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 various names of narcotics. Starting from that, we created four subgroups of narcotics, according to the classification found on DEA[1] website [2]. The differentiation was based on the drug's acceptable medical use and the drug's abuse or dependency potential.

Schedule I	Schedule II	Schedule III	Schedule IV
– cannabis – marijuana – lsd – heroin – methaqualone – ecstasy – peyote – mescaline – mda – mdma	– cocaine – methamph. – hydromorphone – dilaudid – meperidine – demerol – oxycodone – dexedrine – fentanyl – ritalin – methadone – amphetamine – phencyclidine – pseudoephedrine – ephedrine – opium – 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.

In particular, we obtained 4 schedules:

[1] Drug Enforcement Administration of USA.

- Schedule I substances are defined as drugs with no currently accepted medical use and a high potential for abuse;
- Schedule II drugs have a high potential for abuse, with use potentially leading to severe psychological or physical dependence;
- 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.[2]

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 following 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 main 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] In this paper "methamphetamine" is shortened to "methamph.".

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 find relevant terminology and then study the differences between the two classes, we computed the Tf-Idf: it 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 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 decades.

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.[3] 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

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

two pseudo-documents we built previously, setting the window size equal to 5. By distinguishing between the two schedules, we wanted to study if there were any terminological difference between the two sets and find potential distinctive vocabulary related to types of drugs.
After training the model, 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[4] 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", "addiction".
Then, to meet *Step 3* of *research questions* we split the original opinions set 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 words, both 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]
To do so, we built the co-occurrence matrix of single words and bi-grams and then we computed the value of PMI for each couple of elements - respectively single words or two words - and we put the results inside a matrix.

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

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

3 Experimental results

3.1 Tf-Idf results

Studying the distinction between most frequent terms in the two schedules, the top terms are very common words for legal field texts. Some examples are: "defendant", "trial", "evidence", "testify", "officer", "police" and so on.
The results did not change when we computed the Tf-Idf: 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.
As regards time trends, similarly to what we have experienced before, we did

	Schedule I		**Schedule II**	
	word	value	word	value
0	defendant	0.712	defendant	0.726
1	trial	0.265	trial	0.281
2	evidence	0.182	evidence	0.187
3	testify	0.128	testify	0.126
4	officer	0.127	find	0.119
5	find	0.122	police	0.102
6	police	0.107	officer	0.102
7	testimony	0.099	cocaine	0.098
8	search	0.095	jury	0.091
9	arrest	0.087	state	0.091

Table 2: Tf-Idf top 10 terms for schedules.

not notice any pertinent difference between top relevant terms for each set of years.
In Table 3 we displayed the top 10 most relevant terms for four of the decades we built: we almost have the same terms in all of them, sometimes in a different position in terms of ordering.

	50s		**60s**		**70s**		**80s**	
	word	value	word	value	word	value	word	value
0	defendant	0.578	defendant	0.654	defendant	0.692	defendant	0.723
1	evidence	0.242	narcotic	0.258	trial	0.256	trial	0.271
2	narcotic	0.196	trial	0.184	evidence	0.175	evidence	0.182
3	officer	0.185	officer	0.179	officer	0.133	testify	0.123
4	error	0.160	evidence	0.176	testify	0.121	find	0.113
5	testify	0.147	testify	0.170	search	0.120	police	0.105
6	trial	0.142	testimony	0.140	find	0.113	testimony	0.099
7	question	0.127	arrest	0.136	testimony	0.111	officer	0.099
8	record	0.120	witness	0.120	sentence	0.108	state	0.092
9	arrest	0.120	police	0.118	police	0.104	jury	0.090

Table 3: Tf-Idf top 10 terms for decades.

Thus, to have a better idea of the measure of relevance of a term in a specific decade or in a specific schedule, we defined a function named *find_tfidf()* which retrieves the Tf-Idf value and the ordering position of a given word in a given context.

3.2 Word embedding results

As we said in the previous section, initially we collected the top 10 most similar words for the 5 most frequent narcotics of each schedule.

cannabis		**marijuana**		**lsd**		**heroin**		**mdma**	
marijuana	0.57	cannabis	0.57	gram	0.63	narcotic	0.75	tablet	0.60
heroin	0.56	pot	0.54	tablet	0.62	drug	0.63	homogeneous	0.58
sativa	0.56	reefer	0.53	diacetyl	0.62	cannabis	0.56	lsd	0.55
substance	0.54	underage	0.53	barbituric	0.61	tinfoil	0.55	placemat	0.55
manufacture	0.53	liquor	0.50	homogeneous	0.61	capsule	0.53	resin	0.54
contraband	0.53	telegram	0.49	mda	0.61	dope	0.53	randomly	0.53
gram	0.53	junk	0.49	controlled	0.59	gram	0.52	crumble	0.52
plant	0.50	whiskey	0.49	powdered	0.59	pill	0.50	grain	0.51
drug	0.49	dope	0.48	derivative	0.57	powder	0.49	crushed	0.51
methaqualone	0.46	cider	0.48	indole	0.57	tablet	0.48	medicinal	0.51

Table 4: Lists of top 10 similar terms for Schedule I.

cocaine		**methamph.**		**amphetamine**		**pseudoephedrine**		**phencyclidine**	
drug	0.72	pseudoephedrine	0.70	pill	0.70	ephedrine	0.73	mixture	0.69
narcotic	0.69	phencyclidine	0.56	caffeine	0.70	methamph.	0.70	salt	0.68
amphetamine	0.55	manufacture	0.56	tablet	0.69	capsule	0.68	hydrochloric	0.67
powder	0.55	ephedrine	0.52	phencyclidine	0.65	tablet	0.66	acid	0.66
gram	0.54	manufacturing	0.52	capsule	0.65	pill	0.64	hydrochloride	0.66
sixteenth	0.54	mixture	0.51	barbiturate	0.64	filter	0.63	amphetamine	0.65
kilogram	0.52	cocaine	0.51	milligram	0.64	milligram	0.63	isomer	0.64
coke	0.51	simulation	0.50	hydrochloride	0.64	mixture	0.62	caffeine	0.64
methamph.	0.51	pill	0.49	morphine	0.63	salt	0.62	barbituric	0.64
substance	0.50	ingredient	0.49	grain	0.61	precursor	0.61	opium	0.63

Table 5: Lists of top 10 similar terms for Schedule II.

From table 4 and 5 we can observe that for each drug the most close terms are not identical, but we can actually find some patterns. In particular, it is possible to see that in each of the lists we can notice the following elements:

- Other drugs: "methaqualone" , "pseudoephedine".
- Synonyms/ abbreviations/ slang referring to the same drug: "coke", "pot"
- Chemical elements and ingredients being parts of drugs: "hydrocloric", "salt"
- Form of the drug: "plant", "pill".
- Quantity of the drug: "gram", "kilogram".

Even if we cannot find a significant pattern distinguishing the terminology used in the two schedules, we can consider the narcotics singularly. In particular, we can notice some interesting results in lists of terms of "marijuana", "cocaine" and "amphetamine".
Within the terms with highest cosine similarity to "marijuana" we found some alcoholic beverages, namely "liquor", "whiskey" and "cider". Manually exploring the original CDs' opinions, we found out that defendants who have been detained for marijuana possession/use/trade were usually also found under the influence of alcohol or illegally possessing it.
As regards "cocaine" we can see several references to quantity, such as "gram" and "kilogram". Furthermore, also "sixteenth" is linked to this theme, since it refers to the "one-sixteenth of ounce", which is one of "typical" units of measurement of cocaine.
Then, "caffeine" refers to legally tradable tablets/pills of caffeine, that in our case have been used to transport and sell other controlled substances, such as amphetamine. As a matter of fact, by reading our textual data, we got to know that this products - caffeine pills - sometimes hid a few grams of narcotics, or even were entirely composed of illegal substances.

drug		**narcotic**		**substance**		**addiction**	
I	II	I	II	I	II	I	II
heroin	narcotic	heroin	drug	tablet	phencyclidine	addict	drunkenness
narcotic	cocaine	drug	cocaine	cannabis	gram	dependency	addict
dope	buying	buyer	contraband	lsd	cocaine	habit	craving
junk	amphetamine	dope	substance	gram	narcotic	usage	usage
cannabis	occasional	contraband	occasional	manufacturing	profiteer	user	abuser
profiteer	consummate	illicit	amphetamine	barbituric	amphetamine	pusher	alcoholism
shoplift	ongoing	cannabis	middleman	pill	purveyor	abuser	relapse
marijuana	dope	consummate	paraphernalia	pentazocine	drug	overdose	abstain
abuser	profit	junk	methamph.	chemical	powder	withdrawal	overdose
codeine	substance	pusher	ongoing	capsule	isomer	alcoholism	dependence

Table 6: Lists of top 10 similar terms for shared words.

After that, we studied the cosine similarity considering a new point of view: we looked for most similar terms to words that are typical of drug-related texts, namely "drug", "narcotic", "substance", "addiction".
As we can see from Table 6, we found a terminology close to the previous step - looking for words similar to the lists of narcotics. Also in this case we have specific names of drugs, mentioning also form and quantity. But unlike what we saw before, we have references to the trade and consumption - and abuse - of it. Moreover, within terms close to "addiction" we found both for Schedule I and Schedule II words alcohol-related.

3.3 PMI results

The last research question was about measuring the correlation between words and couple of words, by computing positive Pointwise Mutual Information.
We built a word-word - or bigram-bigram - co-occurrence matrix for each of the 2 schedules and one for the whole *opinion* dataset containing narcotic terms split into sentences. Then we computed the PMI value for each cell of the matrices and we defined several functions that can be used to explore the matrices.
By providing *get_co_pmi_value()* two words, it prints the corresponding value of PMI or co-occurrence. While with *max_cooccurrences()* function is possible to find the element - word or bigram - which has the highest value of PMI or co-occurrence in the matrix, with respect to the word or bigram provided.
Lastly, we defined the function *find_word_sorted_cooc_pmi()* which, staring from a given word or bigram, it returns the elements in the matrix for which the crossing cell is different from zero - i.e. we have co-occurrence in the sentences and/or we have positive PMI measure.
We looked for words correlated to the list of common terms we used previously - "drug", "narcotic", "substance", "addiction". Unfortunately we did not find any remarkable result, we only found generic words like "officer", "defense" and "motion" which cannot provide us any evidence of possible interesting association.
On the other hand, when we draw attention to the three bigram matrices, we came across some interesting results.
We got to know that "death sentence" and "death penalty" are correlated with sexual crimes both for the whole dataset[5] and Schedule II data. On the other hand, Schedule I seemed not to have a specific topic that can be linked to death penalty.

4 Concluding remarks

The results we displayed in this paper are only an overview of the whole project: on notebook *04_Results* in GitHub repository, you can find the cleaned datasets, the trained models and the functions that we used to retrieve the outcomes we showed previously.
In particular, it is possible to employ the functions to look for some insights, cosine distances and correlations. In this case, having a strong knowledge on juridical area would be useful in terms of basic notions: these are essential especially for choosing a starting list of terms, which would represent the point from which you can detect relevant terminology in the field.

[5] The dataset composed of all the opinions in which we found at least one of the narcotics listed on Table 1.

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