CS 221 Project Proposal

Derek Yan - zhyan;   Tai Guo - taig

Personal Legal Counsellor and Interpreter of the Law via Machine Learning

Motivation - “The first thing we do, let’s kill all the lawyers” - William Shakespeare, 2 Henry VI, 4.2.59. Any major transaction, legal procedure, patent dispute always require an attorney-at-law in the due process. However, the cost of seeking such legal counseling or hiring a lawyer is often astronomical for everyday people. Even though there are always legitimate cases of people who fall victims to unjust treatment; contractual violation; patent infringement; and etc, these victims often do not have the financial resources to seek justice through legal counseling. Our team looks to create an interactive legal counselor and interpreter of the law for the disenfranchised.

Definition - The tool will take in an user-input legal dispute. It will output a prediction regarding the subject of the law (contract, tort, patent, or etc). The tool will also provide prior art/relevant cases which have already been decided and are similar to the dispute at hand. It will also finally predict the likelihood of winning/settling this case along with any other useful information. Users would receive a vote of confidence regarding their legal concerns before proceeding with legal actions.

Evaluation - Our tool will first classify the user-input case to a subject of the law. At this step, we will use error rate to validate. Then our tool will predict the probability of winning this case. At this step, we will use a calibration plot and the error rate to evaluate the tool’s quality. Finally, we will recommend some prior art/relevant cases. At this step, we will use precision and recall.

Example of input and output - The inputs are legal case briefs crawled through www.casebriefs.com and www.casebriefsummary.com. These websites store over 3800 case briefs of legal proceedings in the United States in a wide range of subjects of law. 90% of the legal case briefs will serve as training data set while the other 10% will be used as test data. For example: the following is a sample case brief:
INPUT-Title: Eisenstadt v. Baird
    Category: Family Law
    Facts: Appellee (Baird) was convicted for exhibiting and distributing contraceptive articles under a law that forbid single as opposed to married people from obtaining contraceptives.
    Issue: Is there a rational ground for the different treatment of married and unmarried persons under the Massachusetts State law?
OUTPUT-Holding: No. The dissimilar treatment of similarly situated married and unmarried persons under the Massachusetts law violates the Equal Protection Clause.

The input would be facts and issue of a case. The output would be proceeding of the legal procedure. In the above example, the defendant would be not guilty because contraceptives distributed to married couples are in similar situation as those who are not married. We would also output prior cases which are similar to the case at hand.

Baseline - The baseline solution would be using linear regression to serve as a predictor of future cases. We would use feature vector to extract word features such as word-grams from Facts and Issues from the training set. The weightings of features extracted from Issues section will be more significant as they deal specifically with a law in question. When the new input is given, we would extract the features, and then use stochastic gradient descent to come up the learn-predictor to estimate the likelihood of the plaintiff winning the case. For the test data, we had an error rate of around 30%. Using word features only got us around 70% of correct prediction.

Oracle - The oracle is manual interpretation of facts and issues given a test case by the group members. We read over the facts of a legal proceeding, the interpretation of law which is under question, and use our common sense to give a probable decision. For legal matters that are not familiar to us, we would research laws of such matter and make human predictions. Since there are two members in the team, for the cases that we do not agree, we would discuss and compromise on one decision. This would serve as our oracle. The test error would be small but still non-negligible because the group members are not of the legal profession. For our test cases of 20, we had an error rate of 10%. This results in a gap of 20% between the baseline and the oracle. It means that more feature extraction and well as labeling can be done to get a better result.

Challenges and Solutions - The civil and judicial laws are different from State to State. Similar legal disputes may result in different proceedings in different States. If indeed this does happen, we will need to build clusters of cases that are of the same State. We will assign more weighting for prior cases in the same State, and less weighting for the cases that are not. Moreover, another challenge is that the holdings for case briefs is not binary. Often, cases are settled out of court as opposed to be decided by a judge. For such cases, we will need to extrapolate what the likely outcome is if the proceeding went to a final hearing based on the facts and issues given.

Related work -
FindLaw.com - A website that allows users to search for lawyers based on legal counseling needs. There is also a forum that users answer each other's questions regarding legal disputes.

This will be a joint project from CS229. Derek Yan (zhyan) who is taking CS221 and CS229 concurrently and Patrick Chase (pchase) and Tianyi Wang (tianyiw) who are taking CS229 will be presenting the results of this project in CS229. The project has received permission from Prof. Ng from CS229.