FINAL PROJECT

OUTLINE PROPOSAL

Paul Dean, Soobin Hwang, Seunghwa Jun

Team Name: Team Supreme

Team Members Paul Dean, Soobin Hwang, Seunghwa Jun

Topic: Text analysis on court cases

Data Set: 200 recent Supreme Court cases (Tax: 100 / Criminal: 100)

Tasks:

* Task 1 – Collect Data
  + Scrape the text, URLs, and supplemental data (case name etc) on 100 criminal and 100 court cases from the Supreme Court website. This is publicly available information.
  + Tools used: requests.get, BeautifulSoup, JSON
* Task 2 – TFIDF
  + Perform TFIDF analysis on the text of the court cases. This will tell us what most relevant terms and topics are for each case. This will allow us to then create a website which allows users to search for cases that are relevant to their search terms (e.g. cases related to share redemptions).
  + Tools used: pySpark
* Task 3 – Naïve Bayes analysis
  + Use a Naïve Bayes model to predict whether court cases are related to criminal law or tax law. I.e., we will pass the text from 70% of the cases as training data, and then will use the text from the remaining 30% of cases as test data to predict whether the model can tell if the case is tax law-related or criminal law-related.
  + Tools used: pySpark or TensorFlow
* Task 4 – Word Clouds
  + We will create word clouds to visualize what the most used terms are in criminal cases, what terms are most used in tax cases, and both.
  + Tools used: Plotly? Tableau?
* Task 5 – Host this information on a webpage
  + Page 1: users can perform the search for documents (using TFIDF scores to sort search terms by relevancy).
  + Page 2: will display the results of our Naïve Bayes classification model – how well did it do on the training and test data?
  + Page 3: will display word clouds
  + Tools used: Amazon Web Services or Google App services, HTML/CSS/Bootstrap/Javascript