Legal Issue Spotting - First Phase Legal Analysis Recommendations

John Phillips
Stanford University
jophilli@stanford.edu

Abstract – Legal analysis is a multistep process that performs the complex tasks of disambiguation of comported real world activities through the lens of a predefined legal code and sets of laws. These activities often may instantiate multiple separate as well as overlapping legal stipulations and criteria thus creating a potential for contextually significant interpretations of how a given code or law applies to specified events. Thus when attempting to achieve a legal resolution given ambiguous circumstances or in a competitively incentivized resolution environment, the employment of large amounts of resources is not unheard of. Despite this market incentive legal analysis has yet to be materially aided by the employment algorithmic reasoning tools, either by NLP or syntactic learning algorithms (though search and visualization tools have proved useful as research aids).

We comparatively employ Naïve Bayes and SVM models in a parallel, non-mutually exclusive alignment towards the generic first phase of legal analysis: legal issue spotting. We do this by training separate supervised learning models against the digested legal opinions of the U.S. Federal Reporter, 3rd Series (F.3d) towards the U.S. Legal Code laws and sub laws to produce testable models capable of ingesting free form scenario descriptions and producing legal code research area recommendations.

Keywords: U.S. Legal Code, law, legal analysis, supersized learning, support vector machine, naïve bayes

INTRODUCTION

Despite efforts in recent years to advance Legal Analysis through various techniques, automated “issue spotting” remains problematic throughout the legal profession. The complex and overlapping implications of legal understandings and a syntactically nuanced legal code create massive challenges for NLP and syntactic learning mechanisms. Moreover the writing style of different legal professionals (and of clients) present real challenges towards the scaling of any successful model. Despite these challenges, the ambition of advancing legal counsel through automated issue spotting and recommended analysis remains preeminent.

The objective of this project is to create a model capable of ingesting text based scenario descriptions and predicting which areas of the Code of Laws of the United States of America (U.S. Code), (U.S.C.).

We should note that by “issue spotting” we refer to the first step in legal research of identifying possibly legal areas which may have been violated and not the specific logistical aspect of legal research of looking up legal documents - where there is a small number of commercialized catalogues of the U.S. Legal Code which aid researchers once legal areas have been identified for research and where a good number of technological advancements have already occurred.

While our initial orientation towards this problem indicated conducting an NLP treatment of the actual U.S. Code itself before applying a machine-learning algorithm against the digested NLP treatment, our investigation of prior research along these lines dissuaded us from continuing down this line of inquiry.

Instead we sought an entirely different dataset which might be distilled into a supervised learning training set(s) and test set(s) and which might help simplify our pre-training data processing. To this end we believe we now have such a dataset, though perhaps more bulky than one would hope.

EXPERIMENTAL DESIGN

A. Background

Prior efforts towards the creation of an algorithm capable of identification of tautological statements within a given text have proved problematic. Nuances of meaning and word associations have prevented unambiguous procedural machine learning efforts to understand a given issue provided its written description. Moreover, when text is compartmented and restricted to registered input values, these associations have proved in some cases overly narrow and thereby restrictive in developing into insight beyond the alternative of direct human observation.

From the standpoint of developing a fully formed legal analysis tool, these results at best could be characterized as having achieved an incomplete mosaic of legal analysis.

B. Dataset

Our effort therefore is to find a dataset which highlight a significant number of complex legal topics without losing the contextual flavoring of associated descriptions. Finding such a dataset we aim to bypass syntactic and referential complexities associated with current impasses in NLP research of the same endstate pursuit.

Towards this end we have scraped the first 100 volumes of opinions of the U.S. Federal Reporter, 3rd Series (F.3d), with the intention of ultimately scraping all 491 volumes – time permitting. In our initial scraping we find 84220 opinions enumerated by the website, with each opinion listing references of anywhere from zero to a handful of U.S.C sections and subsections. The first hundred volumes comprise roughly 756 MB of data after being scraped. The data is stored mostly in plain text, within in a csv file format.
A brief description of the original data may be instructive. Each opinion is written to explain the given ruling of the case. Paragraphs explain the pertinent facts of a given specific set of facts that relate to the case and in some instances these paragraphs are followed by a quick sentence indicating a law or sublaw within the U.S.C. which may be instructive for the reader of the opinion to review. This structure is the heart of constructing our training examples.

By viewing an entire opinion as an X value example we then may view these short references to U.S. Legal Code sections as Y destination mappings.

![Diagram of Opinion to Legal Code law mapping]

**Fig. 1: Opinion to Legal Code law mapping**

An example here is illustrative:

The jury found the defendants guilty of conspiracy to distribute and to possess with intent to distribute cocaine and heroin in violation of 21 U.S.C. Sec. 846 (1988) and possession with intent to distribute and distribution of a controlled substance in violation of 21 U.S.C. Sec. 841(a)(1) (1988). In addition, Thornton and Jones were convicted of participating in a continuing criminal enterprise in violation of 21 U.S.C. Sec. 848 (1988 & Supp. III 1991), and Fields was convicted of using a firearm during a drug trafficking offense in violation of 18 U.S.C. Sec. 924(c)(1) (1988 & Supp. III 1991), and possession of a firearm after having been previously convicted of a felony in violation of 18 U.S.C. Sec. 922(g)(1) (1988). All three defendants were sentenced under the United States Sentencing Guidelines to life imprisonment, and Thornton and Jones were each ordered to forfeit $6,230,000 to the government pursuant to 21 U.S.C. Sec. 853 (1988). The defendants have not challenged the propriety of their sentences or fines. Nor, significantly, have they alleged that the evidence was insufficient to support the verdicts. 7

Thus, we can observe that in only one paragraph of one opinion in one volume of opinions we find no less than six distinct laws under the U.S.C., which our hope is to train our model to correctly predict. Moreover the values of these references do not appear to comport easily to regular expression removal techniques.

Nevertheless, by using these individual references from the original opinions and setting each of these ‘target’ references to be a ‘Y’ destination for each ‘X’ training example of an opinion we fashion numerous positive examples of each law or sub law referenced in the original opinion.

Moreover by applying this method to a large number of opinions, though we may only get a handful of positive examples of a mapping instance to a given law, so long as we take care to ensure that do not submit a training example as both a positive and negative example for a given law, then we simultaneously generate numerous negative examples out of each original instance. We are afforded this benefit through submitting each of the examples in parallel for each of the binary decision models for each specific law, which again is dependent on the non-mutually exclusive nature of our recommendation architecture.

So be using the opinions instead of the U.S.C itself we abstract away from dealing with the NLP issues around the language of various laws and how they interact with given circumstances and rely instead on the word choice of numerous justices as they write their opinions.

C. Concerns

However in doing this a risk arises that we will acknowledge here and attempt to deal with later in our conclusion that the sum total of opinions –even at its fullest articulation may miss some areas of the U.S. Legal Code to which a user might be concerned.

We know that the U.S.C. is itself rather large and complex: In 2013 the U.S. House Judiciary Committee asked the Congressional Research Service to provide a calculation of the total number of criminal offenses contained in the U.S.C. The CRS responded indicating that they lacked the manpower and resources to provide an update to a number from 2008 of 4,500 total crimes, however the Judiciary Committee Chairman characterized the as growing “at a rapid rate of 500 a decade”.

So if we allow that the total number of criminal offenses may only make up a subset of U.S.C sections which an ideal legal analysis function may map to, it is reasonable to be concerned that the overall number of opinions may only generate a subset of training events per U.S.C section.

While we will attempt to address this concern later, it may be that this weakness of the dataset must simply be endured in its children-models.

**MODEL DESCRIPTION**

With this updated dataset the structure of the binary nature of the mapping and similarity to SPAM filter type problems becomes immediately apparent, and this channels our research towards two clear implementation algorithms: Support Vector Machines (SVMs) and Naïve Bayes ‘bucket-of-words’. Additionally, if those algorithms prove easily attainable then we also can follow SPAM filter research towards more complex models using n-grams, etc.

However a simple review of the dataset shows that we first must deal with a lingering dataset concern for both methods, as we first need to remove the ‘Y’ value(s) from the text within the training data.

Thus we worry that a both Naïve Bayes and SVM may identify the actual references of specific legal codes (‘846’, ‘841’, ‘848’, ‘924’, ‘922’, and ‘853’ in our earlier example) and attach a higher value on those raw numbers as the more regularly indicate their broader ‘laws’ than a randomly
occurring number of the same text. Moreover this problem metastasizes when we consider non-U.S. Legal Code references that may occur on a regular basis such as individual state laws, or the titles of specific court cases. Both of these appear to occur frequently enough to ‘corrupt’ our values within our model for those individual words and abbreviation terms. Thus, if removal proves overly complex we might have to simply live with this weakness in our model until a more pristine dataset can be developed.

We set each of our ‘Y’ value laws as separate target filters and allow the collection of the ‘Y’ values to be non-mutually exclusive. We do this with an eye towards our end user as a legal professional and the context of legal research as the user will likely be interested in all types of open legal questions, not simply the highest likelihood given the additive nature of legal complexities.

With these destination ‘Y’ values in place, we can now use the entire training set as both positive and negative training instances depending on the nature of the ‘Y’ mapping for the individual training event and which specific law we are currently training on. And indeed, the complete list of laws would –in theory- be the entire U.S. Legal Code but will -in practice- be the full list of laws referenced in the opinions contained in the full training set.

For training and testing we can see that this amount of data lends itself to a k-fold validation. Though, given the inevitably low number of positive training instances for some referenced laws, it may make sense to treat those specific low number law events as special cases and ensure they are contained by the training set vice the test set.

INITIAL FINDINGS

We have started the initial processing of this dataset and plan on employing a simple SVM and Naïve Bayes implementations which we have from the second homework problem set. A difference will be the parallel and non-mutually exclusive structure which seems a bit more exotic, though hopefully is as intuitive as the models we saw in class.

We have attempted to select an implementation that – assuming the dataset can be properly domesticated- will not prove too unwieldy, this given our recent precipitous decline in the number of project partners to aid with implementation.

CONCLUSION

Concluding remarks will depend on the full implementation, though we feel the discovery and structuring of our dataset to be a vast improvement over the initially proposed line of inquiry of attempting an NLP preprocessing of the U.S. Code. Moreover the parallel structure of our non-mutually exclusive binary decision model seems like an intuitive choice given the additive nature of complex legal analysis.

DIFFICULTIES AND FUTURE WORK

An observer may detect an implicit assumption in the structure of our model as we assume that the considered and clearly articulated legal opinions of appellate court judges as good training data and not be materially different from the ingested description of events from an unknown source which we will use at test time. In practice this difference may vary considerably in professionalism, writing ability, and tone thus producing a model bias that is not inherent in the models’ SPAM filter cousins which both train on and are applied to equivalent subject emails. More research will be needed to discover if this difference in source type proves vital. Nevertheless the use of synonym libraries may reduce this bias should it be discovered.

Separately the use of n-grams (unigrams, bigrams, and trigrams) would likely be preferable to our simple bag-of-words implementation. Though this ambition may be unrealistic given the size of our dataset and the time needed to implement.

Finally, to deal with the earlier addressed issue of our dataset not mapping to the entire U.S. Legal Code as a result of an insufficient number of cases applying to that area of the U.S. Code we offer a short conception of a possible augmentation:

To deal with these low reference values and unmapped laws we might engage a second dataset: the text of the U.S. Laws themselves. With this second data set we might engage an associative clustering algorithm to derive similarity between individuals laws based off of their non-standard terms (a function of tf-idf processing). This output ultimately would feed into a clustering analysis of all U.S. Federal Laws which we use to associate laws to one another. This association of Laws would allow us to thus provide a proximity value for unmapped laws from the first dataset model thereby providing us more laws within our models’ ‘reach’.

Thus by attaching clustered laws to the output targets from the Machine Learning Model we may be able to map to a much broader segment of the U.S. Legal code.