Group project title here

DNSC 6211: Programming for Analytics

Lingyao Meng
Aida Roxas
Yi Yang
Xing Zhang

Abstract

This project aims to address the need for a top list of brunch spots in DC that will always be current. We propose to utilize Yelp and Twitter to ensure our data is current as of social media feeds. The information will then be presented to users via mapping technologies.

Contents

1 Introduction

Sunday brunches is a religion in DC. For girls, a perfect date day should always start with a heart-warming brunch. However, when we look on-line, all seemingly elaborate blog posts about brunches in DC are out dated. With the armory of web-scraping, data analytics and mapping, we are looking for the most popular and twitter-worthy brunch spots in DC in 2016. The advantage of our project is that it should be able to update itself over the time so that it can update the user of their request. The output of the project should consist of two parts: a color coded brunch map and a list most popular brunch restaurants with name, rating, sentiment score and location. The color coded map should cover the area with the center at foggy bottom and a radius of 30 miles. The map contains 1000 most popular restaurants and they are coded by ranking, with different colors based on their rating.

2 Background

In total, the project should include data from two sources: Twitter and Yelp. Each row in the yelp data set should include the information about the restaurants name, location, rating, pricing and reviewers. Then based on the location of these restaurants, our group found the sentiment score for each of the restaurants from twitter. We want to do sentiment analysis on twitter posts and comments. Then for yelp we want import Yelp Hotness directly from yelp. But we also want to experiment and find out if there might be different results from yelp score. Then we are going to combine both score and present them on the Tableau mapping.

3 Method

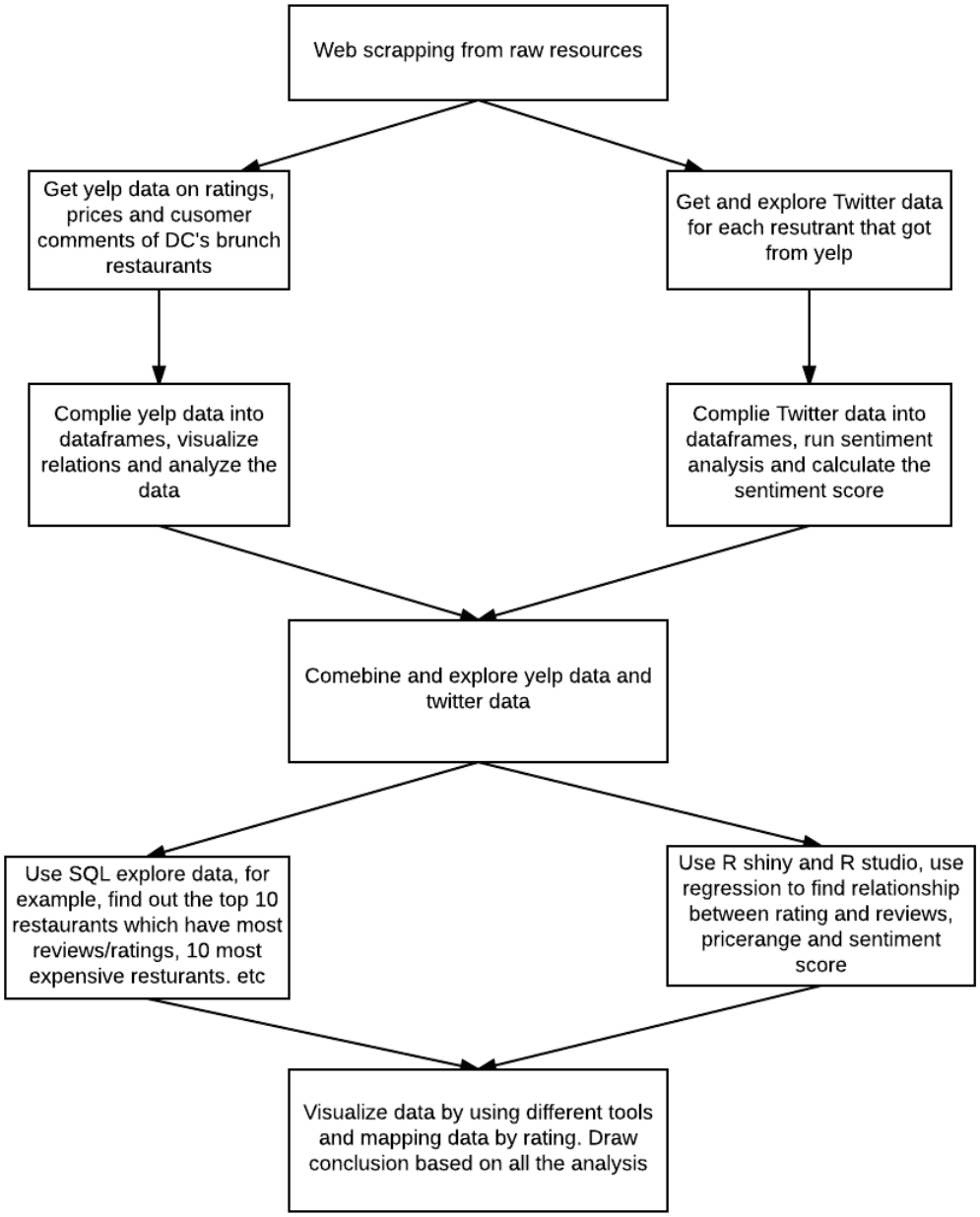
Using web scraping tools, we will get the restaurant information from Yelp. We need the restaurant name, address, ratings, price range, and reviews information. This will be converted to a csv file. Next, we will match the Yelp restaurant list from Twitter to find out if social media is discussing or tweeting about the locations. We will then get the sentiment score for the restaurants mentioned on Twitter. From both social media sites, we can get the geolocation of the restaurants and will now be ready to map the information for visualization. We will use Tableau to map the review ratings and we added the sentiment score as a label for the users. We used geocod.io to map the restaurants for additional information. We will then use R shiny and use rating as the dependent variable and try to find any significant relationship between the other variable in our datasets. We will investigate the possibility of being able to predict a restaurant's rating based on number of reviewers, twitter sentiment, and price range using R programming.

4 Organization

We are going to share the workloads for coding and presentation. Lingyao Meng and Yi Yang will be responsible for getting resources from twitter and yelp, running sentiment analysis and calculating sentiment score. Aida Roxas and Xing Zhang will do the regression part between rating score and other variables, visualize data by mapping based on rating score and finish report. We will be working on the video together and preparing for the final presentation together and try to share all the workloads equally.

4.1 Workflow

We have a process workflow diagram which indicates how we are going to scrap data, utilize API, compile data and present our output. Our data extracting process has two components, one is using web scraping techniques to scrape data from yelp.com, the other is utilizing API authentication method to explore data from twitter. Then we will put the data into data frame and compile the data for analysis using related tools, such as SQL, R program and R shiny. Lastly we will get a color coded map for top brunch places around DC area by rating.


Web scrapping from raw resources
Get yelp data on ratings, prices and cusomer comments of DC's brunch restaurants
Get and explore Twitter data for each resutrant that got from yelp
Complie yelp data into dataframes, visualize relations and analyze the data
Complie Twitter data into dataframes, run sentiment analysis and calculate the sentiment score
Comebine and explore yelp data and twitter data
Use SQL explore data, for example, find out the top 10 restaurants which have most reviews/ratings, 10 most expensive resturants. etc
Use R shiny and R studio, use regression to find relationship between rating and reviews, pricerange and sentiment score
Visualize data by using different tools and mapping data by rating. Draw conclusion based on all the analysis

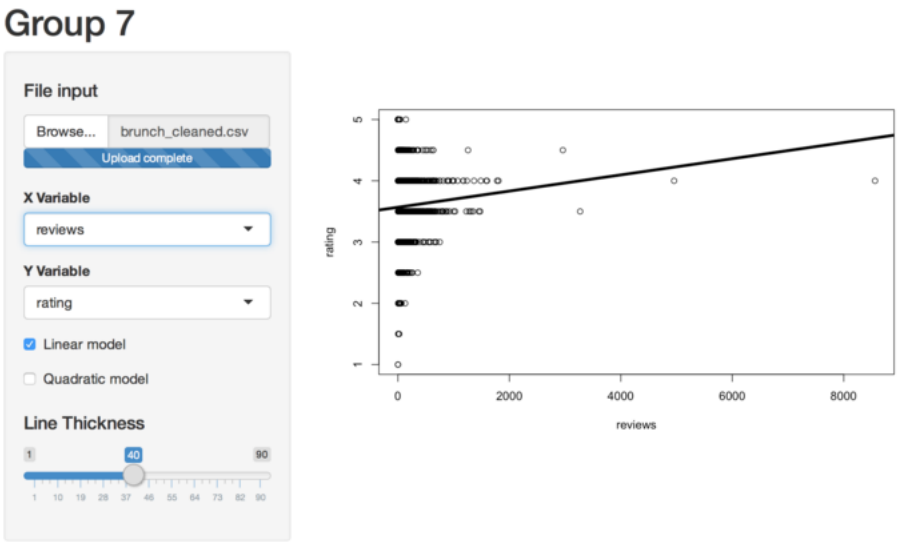
4.2 Project structure

We used yelp top brunch and breakfast and twitter for our data sources. They are related to each other because for people like to post their comment and how their feel about the restaurants that they had. While yelp has the information about each restaurants, we are going to find out how these restaurants perform on twitter and calculate their sentiment score, also we are going to find if there is any relationship between rating and other variables.

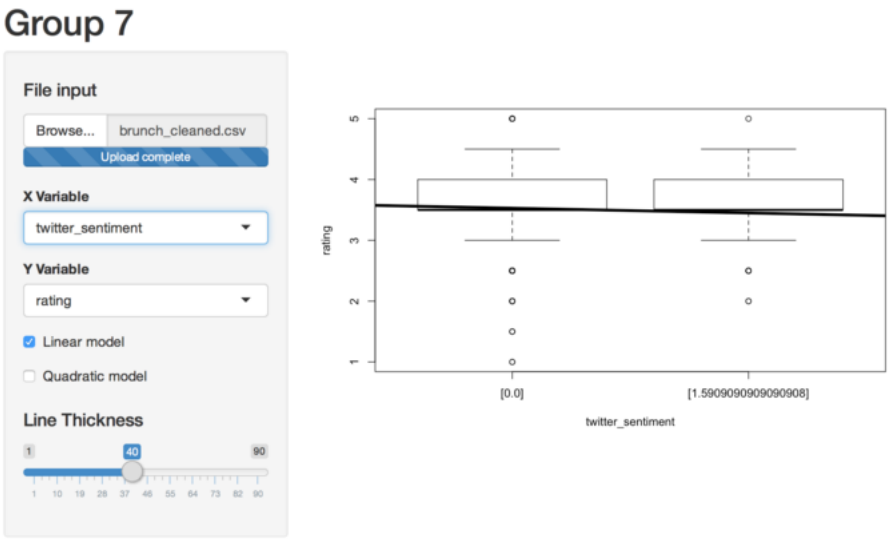
4.3 Figures and Tables

There are regression table and figures that we got by using r shiny and r/r studio to find some relationship between rating and reviews, price range and sentiment score.

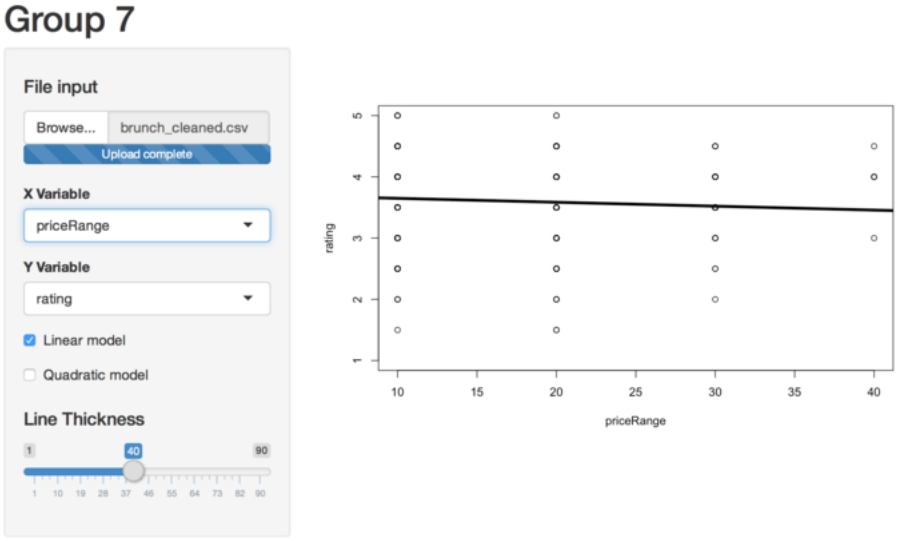
rating and reviews



rating and twitter sentiment score



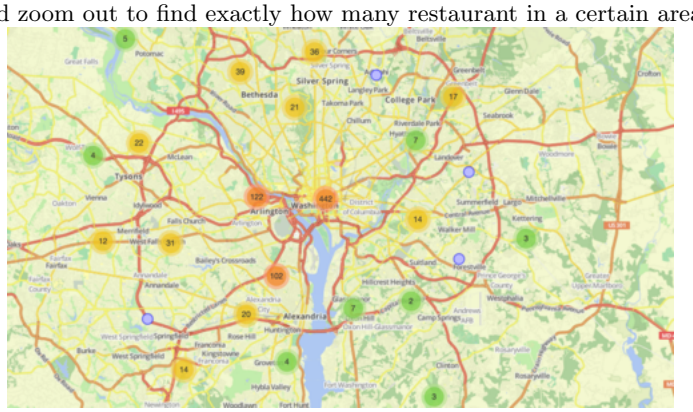
rating and pricerange



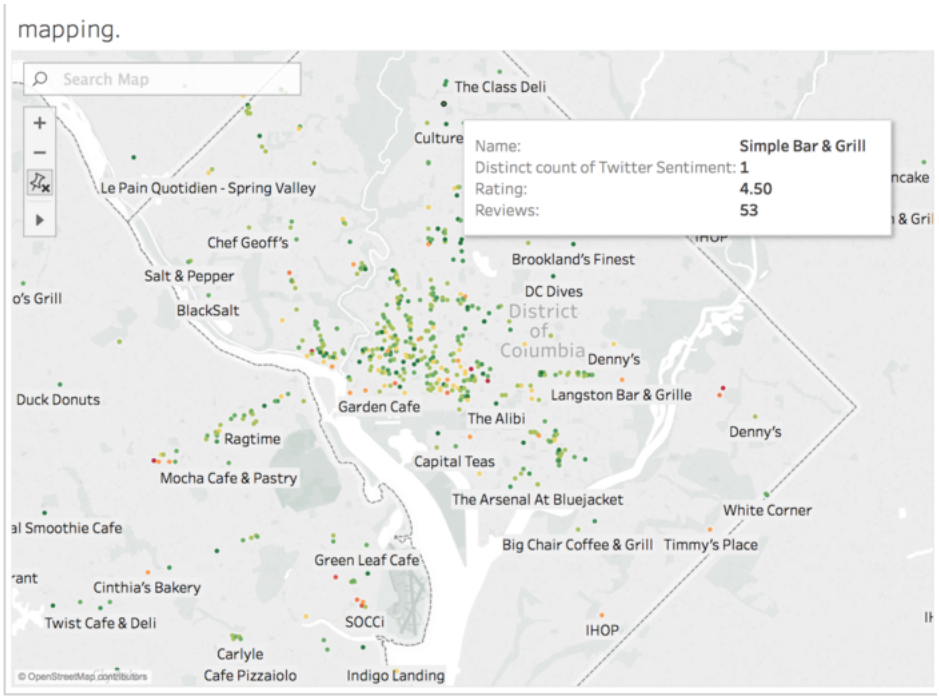
```
> fit3=lm(rating~priceRange)
> summary(fit3)

Call:
lm(formula = rating ~ priceRange)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15001 -0.15001 -0.08567  0.41433  1.41433

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.714357   0.055722  66.659   <2e-16 ***
priceRange  -0.006434   0.003009  -2.139   0.0327 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5797 on 963 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.004727,  Adjusted R-squared:  0.003693
F-statistic: 4.574 on 1 and 963 DF,  p-value: 0.03272
```

Taking rating and price range as example, based on the significant p-value and graph, we can see that there is negative linear relationship between rating and price range.

There is a mapping that shows how many restaurants in each area, and we can zoom in and zoom out to find exactly how many restaurant in a certain area.



There is another mapping that we draw by rating, each point represent a unique restaurant. Color green to red represent rating from 0 to 5, we also label the name, sentiment score, rating and reviews for each restaurant.



5 Discussion

A wonderful brunch is all we want after waking up in the morning, also as a group of girls, we like to find nice food on twitter since people like to post pictures of nice food that they had in their day. So the value of our project is we list the 1000 best brunch places based on yelp reviews, price range, number of posts and their sentiment score from twitter. One of the selling point of our project is our data is based on two of most popular social media: Yelp and Twitter so we have large data set and updated in every minutes, so our result is very accurate. Also, the results provided from our project is based on these elements: Yelp rating and reviews, twitter sentiment score, so our result is very persuasive, we let the data talk, not the objectives of food bloggers.

5.1 Learnings

From the data that we got from yelp, we can see most of the rating for the 1000 brunch restaurant is between 3 to 4, the price Range for majority of the restaurants are between 10 to 20. Also form the map, we can see most of the brunch restaurant we find are located in DC area, and foggy bottom is in the center of this area, some of them located in Virginia and Maryland. More than half of the restaurants have sentiment score 0, chain catering enterprises, such as IHOP has much more advantage to have higher sentiment score. Based on all regression analysis, we make a conclusion that rating score actually is not much depend on any of other variables that we considered, only little bit negative linear relationship with price range, which our group believe that rating score is fairly based on the quality of the food.

5.2 Challenges

We were not able to find the best way to come up with the overall point for each restaurant in a fair way based on the elements that we considered. We were still figuring out what weights for each element should be considered in the overall score for restaurant. No matter which way we decide to calculate, the result is not objective enough.

6 Conclusion

We believe in the importance and value of the information that we will be able to give access to users. The amount of time and finances spent on trying out places to have brunch at, is just not effective. We aim to put together, at a minimum, twitter and Yelp data, to provide a real time picture for people in the area, locals and tourists alike. On a project perspective, our group believes that this is challenging enough to put together all lessons we have learned this semester. We also like the idea of making analytics work in a practical sense.

Another value of what we are trying to achieve is to evolve to include other social media sources that are updated live by users. Everyone of us are sensors, the challenge is how to maximize the flow of information right in our fingertips.