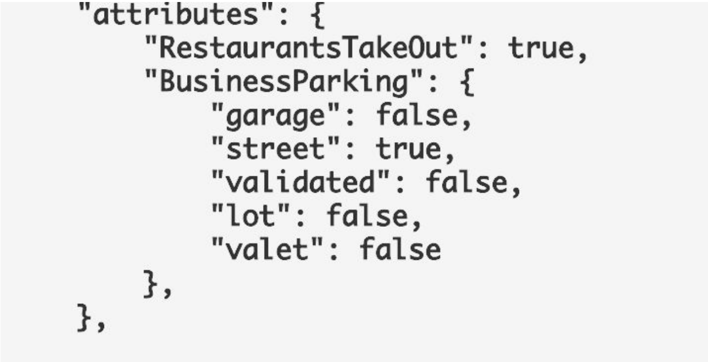
Yelp Restaurants

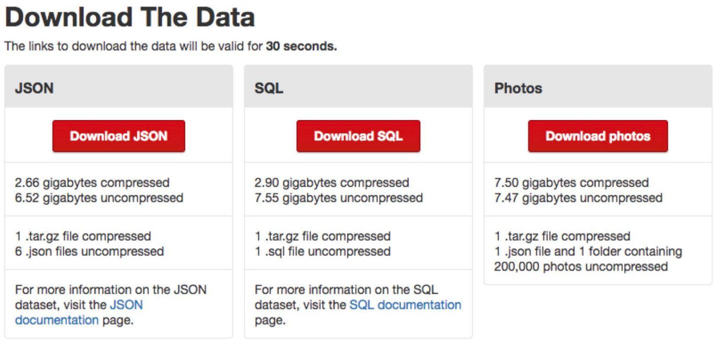
CMSC12300 Project
Group: WACC

Dataset

- Yelp public dataset available on: https://www.yelp.com/dataset/download
- Size: 6.52 gigabytes JSON file
- Contains information about 174,000 businesses (over 1.2 million business attributes like opening hours, parking, availability, and ambience) and 5,200,000 reviews contributed from 1,300, 000 users.



```
    "attributes": {
        "RestaurantsTakeOut": true,
        "BusinessParking": {
            "garage": false,
            "street": true,
            "validated": false,
            "lot": false,
            "valet": false
        },
    },
```



Download The Data

The links to download the data will be valid for **30 seconds.**

JSON	SQL	Photos
Download JSON	Download SQL	Download photos
2.66 gigabytes compressed 6.52 gigabytes uncompressed	2.90 gigabytes compressed 7.55 gigabytes uncompressed	7.50 gigabytes compressed 7.47 gigabytes uncompressed
1 .tar.gz file compressed 6 .json files uncompressed	1 .tar.gz file compressed 1 .sql file uncompressed	1 .tar.gz file compressed 1 .json file and 1 folder containing 200,000 photos uncompressed
For more information on the JSON dataset, visit the JSON documentation page.	For more information on the SQL dataset, visit the SQL documentation page.	

How neighborhoods affect success of restaurants

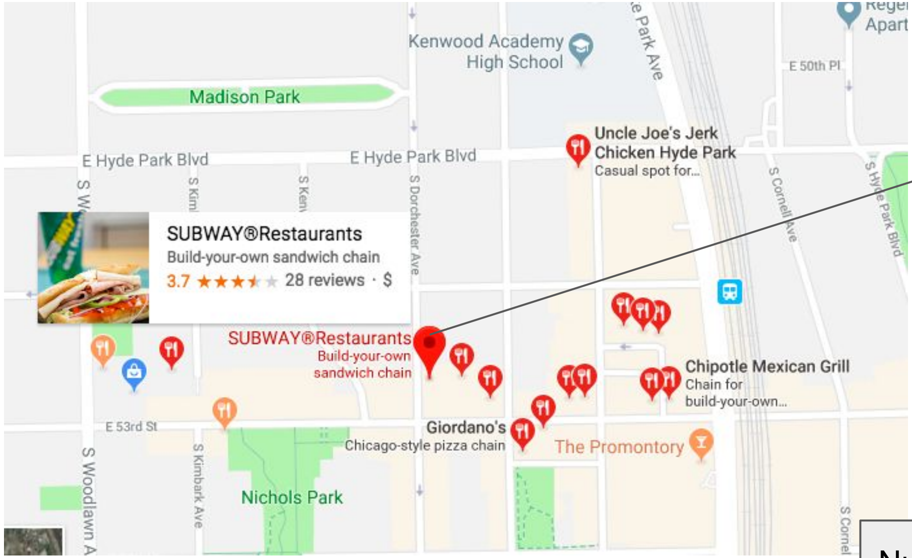
Hypothesis:

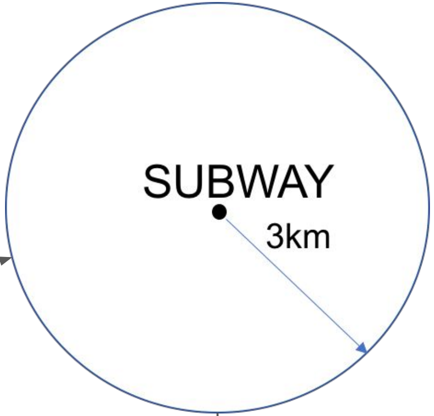
Features of neighborhood affecting success of restaurants within it.

Features include:

- Number of restaurants
- Numbers of categories
- Category similarity
- Reviews similarity
- Average rating
- Price range
- Number of reviews
- Average sentiment of reviews of all restaurants in the neighborhood

Identifying a Neighborhood





Step 1: Determine the neighborhood - Calculate Haversine Distance for each pair of restaurants

Number of Competitors	Categories Similarity	Reviews Similarity	Average Rating/Price Range	Sentiment of Reviews

Step 2: Explore restaurants within the identified neighborhood by different dimensions

Measuring Success

- Use **number of reviews** and **stars** to measure success
- **Review score**: measure the number of reviews of [0.25, 0.5, 0.75, 1] quantiles in different cities. Restaurants that have number of reviews in the first quantile have review score 1, those in the second quantile have review score 2, etc.
- **Rating score**: star scale is the same for all restaurants, so we directly use star to represent rating score
- **Success score** = review score * rating score

Competition vs. Success

- Project Objective: reveal the secrets for success
    - Hypothesis: does competition affect the success of a restaurant?
    - For each restaurant, we count the number of restaurants nearby
    - For each pair of restaurant, we compute their similarity in **categories,** and then take the average of this similarity index
- Two measures of similarity in categories
    - Overlap: whether there is at least one overlapping category
    - Levenshtein: a distance measure of the similarity of two lists
- Calculate the average **star rating**, **price range** and **number of reviews**
- Compare & Contrast
    - Homogeneity in category -- do successful restaurants enjoy the economics of agglomeration?
    - Will success lead to success ?
- Construct a linear model **(Cont’d)**
    - $$\begin{aligned}\text{success}_i = \beta_0 + \beta_1\text{restaurant\_nearby}_i + \beta_2\text{average\_rating}_i + \beta_3\text{average\_popularity}_i \\ + \beta_4\text{same\_category}_i + \beta_5\text{levenshtein}_i + u_i\end{aligned}$$

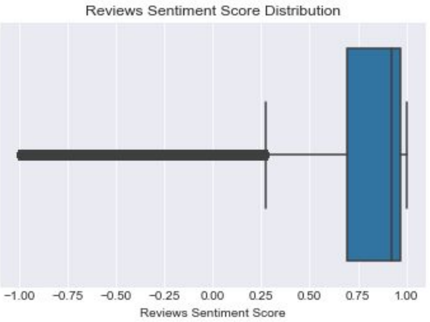
Reviews Textual Analysis

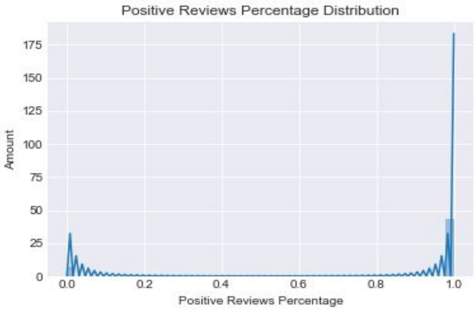
Sentiment Analysis -- using NLTK

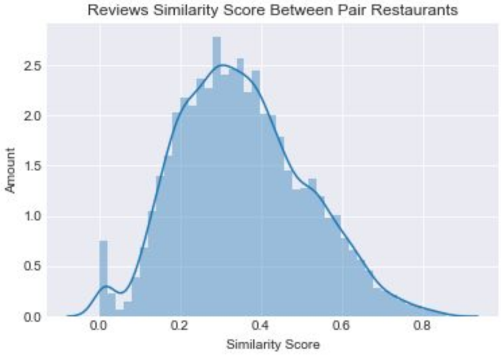
- Calculate the general sentiment score (-1 to 1) of each restaurant's reviews, and take the median as the sentiment score of the restaurant.
- Calculate the percentage of positive reviews (score > 0) of each restaurant.

Reviews Similarity -- using Bag of Words model

- Construct bag-of-words sparse matrix of each restaurant's reviews
- Calculate the reviews similarity score (cosine similarity) between each pair restaurant within the neighborhood. Take the average similarity as the reviews similarity of the neighborhood. This average similarity is a proxy of diversity.

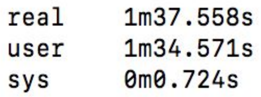






Big Data Methodology

- We use Mapreduce to do the computation
- Iterate the dataset to calculate the distance between each restaurant pair
    - After filtering our dataset, there are 54618 observations of restaurants
    - Form a set of restaurant pair, a combination of 54618 times calculation needs to be done
    - A single task, calculate the distance for a set of 1000 restaurants on a single machine takes



```
real	1m37.558s
user	1m34.571s
sys	0m0.724s
```

- Compute features of neighborhoods (e.g., reviews similarity, category similarity etc.) by iterating pairs of restaurants within the neighborhoods. For example, if there are 100 restaurants in the neighborhood, 100 * 99 / 2 = 4950 times of calculation should be done to calculate the similarity scores

Thank you for listening!