CS 838 (Spring 2017): Data Science
Project Report - Stage 2 (Group 12)

Deepanker Aggarwal
deepanker@cs.wisc.edu

Saket Saurabh
ssaurabh@cs.wisc.edu

Vishnu Lokhande
lokhande@cs.wisc.edu

1 Objective

In this stage of the project, we do information extraction from text documents using supervised machine learning approaches. We have multiple text documents and the goal is to extract specific kind of entity from these text documents.

2 Procedure

We follow supervised learning approach to do entity extraction from the text documents. Since this is a supervised approach, we manually labeled the required entities in all the text documents. We divide the data-set into two sub-datasets, the first one being a training (or development) set and the second one is a test set.

3 Dataset description

Each text document pertained to a restaurant review in the New York City. Development set had 215 documents and test set has 100 documents.
The entity type which we extracted are the names of the dishes mentioned in the text document. Some examples of dish names is as follows ??. A summary of dataset description is given as follows.

Text documents type =	Restaurant reviews
Total number of text documents =	315
Total number of mentions =	??
Number of documents in development set, Set I =	215
Number of mentions in set I =	??
Number of documents in test set, Set J =	100
Number of mentions in set J =	??

In this project, we collect different kinds of data related to restaurants in New York City. The data science problem that we want to solve in this project is to answer how well does a Yelp rating of a New York restaurant correlates with various factors like demographics and city restaurant inspection results. The insights from this project can then be possibly used to predict where to open the next popular restaurant within New York City. Specifically, we want to answer the following data science questions:

- Is there a correlation between city restaurant inspection result and Yelp rating for a restaurant? That is, based on the inspection result can we predict how good or bad the restaurant rating would be?
- Based on the median household income for a given zip code within New York City, can we predict whether the restaurants within that zip code are more pricey/upscale or not?
- Can we predict where in New York City should one open a new restaurant so that it more profitable?

4 Set of data sources for our project:

4.1 Yelp Restaurant Data:

The Yelp restaurant dataset contains the information of different restaurant businesses acquired using the search api provided by Yelp Open Link. The dataset has been narrowed down to focus on only the restaurants in New York. For each restaurant in the dataset, several details are provided such as the restaurant's name, it's location, zip, rating, price, etc. The dataset contains 19,287 records. We use this structured dataset to do entity matching with NYC City Restaurant Inspection Dataset (below) based on the restaurant name.

4.2 NYC City Restaurant Inspection Dataset:

The New York City Restaurant Inspection Dataset contains inspection results for various NYC restaurants that are conducted on regular basis by New York City Department of Health and Mental Hygiene (DOHMH). The DOHMH records all the health-related violations at a given restaurant and assigns it an inspection score and a grade.
This dataset was collected from NYC Open Data website. Open Link
This dataset contains 232,385 records with each tuple having 18 attributes. We use this structured dataset to do entity matching with Yelp Dataset (above) based on the restaurant name.

4.3 NYC Demographics Dataset:

The NYC Demographics dataset was collected by web scraping the data from city-data.com, which records various demographic information for each zip code

within New York.
The link to the webpage itself can be found here. Open Link
This dataset contains data for 179 zip code areas within New York City. For each zip code, it records the population, population density, cost of living, median household income, and median real estate value for that zip code area.

4.4 NY Times Restaurant reviews:

The NY Times dataset has reviews of restaurants in New York. We extracted the data by scraping the New York Times website from here . This dataset has data for 673 restaurants. For each restaurant, this dataset has restaurant name, summary, website link to the review itself, review, address, telephone and cuisine.

5 How did we extract structured data from the data sources:

5.1 Yelp Restaurant data:

We extracted structured data directly from Yelp using their exposed API service. The Yelp API allows us to query information about restaurants based on the city name (New York in our case) and the city zip codes. Structured data can be retrieved in JSON format. The description of Yelp API is available here.

5.2 NYC City Restaurant Inspection Dataset:

This data was already available for download in structured CSV format. The dataset can be download from here.

5.3 NYC Demographics Dataset:

We obtained this dataset by manually web scraping the web page of city-data.com using Scrapy, a Python-based web scraper. Scrapy allows one to define a set of rules using CSS/XPATH selectors to extract certain fields from an HTML document. The extracted data can then be easily converted to a structured CSV format.

6 What is that we want to extract from the text documents:

In this project, we plan to extract various dishes that a restaurant provides from the review text. Based on names of dishes extracted from critics' reviews in New York Times, we want to predict what dishes a restaurant should sell in order to gain popularity within its locality.

7 Open source tools used to extract:

7.1 WebScraper.io Chrome Extension

Using this extension, we created a sitemap that specified how the website should be traversed and what all elements should be extracted from it. Using these sitemaps, the Web Scraper traverses the site and extracts the data which can be exported as CSV. Sitemap for our scraper can be found here. OpenLink

7.2 Scrapy

Scrapy is a fast high-level web crawling and web scraping framework in Python that is used to crawl websites and extract structured data from their pages.