nlp/document-similarity/.
4. Mattia Falduti: Law and Data Science: Knowledge Modeling and Extraction from Court Decisions (a.y. 2019/2020). https://air.unimi.it/retrieve/handle/2434/799875/1659203/phd_unimi_R11975.pdf.
5. Kavita Ganesan: Extracting Keywords with TF-IDF and Python's Scikit-Learn. https://kavita-ganesan.com/extracting-keywords-from-text-tfidf/.YSfMJY4zZPY.
6. Valentina Alto: Understanding Pointwise Mutual Information in NLP. 2020. https://medium.com/dataseries/understanding-pointwise-mutual-information-in-nlp-e4ef75ecb57a.
7. Chris McCormick: Word2Vec Tutorial - The Skip-Gram Model. 2016. http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
8. Martin Riva: Word Embeddings: CBOW vs Skip-Gram. 2021 https://www.baeldung.com/cs/word-embeddings-cbow-vs-skip-gram
9. Ria Kulshrestha: NLP 101: Word2Vec — Skip-gram and CBOW. 2019. https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314