Yelp Data Extraction and Exploration

1 Problem Description

Yelp is a local-search service publishing crowd-sourced reviews on businesses, which heavily influence customers' decision-making, and in turn affecting business performance. The goal of this project is to extract key data from the Yelp dataset and to perform data explorations based on practical needs. This dataset is aggregated and provided by Yelp for use in personal, educational, and academic purposes [3].

The dataset is interesting from a customer's perspective such as finding popular restaurants in the neighborhood and searching restaurants by categories. From a research perspective, the dataset can reveal more insights such as locating fraud users, whose reviews and ratings might be misleading. By extracting necessary data fields, we could query abnormal user ratings to deduce Yelp's rating algorithm on businesses. Another interesting query could be relating user reviews and tips to their genders. Through analysis over the dataset, we hope to provide more insightful information for both customers and researchers.

1.1 Data

The Yelp dataset consists of 6 JSON formatted files, each of which is composed of one JSON object per line [5]. The 6 JSON files contain information about business profiles, reviews, user profiles, check-in time, tips, and business photos.

Based on the name-value pair nature of JSON objects, fields of interest are easily located. For instance, the business data include locations and categories. Hence, it is feasible to find the number of businesses in certain categories within an area. Queries can also be performed over different data such as comparing the average rating based on the review data and the rating in the business data to see whether there are deviations. Because the task list and detailed assignments are not determined at the moment, the final fields of interest are not decided yet. Those fields will be finalized when the task list is determined.

The Yelp dataset is originated in the Yelp Dataset Challenge [4]. There have been 11 rounds of competitions in the past and a great many explorations have been performed over this dataset. Over the past winners, some team predicts potential customers using machine learning and natural language processing techniques [1]. Another team presents new techniques of recurrent neural networks over this dataset [2].

1.2 Database System

The data is loaded into a Postgres database, version 10.5.

Since the data is stored in JSON, it is necessary to convert the data into a format which is easily loaded into the database. Therefore, before performing data manipulations and queries over the dataset, extra work needs to done to load the data into the database.

2 Team

This work is completed by the following three people. Everyone is assigned an even amount of work.

Name	NetID	Department	Degree Program
Luoqi Wu	lw31	Computer Science	Bachelor
Yan Xu	yx28	Chemical and Biomolecular Engineering	PhD
Zheng You	zy24	Computer Science	Master

Reference

[1] Li, R. et al., “CORALS: Who are My Potential New Customers? Tapping into the Wisdom of Customers’ Decisions”, in: (2017).

[2] Ming, Y. et al., “Understanding hidden memories of recurrent neural networks”, in: *arXiv preprint arXiv:1710.10777* (2017).

[3] *Yelp Dataset*, URL: `https://www.yelp.com/dataset`.

[4] *Yelp Dataset Challenge*, URL: `https://www.yelp.com/dataset/challenge`.

[5] *Yelp Dataset Documentation*, URL: `https://www.yelp.com/dataset/documentation/main`.