Yelp Data Extraction and Exploration

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

1 High-Level Statement

The goal of this project is to extract data from the Yelp dataset and to perform data explorations. The overall project can be divided into five phases: data import, data comprehension, data cleaning and profiling, ERD diagram design, and data explorations. The following tasks are categorized according to the five phases.

2 Task Assignment

2.1 Data Import (Zheng You)

1. Conduct researches about loading JSON data into PostgreSQL.

   There are possibly multiple methods to load the data. One option is to convert the JSON data into CSV format and then load the CSV data into PostgreSQL. Another way is to operate JSON data directly with PostgreSQL because it supports native JSON data type since version 9.2 [2].

2. Import the JSON data into PostgreSQL to create relations based on a chosen method.

2.2 Data Comprehension (Yan Xu)

1. Identify a list of interesting questions from customer and business perspectives.

   There are generally three categories of the key questions:
   - identify popular restaurants in a particular neighborhood;
   - identify possibly fraudulent users;
   - understand how customer genders relate to reviews and tips.

   Note a key question might involve multiple sub-queries. For example, the first key question could be divided into the following sub-queries:
   - What are the top 10 cities in the country that have the most business of restaurants?
   - What are the restaurant categories in each city?
   - What are the top 10 restaurants in a category that have the most review_counts?
   - What is the star of the restaurants that are obtained in the last sub-query?

2. Understand relations and identify fields of interest.

   The fields of interest are chosen based on the key questions. For instance, attributes "categories", "city", "state", "business_id" from relation "business" are needed to answer the first key question.

2.3 Data Cleaning and Profiling (Luoqi Wu)

1. Clean the raw data.

   Before we dive into profiling the data and extracting derived relations, we need to perform some clean-up. This process is mainly based on the understanding we have developed on the dataset. For example, suppose we identify the "text" field in the "yelp_academic_dataset_review.json" file as relevant, and also want to limit its character length within 5000, the Yelp max character length, then we could clean up the file by adding a column "text_len" that contains the length of the review body text. Based on this column we can easily filter out reviews that do not fit our standard. Similar procedures will be performed for other fields of interest.

2. Profile the data.

   We will mainly perform three types of data profiling: single column profiling, cross-column analysis, and inter-table analysis. Single column profiling will focus on improving understanding of the frequency distribution of different values, types, and use of each column (e.g. average length of Yelp reviews). Embedded value dependencies can be exposed in a cross-columns analysis (e.g. relationship between a business's average rating and its location). Finally, overlapping value sets possibly representing foreign key relationships between entities can be explored in an inter-table analysis (e.g. the set of businesses that a user has reviewed) [1].

2.4 ERD Diagram Design (All team members)

1. Identify the relations and relationships.

   The relations consist of the raw data imported and derived relations. Referential integrity will be enforced with foreign keys.

2. Draw the ERD diagram based on defined relations and relationships.

2.5 Data Explorations (All team members)

Explore the data based on listed key questions. Note the key questions are tentative and there might be more aspects of the data we would like to explore. Every team member will be assigned some key questions to answer.

We believe the work is balanced among team members and is distributed by each one's advantages. Zheng and Luoqi are major in Computer Science who have relatively solid programming backgrounds. Hence the data import and processing are assigned to them. Yan is a PhD candidate with strong analytical ability who will bring insights to the data. In addition, all team members will participate in the ERD diagram design. The final data exploration tasks will be distributed evenly through key questions.

Reference

[1] *Data Profiling*, URL: `https://en.wikipedia.org/wiki/Data_profiling`.

[2] *PostgreSQL JSON*, URL: `http://www.postgresqltutorial.com/postgresql-json/`.