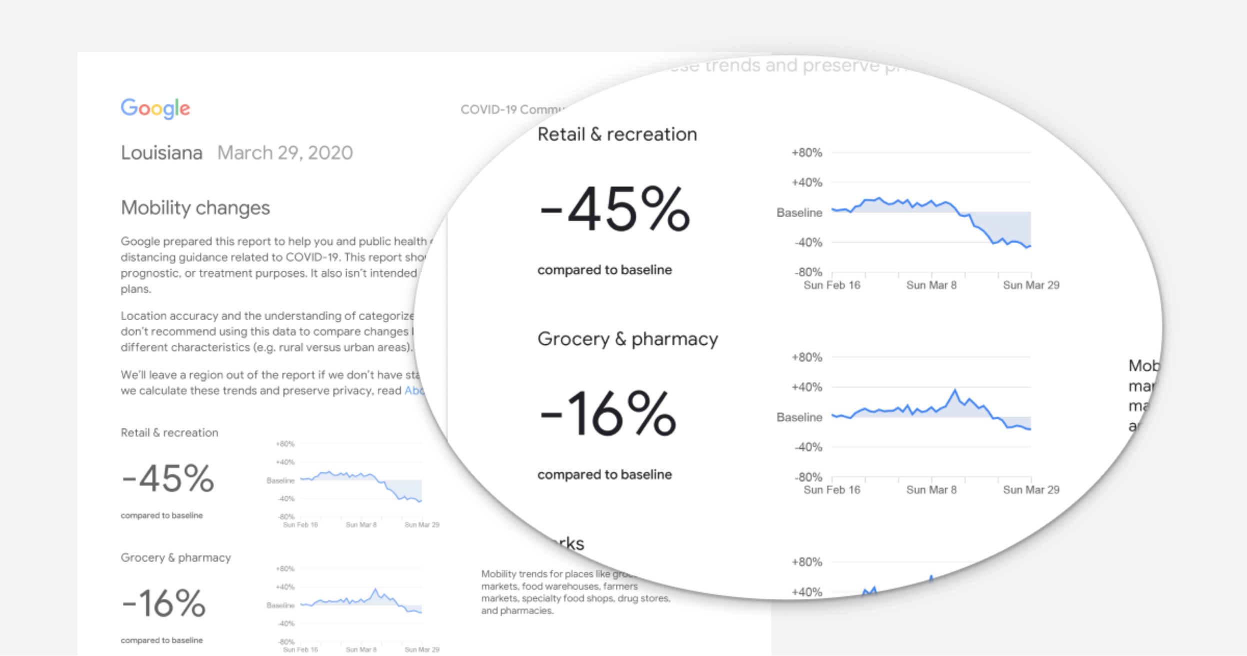
Quang Pham
Project 2

Report

Introduction

The main goal of this project is to download data from the Google COVID-19 Community Mobility[1] report website. By using a python application, 182 pdf files were downloaded, contained information of different countries and also different states in the U.S. Moreover, an extracting tool had been designed with the help of "Tika Parser" [2]with the goal to extract information from the 6 attributes of each pdf file.

Downloading data from website



Google
Louisiana March 29, 2020
COVID-19 Comm

Mobility changes

Google prepared this report to help you and public health
distancing guidance related to COVID-19. This report sho
prognostic, or treatment purposes. It also isn't intended
plans.

Location accuracy and the understanding of categorize
don't recommend using this data to compare changes
different characteristics (e.g. rural versus urban areas).

We'll leave a region out of the report if we don't have sta
we calculate these trends and preserve privacy, read Abc

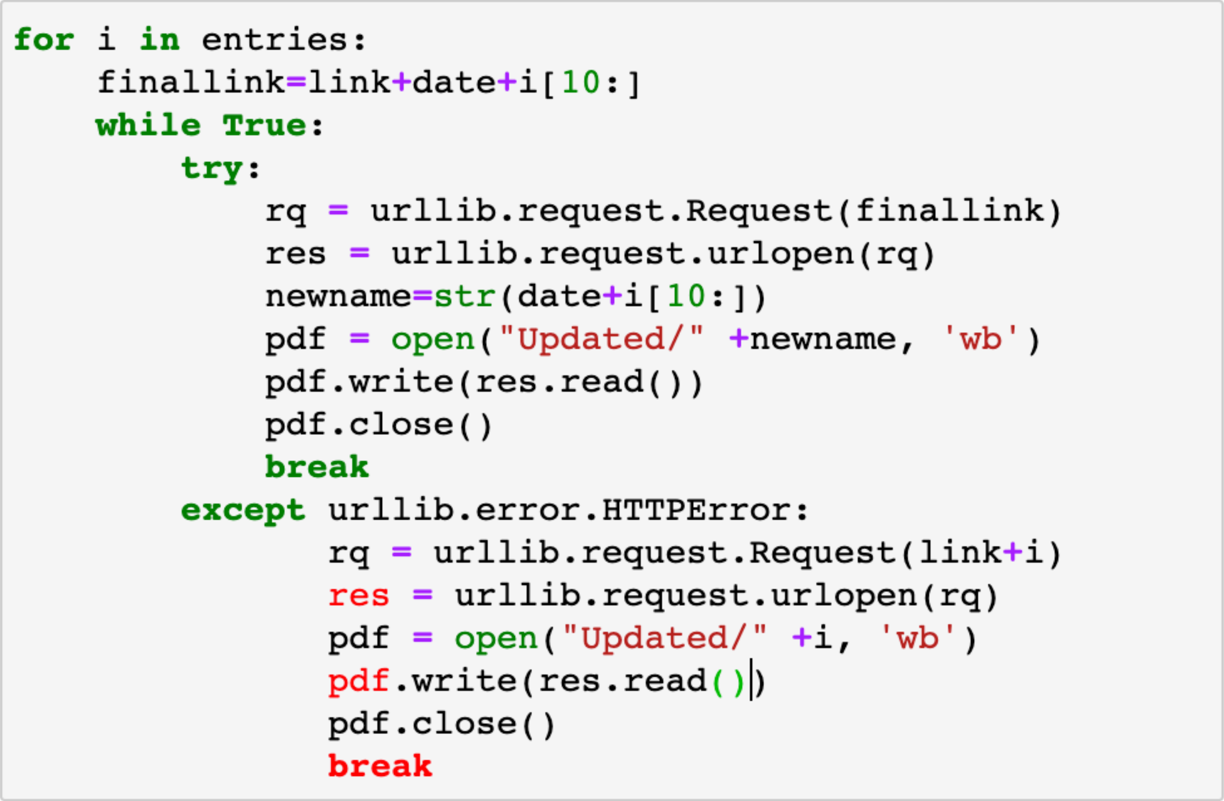
Mobility trends for places like gro
markets, food warehouses, farmers
markets, specialty food shops, drug stores,
and pharmacies.

The website was uploaded frequently, so different approaches were performed to update and maintain the data. The first attempt started with sending a request to the url, then get the response and the detail of the website in text format. A filtering algorithm was used to search for all the download links ended with ".pdf". Using this method, a dataset of 182 files was downloaded on April 14,2020.

[1] https://www.google.com/covid19/mobility/

[2] https://github.com/chrismattmann/tika-python

However, since the website was uploaded frequently, the HTML structure of the website had been changed a lot when I check my code again to upload news file. After observing that the download link followed a strict regulation, the downloading program was modified with the second part, where the tool will check for new files on the web based on the date, and if new file has been uploaded, the program will download that new update.



```
for i in entries:
    finallink=link+date+i[10:]
    while True:
        try:
            rq = urllib.request.Request(finallink)
            res = urllib.request.urlopen(rq)
            newname=str(date+i[10:])
            pdf = open("Updated/" +newname, 'wb')
            pdf.write(res.read())
            pdf.close()
            break
        except urllib.error.HTTPError:
               rq = urllib.request.Request(link+i)
               res = urllib.request.urlopen(rq)
               pdf = open("Updated/" +i, 'wb')
               pdf.write(res.read())
               pdf.close()
               break
```

As in the above figure, the program will check for update special 'date' and download pdf file if new update existed. New updates were saved in the Updated folder (182 files).

Extracting data from CSV

Tika Parser was used as the main tool to extract information from a pdf file, especially data in text format. For every file, 6 attributes in Table 1 was mined and saved into a data frame. A algorithm was designed to search for required values by checking special patter in different case. For the first case, the changes in mobility of a country was showed in the first page, so a searching tool was used to find pair of attributes and value. On the other case, a pdf file could have information of a country and many states or regions belong to that country, so the searching algorithms also have to search for the regions' names, besides 6 attributes in the Table 1. Normally, this name often stands right before "Retail & recreation" key words.

Attributes
Retail & recreation
Grocery & pharmacy
Parks
Transit stations
Workplaces
Residential

Table 1

Results

The project has successfully in both downloading and extracting data from the Google Covid 19 Mobility website. For the downloading process, a method to download and update data based on date and time was created and this tool was used to download and update 182 files. Secondly, the tool to extracting information related to the 6 attributes in Table 1 had been designed and tested, and it had been used to extract data from 182 pdf files into 182 csv files. These csv files can contain information not only from a country, but also from every states from that country, or just a single state in the U.S.