Homework4

Jae Young Kim

2918465924

# Section1. Motivation

As I was new in LA, I want to look for great hamburger restaurants in LA. To find the restaurants, I searched a lot of web pages. However, it was not easy to find a webpage which shows a map displaying best hamburger restaurants of overall LA. Therefore, I want to make a gourmet map of LA hamburger restaurants. If so, I can see overall distribution of famous hamburger restaurants and notice where they are concentrated. In addition, as I included the weather information of the area around the restaurants, it is possible to decide whether it is a good time to visit the restaurant or not. And also, I would like to analyze the restaurants’ reviews and comments and visualize them as a scatterplot to compare the analyzed reviews and comments with the blog’s rating. After that, I will make a regression model between the number of comments and the rating scores to check whether the number of comments is an influential variable in predicting the rating score.

# Section2. Data Source Info.

To get the information of the restaurants, I scraped a blog introducing best burger restaurants in LA. I scraped <https://www.timeout.com/los-angeles/restaurants/the-best-burgers-in-los-angeles>. Since for each burger restaurants introduced, there are following introducing pages such as <https://www.timeout.com/los-angeles/bars/everson-royce-bar>, I also scraped those following introducing pages. I scraped name, famous burger, rating, review, address, website from the blog.

After that, I used Yelp api to get the number of reviews, rating in Yelp, open hour list and comments.

Then I used Google natural language api to do the sentiment analysis of reviews and comments of the restaurants.

Then by using Google Geocoding api, I got the longitude and latitude of the restaurants.

Then by using Government weather api, I got the weather forecast of the longitude and latitude.

# Section3. Frequency of data update of the data source

As I scraped data from an article of the blog, it is static.

Yelp api updates data at least once a week. I checked the information change.

Since Google natural language api computes the result immediately, it is not adequate to say whether it is static or not.

I guess Google Geocoding api is static since the geographic information does not change.

Government weather api updates data at least once a day since it forecasts weather daily.

# Section4. Procedure of extraction and description of my code.

In project1\_scrape\_web\_page.py, I scraped data from <https://www.timeout.com/los-angeles/restaurants/the-best-burgers-in-los-angeles>.

From the article, I scraped the subtitles of the article. It looks like “Single Burger at Everson Royce Bar”. Then as some strings had non-English characters, I changed them. From these subtitles, I divided famous burger and the name of the restaurant.

There were some hyperlinks in the subtitle. Therefore, I scraped them. Then I scraped rating, review(description), address and web address from the links.

After that I merged scraped data to a data frame and made it to csv file named “1\_crawled\_data.csv”.

In project2\_yelp\_api.py, I got data from yelp api. To use this api, api\_key is needed. In the python file, I already declared my api\_key. If you want to use your own api\_key, you can make your own api\_key from <https://www.yelp.com/developers/v3/manage_app>.

With the api\_key, I got business id, the number of comments(review\_count), rating and open\_hour of the restaurants I scraped from the page. As there were some restaurants which does not open, I used try, except to express not-open status.(On 4/19, all of the restaurants had open hour but on 4/21, some of them disappeared. I guess some restaurants closed due to the corona-virus.)

Then I got 3 comments for each restaurants by using yelp review api and stored as a list.

After that I merged all the data to a Data Frame. Then stored it to csv file named “2\_crawled\_data\_with\_yelp.csv”.(Since this api is unstable, if there is error which says, there is internal problem, you may try 1 minute later. Then it will work.)

In project3\_nlp\_google\_api.py, with the data I collected, I implemented sentiment analysis. I used Google Natural Language api. To use this api, you have to authenticate your google account. It was not possible to use USC google account to use this api. So I used another ID. You can follow the following steps to run my code. I also included my api\_key json file in the zip file. If you want, you can use my api\_key json file.

Google Cloud Console🡪Create a project 🡪Search “Cloud Natural Language API”🡪Create Credential🡪Create a Service account🡪Create Key🡪Key type JSON🡪JSON file will be downloaded🡪register the path of the JSON file as an environment path of [GOOGLE\_APPLICATION\_CREDENTIALS](It can be different in MAC pc. I used Windows)

Then it is ready to run the code.

In the code, I defined sentiment\_analysis function, which does sentiment analysis of the given text. Then I used sentiment\_analysis on reviews and comments of the restaurants. After that I merged all the sentiment score as a Data Frame. Then stored it to csv file named “3\_crawled\_data\_with\_sentiment.csv”.

In project4\_geocoding\_api.py, I used Google Geocoding API. When the google account is authenticated in google api platform, there will be api\_key and you can use it to run this code. In my code, I included my api\_key. With the addresses of the restaurants I scraped, I got latitude and longitude of the restaurants by using geocoding api.

Then with government weather api, I got weather forecast and temperature data. Then merged the data with original data and saved as a csv file named “4\_crawled\_data\_with\_geo\_data.csv”.

# Section5. Discussion of how its model would be extended.

My first data source is an article from a webpage. Therefore, my data source is subjective and limited. If there is a website displaying more famous hamburger restaurants or different categories of restaurants, my result will be more dynamic.

Secondly, yelp api allowed only 3 comments for each restaurant. Therefore, the sentiment score I calculated can be skewed. If I can get more comments, the sentiment score will be more reasonable.