Blackstone Case Study on NLP and Law

Regan Meloche - Dec 11, 2020

The following is a summary of a Blackstone case study on using Natural Language Processing in the field of law, and how it might be relevant to a **policy difference engine.**

Link to Full Article

ICLR&D and Blackstone

The International Council of Law Reporting for England and Wales (**ICLR&D**) is a research lab dedicated to exploring and improving the law ecosystem using design, data science, and working in the open.

One area they've researched is how Natural Language Processing (NLP) can be used to impose control and structure on legal content generated in uncontrolled environments. To this end, the Blackstone project was devised in August 2019. A key deliverable is an open source library that allows the automatic extraction of relevant info from unstructured legal texts.

Two Strategies

There are two general strategies that were considered for this project. The first is a rules-based approach, where specific rules are devised around the information you want to extract. This may involve encoding some specific processing around certain phrases or words (for example if you want to identify responsibility, you may target "obligation" words such as "should", "shall", "owes", etc. This approach can work well with predictable raw data, but as the problem space expands, it can be difficult to capture all the different permutations of the extraction targets.

The second approach is prediction-based. This involves training a statistical model to form a generalized view of of concepts we are interested in extracting. This data-driven approach has high potential and has many use cases in other environments such as spam email identification. However, it is a little bit of a blackbox, since the methods of extraction are buried in a complex processing pipeline. It is also a statistical approach, so we can never get 100% accuracy.

The Blackstone project took a blended approach, where a prediction-based model performs the main extraction, while a rules engine can support it by filling in the gaps.

Building the Product

The product was built in part using **SpaCy** , a well-documented and feature-rich Python library with lots of built-in NLP functionality.

As noted, the predictive approach is data-driven and relies primarily on pre-processed training data. The training data for this project consisted of archives of various law reports and judgments.

The training phase involved many standard NLP techniques, such as cleaning the data, extracting sentences, identifying parts of speech (nouns, verbs, punctuation), and identifying relationships between the different words (e.g. formally linking an adjective to the noun it describes)

The model could be trained to recognize specific named entities, such as cae names, citations, acts, courts, judges, etc. Furthermore, it could be trained to categorize a sentence into concepts such as axioms, conclusions, and issues. This was done using a supervised learning approach, where the model is trained on a large corpus of examples that has been pre-labelled with these concepts. The model learns how to identify a concept based on the features of the text it is given.

Key Takeaways for a Policy Difference Engine

This is a very valuable case study for the development towards a PDE, which may involve using similar NLP techniques to analyze legislation and policy. Some potential goals may include extracting logical rules from written legislation to facilitate the translation to **Rules as Code**.

One component of the PDE is the “Rules Module”, which seeks to analyze the linkages and conflicts between different rules. At this point, this is very much a manual process, so if a policy is changed, then a rule writer may need to pore over many other rules to find out if there are any unintended side-effects of the rule change. NLP may be used to help identify rules that are closely related to a target rule to make this task more focused.