Samuel_Woolledge_1

December 10, 2019

IMPORTANT

Please read the README file before trying to run this .ipynb. This project relies on the library “pypatent” for scraping data, which can easily be installed by running “pip install pypatent”.

GitHub link: https://github.com/SamWool1/cse184-FinalProject

Scraping and Cleaning Data

Everything in this section does not need to be ran again. It was used in order to construct the dataset from which our visualizations were made, and this data has been saved.

Generating A Sample

In order to scrape our patents, we first needed to randomly generate a set of patent numbers which we would then scrape and use for our data wrangling. This step does not need to be run again, as the dataset has already been scraped and generated. It is assumed that a folder named “samples” already exists prior to running this code.

[ ]:

```
import gen_sample
for i in range(1980, 2019):
    gen_sample.sampleByYear(str(i))
```

Scraping

Once our samples were generated, we scraped our target website (http://patft.uspto.gov/) by constructing the URL for individual patents using the generated patent numbers. This scraped data became the dataset we used in order to construct our visualizations. This step does not need to be run again, as the dataset has already been scraped and generated. It is assumed that a folder named “scrapes” already exists prior to running this code, and that the patent numbers have already been generated.

[ ]:

```
import scrape_with_sample
for year in range(1980, 2019):
    scrape_with_sample.main(str(year))
```

Cleaning

Some data cleaning and minor wrangling was necessary before the data was ready for visualization. While cleaning for the racing bar chart for companies was handled within the function that generates the dataset for visualizing, additional cleaning and wrangling was used for creating a unified .csv file for the dataset for our other two racing bar charts. This step generates a single .csv file using the scraped data. It contains cleaned information pertaining to fields and locations for our scraped patents, and identifies the year of each patent as well.

[ ]:

```
import wrangle_data
for year in range(1980, 2019):
    print('Wrangling ' + str(year))
    wrangle_data.main(str(year))
print('All wrangling finished')


import wrangle_loc_fields
wrangle_loc_fields.createLocFieldSheet()
```

Wrangling and Visualizations

Wrangling

We used Flourish and Datawrapper to create our visualizations, which required us to have specifically formatted .csv files. We used functions in wrangle_data.py to wrangle our data into formats that these websites would accept. These functions rely on the scraped dataset in the folder "scrapes", as well as the cleaned and unified dataset created in the "Cleaning" section.

[ ]:

```
# import wrangle_data
# This should already be imported from the previous step

wrangle_data.makeRacingBarCountries()
wrangle_data.makeRacingBarFields()
wrangle_data.getTimeDiff()
wrangle_data.makeLineCSV()
wrangle_data.makeRacingBar()
wrangle_data.makeBarCumSum()
wrangle_data.createUnifiedScrape()
```

Visualizations

All of our visualizations are hosted at https://people.ucsc.edu/~tcpappas/cse184/home.html. Please click the link to view them. You can also view them using the "home.html" file in the "website" directory on our GitHub.

[ ]:

```

```