Virginia Courts Data Scraping Project

A Technical Report
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Andrew Kim

with

David Alves
Matthew Bacon
David Stern
Jessie Shen

May 13, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Andrew Kim

Capstone advisor: Jack Davidson, Department of Computer Science

How can the accessibility of Virginia Court Data be improved for the public? The state of Virginia understands the importance of the accessibility of online court data. However, their current implementation lacks a friendly user interface, is unintuitive to use, and lacks key features that allow users to search for and analyze data. A small team and I worked with UVA Law's Legal Data Lab to optimize their system for public use. This project involved further developing a web-scraper that pulls court data and gets converted into JSON format. The JSON data then gets ingested into a MySQL database. MySQL allows us to easily store and query data that users can quickly access. Now, users can have the ability to conduct filtered searches within the Virginia Court database and export the data that they need for further manipulation and analysis. The final product would be to host the scrapped court data on our website for users to conduct easy searches of court data within the state of Virginia.

Because this project deals with sensitive legal data, the team operated while keeping privacy and security a top priority. Therefore, the first couple weeks of the project consisted of completing a thorough Institutional Review Board (IRB) training seminar, properly setting up the coding environment, and examining the existing codebase by Ben Schoenfeld. One such high-security decision was deciding how to handle public and private data. Private data include full names of defendants and plaintiffs, as well as their addresses. In order to address this, the team came up with the idea to hash these values to minimize any potential infringements of privacy.

My chief contribution was developing a query function in Python for the data that gets stored in the MySQL database. This query allows the user to make a refined search within the database, ouputted in a comma separated value (csv) file. Some of the search parameters for the

query function involve name, date of birth, and sex. The querying function provides flexibility where the user can find their desired results given one parameter to all possible parameters.

In the future, it is possible for this project to be further implemented to other states. Additional enhancements to the user interface could be made as well.