PRELIMINARY DATA REPORT

BENCHMARKS: A Citizen's Scorecard on Judicial Accountability in Massachusetts

CS 506 Instructor: Professor Lance Galletti

BU Spark! Advisor: Professor Maggie Mulvihill

BU Spark! Mentor: John Merfeld

Fall 2019

TEAM MEMBERS

NAME	EMAIL ID
Xiao, Yuwan	yuwan@bu.edu
Jain, Abhimanyu	jainabhi@bu.edu
Njavro, Anton	njavro@bu.edu
Das, Yashvardhan (Team Leader)	yashvdas@bu.edu

INTRODUCTION

In order to get a comprehensive analysis of civil and criminal case reversals in Massachusetts over the past decade, first, we need to get all the cases (2008 to 2019) from the Massachusetts Appellate Court Website -

https://www.mass.gov/orgs/massachusetts-court-system

The previous team affiliated with this project had completed the scraping of cases that occurred during the time period of 2008 - 2018. The analysis was done mostly in criminal cases. The current team's focus is to scrape the related web pages pertinent to the cases that took place between the mid-2018 and late-2019 and analyze the reversal of the civil cases.

Based on the scraped pages, we are leveraging the keywords provided to perform text mining to find all the reversed cases. We mainly have two datasets to further perform the analysis on:

i) one for the reversed criminal cases

ii) one for the reversed civil cases

DATA PREPROCESSING

- **SCRAPING THE RESPECTIVE WEBSITE**

  We have used the code base of the previous team to scrape the recent year cases (2018 - 2019). We already have the data ranging from 2008 - 2018 through the work done by the previous team. We use the BeautifulSoup

library of Python to scrape the data from the Massachusetts appellate court website. Since, the website has a limit for viewing and extracting details, for now, we only scrape part of the cases to do the basic analysis.

- **FINDING THE REVERSED CASES**

  In order to make a determination about whether each case was reversed or not and to gather identifying information about each case, we use the list of key terms provided to decide.

  Each page contains the parties involved, the nature, title, and status of the case, the legal counsel present, information about the lower court that initially decided the case, and the list of docket entries detailing the lifecycle of the case.

- **PREPARATION OF THE DATASET**

  To accomplish this, we used the HTML parsing package of BeautifulSoup to parse the name of each bolded tag on the page and the value next to it, interpreting these as column titles and entries. The docket entries were parsed separately, with each docket entry being recorded as the case ID, entry date, and entry text. Thus, each case could be converted to a dictionary of key-value pairs, including case id, title, nature (criminal or civil), status (e.g. closed), plaintiff, defendant, relevant dates, and a list of docket entries. We also added two new columns that indicate the docket entries containing the strings "affirm" or "revers", respectively (the latter to capture both "reversed" and "reversal"). This allowed each case in the appellate court system to be represented as a dictionary object which could be used for later

analysis. For the scope of this project, we limited the analysis to cases whose type was *Criminal* and status was *Closed: Rescript* or *Decided*: *Rescript* (indicating the appellate court had decided the cased and issued its decision back to the lower court).

CHALLENGES FACED DURING SCRAPING

- It was quite a difficult task to gather the entire data accounting for the last 1.5 years. Since the concerned website is associated with the judiciary, there can be several restrictions regarding the extraction of information from the requisite web pages. Because of the website limitation of preventing scraping a lot of data per day, for now, we still have a part of the entire data which is not scraped yet, which will be taken care of soon.
- All though all of the web pages have been scraped, some of the scraped web pages show a forbidden access error. This has come into existence because of sending too many requests to the website.
- Finding all the reversed cases is also somewhat cumbersome. To decide if the case is reversed, we might need to add more keywords to make sure we get the whole list of reversed cases. We need to make sure that the reversed cases we find are validated. This website has fairly sophisticated anti-scraping techniques applied to it, so we manually extracted the data.

PROBLEMS ENCOUNTERED DURING TEXT ANALYSIS

Another important challenge we encountered is the large and relatively complex legal text. Each case and appeal comes with the title of the case and lengthy text/headnote. Legal text mostly describes the case in detail and due to current lack

of our legal experience we are not able to extract much of the meaningful data for now. The interesting problem will be to see how can we standardize context extraction out of such complex texts and how can we extract certain features from it such as conviction status and details of involved parties.

NEXT COURSE OF ACTION

- Exploring the reversed cases to find the key factors that may affect the reversed decision, eg. a specific judge or weak witness.
- Extracting sentiment/context from legal texts.
- Extracting specific case-related data from legal texts.
- Analysis of appealed cases.
- Performing exploratory analysis based on certain correlations amongst certain factors.