Data Wrangling – Google Mobility Index

Khai Nguyen (カイ·グエン)
khainguyen@temple.edu

CIS 4517

Data Intensive and Cloud Computing

1. Overview

This folder consists of a Web Scraping and PDF Scraping tool, generating csv data from Google's COVID 19 Mobility Report https://www.google.com/covid19/mobility/ to serve ongoing research efforts to fight against COVID-19.

2. Requirements

- BeautifulSoup4 : https://pypi.org/project/beautifulsoup4/
- python tika : https://pypi.org/project/tika/

3. Building blocks

- *scrapeWeb.py*
- *scrapePDFtoCSV.py*

Important functions in *scrapePDFtoCSV.py*:

- *processBigTerritory(metric_list, file_name)*
- *processSmallTerritory(stop_index, metric_list, file_name)*
- *scrapePDFtoCSV(path_to_file)*

4. Workflow

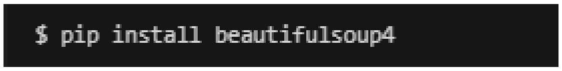
- Implement scrapeWeb.py with BeautifulSoup to pull all the pdf's down
- Start building scrapePDFtoCSV.py by searching for a pdf extraction tool
- Break down into 2 steps, inner-file processing and file-name processing
- Build *getCountryName()* function
- Build *processBigTerritory()* function
- Build *processSmallTerritory()* function
- Build *scrapePDFtoCSV()* with a sample pdf
- Build *scrapePDFtoCSV()* to process all pdfs in a folder
- Format csv output files

5. Testing

- Incremental tesitng when implement new feature, no assumptions

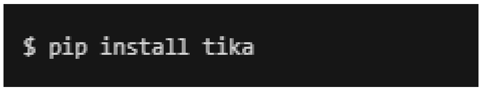
6. **Installation & Usage guide**:

- BeatifulSoup4



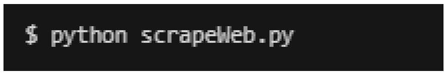
```
$ pip install beautifulsoup4
```

- tika



```
$ pip install tika
```

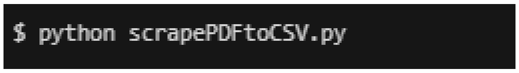
- For downloading all pdf file from Google's site



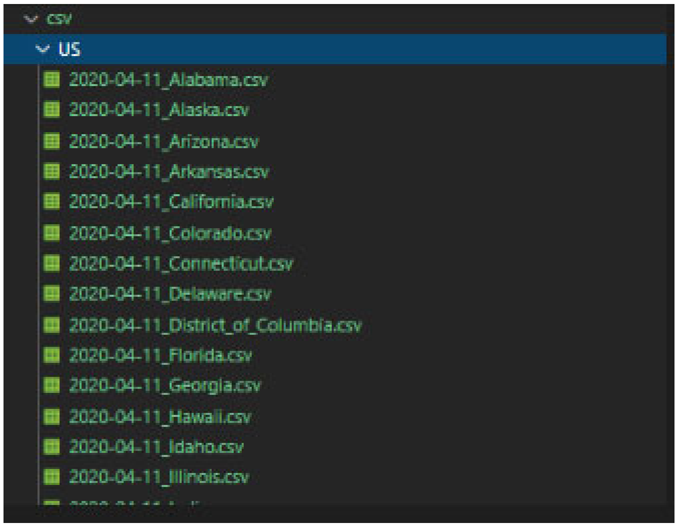
```
$ python scrapeWeb.py
```

NOTICE! Remember to pre-create the **"./csv"** and **"./csv/US"** folder before proceeding!

- For scraping data from PDF to CSV, you can either execute the program through Jupyter notebook tika-extraction.ipynb, or cmd line as below



```
$ python scrapePDFtoCSV.py
```



```
v csv
  v US
    2020-04-11_Alabama.csv
    2020-04-11_Alaska.csv
    2020-04-11_Arizona.csv
    2020-04-11_Arkansas.csv
    2020-04-11_California.csv
    2020-04-11_Colorado.csv
    2020-04-11_Connecticut.csv
    2020-04-11_Delaware.csv
    2020-04-11_District_of_Columbia.csv
    2020-04-11_Florida.csv
    2020-04-11_Georgia.csv
    2020-04-11_Hawaii.csv
    2020-04-11_Idaho.csv
    2020-04-11_Illinois.csv
```

7. Issues

- Google can change their web format, the format used to scrape the pdf files were initially <xml> (as in file *covid19_mobility.html)* on **04/11/2020**
- Error might be causes not declaring encoding type when read/write files, the choice would be UTF-8
- processing strings: wrong path, or "\n" not truncated
- text processing issues, delimiters, regex
- csv formatting
- Time management