UCLACaseLawCite

Final project for UCLA's Data Science - Exploratory Data Analysis and Visualization

Obtaining the Data

Bulk case data

Bulk case data can be downloaded from the following URL:

- Download Page
- Direct Link (465 MB)

```
mkdir Data && cd Data
curl https://api.case.law/v1/bulk/22341/download/
```

Case citations

Citation can be found here:

- Download Page
- Direct Link (165 MB)

```
cd Data/Illinois-20200302-text/data
curl https://case.law/download/citation_graph/2020-04-28/citations.csv.gz
```

Preparing the Data for analysis

First be sure to install the required Tools as listed here:

- Tools

After downloading the data into the `Data` directory we can use the python script included in `./ETL/hcapetl.py` directory to transform, clean and insert the data into a SQLite database that will simplify our analysis.

The data must be extracted first with these commands:

```
DATA=Data/Illinois-20200302-text/data
DPROC=Data/Processed

xzcat $DATA/data.jsonl.xz > data.jsonl
jq -s $DATA/data.jsonl > $DPROC/data.json
```

The database will be named `hcap.sqlite` and it can be created by the following commands:

```
dbpath=./hcap.sqlite

./ETL/hcapetl.py create tables "$dbpath" ./Database/*.ddl.sql
./ETL/hcapetl.py create attorneys "$dbpath" ./Data/Processed/data.json
./ETL/hcapetl.py create cases "$dbpath" ./Data/Processed/data.json
./ETL/hcapetl.py create citations "$dbpath" ./Data/Processed/data.json
```

Running the full ETL pipeline should take about 10 minutes (excluding data download).

The previous commands can be found in the `gendata.sh` script at the root of this project.

The `ETL` directory has all of the python source necessary to work with the data. To aid with the exploration and cleanup we have the following Jupyter Notebooks:

1. Attorney Name Parsing
2. Attorney Record Parsing

3. Attorney Career Overview
4. Influential Case Attorneys

The Data Exploration file has information about the commands used to gain insights to fragments of the data and to determine a SQL db schema.

Analysis

As mentioned in the project report we seek to answer the following questions:

1. Who is the attorney that has had the most participation in cases?, from private parties?, from the government?
2. How much the work in which an attorney is involved is cited, e.g. how influential was the work.
3. What is the page count of cases in which an attorney has participated?

Those questions are answered by the following respective Jupyter Notebooks, and the findings presented in the project report:

1. Q1 Most Influential Attorneys
2. Q2 Attorney Case Citations
3. Q3 Average page count for cases