Proposal for Data Mining Project

Ian Blanchard
Department of Computer Science
Kennesaw State University
Marietta, United States
iblanch1@students.kennesaw.edu

***Abstract*—This document proposes that a dataset of 1.48 million Google Restaurant reviews be the subject of interest in this semester's Data Mining project. This paper further proposes that multiple data mining techniques be used to discover patterns relating (1) a person's review scores to the quantity of reviews that person has given (2) the quantity/quality of reviews to the geolocation of the business_id and (3) the type of restaurant to the number of pictures included in the pics attribute.**

I. Introduction

In today's world we have a lot of data that has been generated over the years. For example, Google Maps provides us with the location of almost every business, street, and house in the U.S. In addition, it also contains a catalog of reviews for each business. In the pursuit of equipping myself with a greater understanding of what Data Mining is, how it works, and how it is useful, I am using a dataset related to these reviews to find useful and interesting patterns.

Through analyzing this dataset, I shall gain a better understanding of how the process works. In addition, I shall uncover interesting patterns, allowing myself a better ability to detect them. From manually looking at the review text alone, I have been able to identify the location and name of the businesses included, which helps prove how useful the dataset can be.

II. Dataset Description

The dataset I have chosen is called "Google Restaurants" and it is sourced from *Personalized Showcases: Generating Multi-Modal Explanations for Recommendations*. It contains 1,487,747 Google Reviews of 64,527 different restaurants and is just over one Gigabyte in size. It contains five attributes, the *business_id*, the *user_id*, the *rating* (which is from 1-5), the *review_text*, and *pics*, which includes the id and link to each picture included in the review. I shall be using every attribute except for review_text (which I would only use as a last resort). There appear to be no issues with the dataset. Everything is consistently formatted: apostrophes are replaced with `\u2019`, each section is clearly separated with a comma, and every attribute is neatly enclosed in quotation marks except for the rating. On its own, this is a fantastic dataset to start with. However, the *business_id* is useless on its own. In order for me to properly utilize it, I will have to also purchase a Google API allowing to cross-reference the *business_id* with data associated with it, such as Geolocation or category of restaurant. For each business that is no longer functioning or has missing data on the API end, my journey becomes even more difficult

Below I have included a sample of the *review_text* and *pics* attributes, below that is a table of the other attributes in the first five reviews.

review: "We came for a birthday brunch and this place is so much bigger than it looks from the outside! It was totally packed and loud. Service was on the slower side. I ordered 2 mojitos: 1 lime and 1 mango. The ingredient weren\u2019t really fresh, there wasn\u2019t even any visible mango in the second one. Tasted like mango juice from concentrate. My food was really good though. I ordered the steak and eggs and I usually order my steak rare but this was skirt steak so they did a perfect medium. The sunny side up eggs were more cooked than I would have preferred but still good. They actually cooked the breakfast potatoes too, which was awesome. Will likely be back to try something else!"

pics:[{"id":"AF1QipPrls2G30PS3tyC55KBxUrKgy3ER0AB5UJY57BZ","url":["https://lh5.googleusercontent.com/p/AF1QipPrls2G30PS3tyC55KBxUrKgy3ER0AB5UJY57BZ=w150-h150-k-no-p"]},{"id":"AF1QipOBdiu4hEPC4538Q9iRzl1gAyNuMxThTVRKubSX","url":["https://lh5.googleusercontent.com/p/AF1QipOBdiu4hEPC4538Q9iRzl1gAyNuMxThTVRKubSX=w150-h150-k-no-p"]}]

TABLE I
Sample of Google Maps Reviews Dataset

Business ID	User ID	Rating
"605730f68cd0e3d69a52284b"	"113890892872599852766"	4
"605730f68cd0e3d69a52284b"	"100158382312395993536"	5
"605730f68cd0e3d69a52284b"	"113495161718980109602"	2
"605730f68cd0e3d69a52284b"	"111259544401075262963"	3
"605730f68cd0e3d69a52284b"	"104285463275063919410"	5

III. Discovery Questions

A. Is there a pattern linking the review scores and the quantity of reviews that a person has given?

This question is interesting because it could point out

- Fake Reviews
- Reviews that are only made because the restaurant is especially good or bad
- People who review every restaurant they visit

Since our reviews are grouped by restaurant and we only have approximately 1.49 million reviews, we will likely not see all of an individual user's reviews. In addition, restaurants appear to not be grouped by location, as the first restaurant

listed, according to one review, is "One of my favorite place in Orlando" but the second restaurant is described as the "Best New York bagels west of the Mississippi." Another review I saw claimed that the restaurant was located in Denver, so the restaurants aren't restricted to one region of the U.S. From what I've seen, all restaurants are located in the U.S.

B. Do restaurants in urban cities have a different quantity/quality of reviews than other areas?

This will be difficult because I have to use Google's API (probably in addition to another) to get both the Geo-location of the restaurant from the *business_id* and the city if the Geo-location is located within another location's city limits.

C. What ratings at what kinds of restaurants tend to have more pictures posted in their reviews?

This requires me to use Google's API to relate *business_id* to the category of restaurant and relate that to pics and rating. Considering that we have many different kinds of restaurants listed here, it will be the most difficult to discovery question to answer.

IV. Planned techniques

For the first discovery question I shall use anomaly detection and K-means. Anomaly detection should help for detecting unusual/interesting cases for the number of reviews. K-means is a technique in the clustering category that is chosen due to the size of the dataset.

The second discovery question is similar enough to the first that k-means and anomaly detection will likely be used. In addition, I'll also be using Decision trees to garner further information.

For the third discovery question, association rules apply more because the tags related to the restaurant may be associated to different places. I shall use apriori over FP-Growth due to it being more thorough. In addition, I shall use K-means again.

V. Preliminary Timeline

A. Further preparations

I am officially done with this section, but I will double-check with the professor on what techniques I should use. In addition, I should use SQLite or short Python programs to learn more niche things about the dataset, such as "which user within the dataset has submitted the most reviews?"

B. M2

After this, I shall read the additional class material and ask the professor at the end of every class session to build an initial implementation as soon as possible. The ideal period to pursue this goal is the middle of this month, because I have few assignments due at the time. I am certain of what techniques I need to use, I am still unsure of what context that shall be in. As stated previously, the best way I can be confident in my path is to read relevant material and ask for help.

C. M3

The most opportune time to finish the complete implementation of the project is early in March, after turning in M2. If not, then I have a 5-day window from March 27th to April 1st to work on the project without having other assignments or exams to worry about. I'll have to once again ask about how far apart an incomplete implementation and a complete implementation is.

D. M4

The final deliverable is mentioned in the Syllabus and not in D2L, but I know that if it exists, it is due around May 3, which is after finals. There is also a Presentation due on April 30. Most of the middle of April is a good time to work on this stage of the project, if there is more work to be done. I am excited for the presentation since that does mean I get an excuse to dress up and discuss the project. I'll work on it after I finish the final deliverable.

VI. Acknowledgment

I know that there are 64,527 different restaurants because I made a Java program that scans line-by-line.

References

[1] Google Restaurants Personalized Showcases: Generating Multi-Modal Explanations for Recommendations An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, Julian Mcauley arXiv:2207.00422, 2022 pdf