Data Science Project Stage-1

Group ID: 17
Group Members: Gautam Singh, Harsh Singhal, Naman Agrawal

Objective : Defining the Data Science Problem and Data Collection

Data Science Problem

We want to analyze the two structured data sets that contain information about various restaurants and also the restaurant review text documents to find out the influence of the various attributes on the rating of the restaurants.

We aim to gain insights on :-

- Impact of food quality on rating
- Impact of pricing on rating
- Correlation between food quality and pricing
- Food quality in restaurants providing home delivery
- Food quality in restaurants providing take out services
- Areas (zip code) where restaurants are likely to have better rating

Data Sources

We have collected data from two popular sites for restaurants review.

(1) Zomato.com
(2) Yelp.com

Data Extraction

Used scrapy web crawling library in python to crawl the web pages and extract the structured data in CSV format. Structured data has various attributes such as address, zip code, telephone number, delivery and take out services, outdoor seating and Wifi etc.

Text Documents

Text documents consist of restaurant reviews and will be used for sentiment analysis. We are interested in extracting following information from text files.

- Food quality
- Best cuisine served in a restaurant
- Variety of cuisines in a restaurant
- Customers view about pricing

Open Source tools

Scrapy

Scrapy is an open source web crawling framework for crawling websites and extracting data from websites. This tool is written in Python and currently maintained by Scrapinghub and many other contributors. Scrapy provides an interactive web crawling shell to test CSS and Xpath operations to scrape data, this shell is often useful to debug spiders during development.