CS 839 – Data Science Project – Stage 2

Team Members:

- Madan Raj Hari – mhari2@wisc.edu
- Raghavan Vellore Muneeswaran – velloremunee@wisc.edu
- Shadana Subramanian – ssubramani23@wisc.edu

Web Data Sources:

- **Zomato:** The first data source selected was **(www.zomato.com).** Zomato is one of the best Restaurant search and discovery application. Zomato gives information about nearby restaurants, menus, prices of each item, reviews and other information related to the restaurant.
- **Yelp:** The second data source selected was (**www.yelp.com**)**.** It is a local search service that develops, hosts and markets which publishes reviews about local businesses as well as the online reservation service. Yelp is great for finding a restaurant in the city. Yelp also provides information about the restaurants, people's rating, menus, location, opening and closing hours.

Extraction of Data:

- The structure and the format of the HTML pages of Zomato and Yelp were analyzed first. We decided to extract the entity **Restaurant** from the above two sources. The locations that were chosen includes **Madison, Phoenix, Atlanta, Minneapolis**. The attributes that we targeted includes **Name, Address, Price Range, PhoneNumber, Cuisine**. We used **Scrapy**, an open source web scraping tool.

Steps to extract the data:

- The starting URLs were hard coded to start the crawling.
- In the search results page of Zomato and Yelp filtered by restaurants, all relevant details to be extracted were analyzed. To ensure that similar data gets captured in both the sites the restaurants were ordered by ratings so that we crawl the top-rated restaurants.
- The data was captured by using Xpath and css selector to choose the required entities.
- The extracted information was stored in a CSV file.
- This process is repeated for a series of pages for a single location.
- The same steps were repeated for a different location.

Entity Selection:

- **Table A – Zomato (Zomato_data.csv):** The table from Zomato contains the extracted details about the restaurants – Name, Address, ZipCode, Price Range, Phone Number and Cuisine.
- **Table B – Yelp (Yelp_data.csv):** The table from Yelp contains the extracted details about the restaurants – Name, Address, Phone Number, Price Range and Cuisine.

To arrive at a common schema for both the sources following entities were used:

Attribute Name	Data Type	Description
Name	String	Name of the restaurant
Phone number	String	Phone number of the restaurant. It includes the area code and phone number of the format (XXX) – XXXX – XXXX. Blank value indicates missing of phone number.
Price Range	Integer	Price range of the restaurants. It ranges from 1 to 4 with 4 being the costliest.
Address	String	Address of the restaurant. It includes only the first line of the address (usually the block number and the street name).
Cuisine	String	Includes the list of all cuisine types that the restaurant provides. The value is expressed as a single string separated by a comma.

Table Information:

- For the purpose of this project, we consider the restaurants in Zomato and Yelp at specific locations – Madison (WI), Atlanta (GA) and Phoenix (AZ), Minneapolis (MN). The exact count of the number of records extracted from each of the location are shown below:

CSV File Name	Total number of Tuples
Zomato_data.csv	5644
Yelp_data.csv	3809

Open Source Tools Used:

- **Scrapy:** Scrapy is a web crawling framework in Python. It is a general-purpose web crawler and can be used for web scraping and extracting data using APIs. Using Scrapy, we define spiders for each web source. It provides the framework for processing the response for each request.

Codebase can be found here:

- **Zomato:**https://github.com/raghavan94/DataScience/blob/master/Stage2/Code/cs839/spiders/zomato_spider.py
- **Yelp:**https://github.com/raghavan94/DataScience/blob/master/Stage2/Code/cs839/spiders/yelp_spider.py
- **Zomato_data.csv:**https://github.com/raghavan94/DataScience/blob/master/Stage2/Data/Zomato_data.csv
- **Yelp_data.csv**https://github.com/raghavan94/DataScience/blob/master/Stage2/Data/Yelp_data.csv