MAST30034 Applied Data Science

Final Project Proposal

Build a recommender system for users to live better with Yelp

Team Members

- Jiachen Zhou 901091
- Xintong Yao 991477
- Yiran Wang 987751
- Ruoyu Wu 947132

Introduction

Yelp is an online directory for users to discover local businesses by searching for and reviewing businesses around them (Trefis Team. 2019). It provides a platform for both users and businesses to gain advantage from online information exchange. Figure 1 shows a wrapper for the business categories and functionalities. In the last year, Yelp allowed users to filter business based on their personalized preferences. Therefore, only places that meet their individual requirements will be surfaced in search results (Souther, 2019). This new feature will help match up consumers and businesses to increase the number of loyal customers while improving user experience.

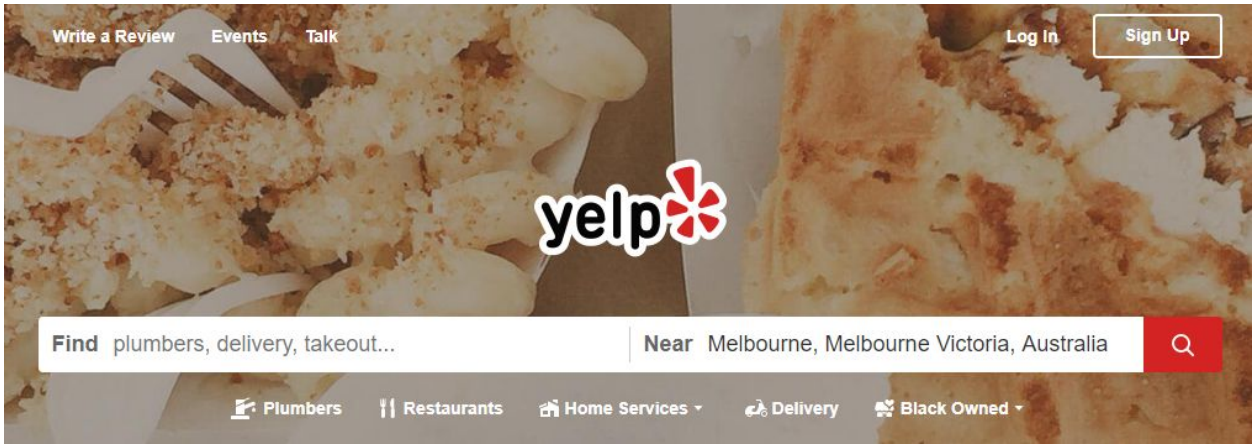

Figure 1: A screenshot of Yelp official website (https://www.yelp.com/)

In this project, a recommender system will be built to provide users with more precise suggestions on suitable businesses. The current settings in the Yelp website are customized, where a list of locations, time and food categories are available for user selection. However, it is still time-consuming to look through the long reviews of each business in order to find the

desired ones. Moreover, the current preference settings of Yelp fail to help indecisive customers that have trouble deciding on the initial category choices. However, the system that will be built in this project can hopefully inform customers about their happiness index on different choices, then recommend business accordingly.

This system proposed in the project aims to rank different businesses based on specific user preferences and habits and present individualized recommendations to each user. This could be achieved by analyzing the combination of the reviews and stars left by previous customers. Other users' reactions (attributes like "compliment_cool", "compliment_funny" etc.) to these existing reviews can also be studied to determine the usefulness of reviews, hence to give better user recommendations.

Dataset

Yelp Dataset: A trove of reviews, businesses, users, tips, and check-in data! (Yelp, Inc. 2020)

The structure of this dataset:

- Dataset_Agreement.pdf
- yelp_academic_dataset_business.json
- yelp_academic_dataset_checkin.json
- yelp_academic_dataset_review.json
- yelp_academic_dataset_tip.json
- yelp_academic_dataset_user.json

This high dimensional 10 GB dataset has approximately 8 million instances and more than 50 columns, including primary keys and foreign keys. Tasks like classification, regression and NLP etc. can be performed on this dataset flexibly. Although images are not included, text processing is just as intricate and sexy since NLP and recommender systems play an important role in the real-world application. The temporal structure is clear in this dataset as all the reviews even businesses have a Datetime attribute to record the action.

Data cleaning in NLP tasks is also challenging and interesting. Reviews/texts can be looked at to determine which word, sentences or even punctuation symbol to be removed to reduce noises. After a simple review of the dataset, there is no missing data. However, the innate complexity of human emotions often cannot be expressed thoroughly in the review despite its long paragraph, suggesting that any sentence could possibly be interpreted as missing data or even censoring and extremes, which is the charming part of NLP. Hence, this is an overall complex dataset that is challenging in all aspects including processing, model-building and the team is excited to take on the project.

Figure 2 shows the structure of the chosen dataset with the relationship of primary keys and foreign keys added. It can be seen that this dataset complies with the Tidy structure along with high dimensional structure, which tends to be a good dataset for analysis.

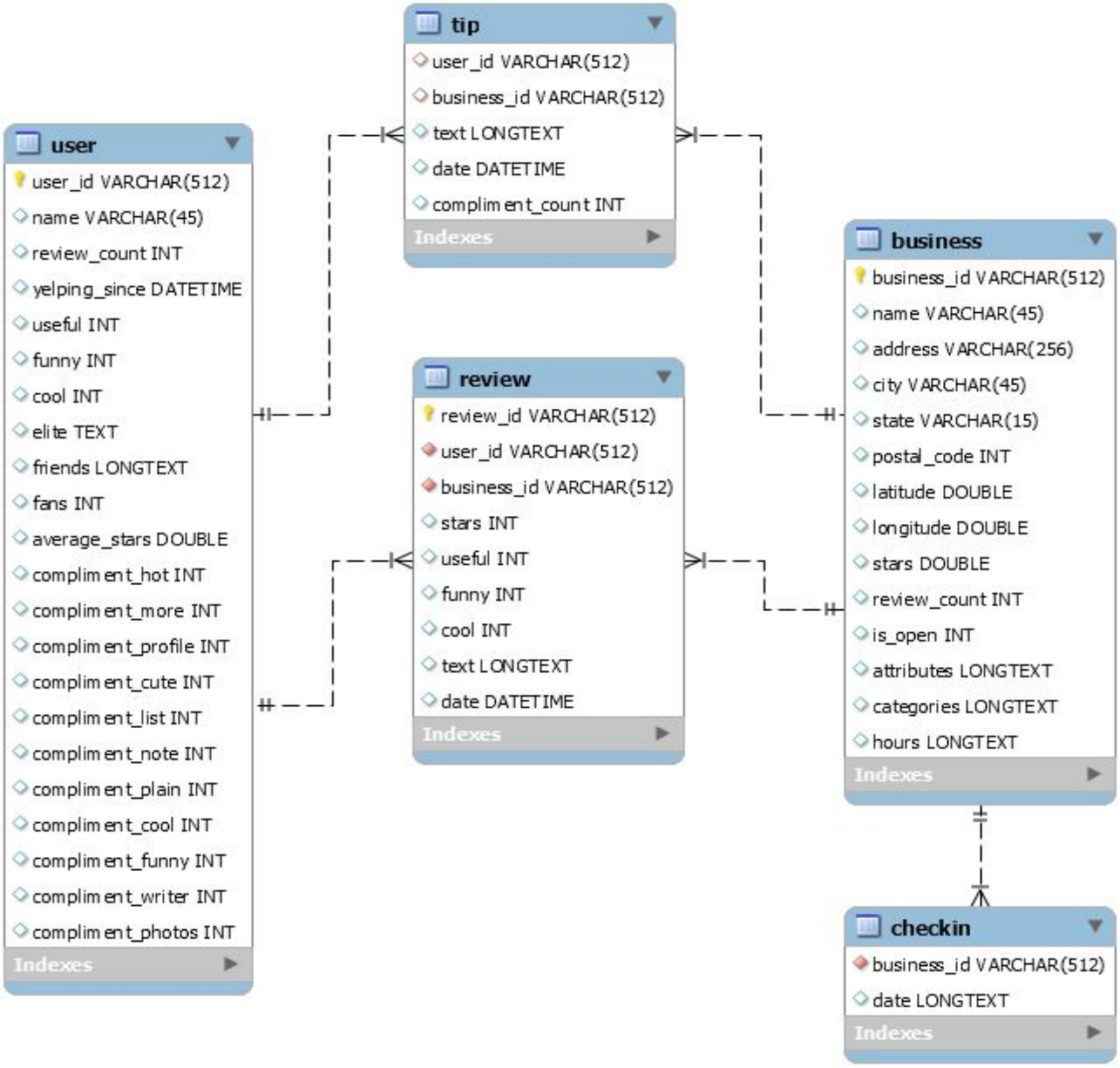

Figure 2: A SQL EER diagram for Yelp dataset

Tasks

- Build prediction models using some latest techniques such as Bert and compare performance with older ones like LSTM.
- Choose well-performed models to build a recommender system for users to live better with Yelp.

Important Methods

- Word embedding by Tensorflow **Tokenizer**
  A tokenizer can convert input text to numerical representations or convert sentences to a list of vocabulary indices from strings. The advantage of Tensorflow Tokenizer is that it is already documented so that it is unnecessary to find a vocabulary dictionary somewhere else.
- Text cleaning by **Spacy** and **Nltk**
  As two support methods, Spacy and Nltk are good at cleaning text such as removing stopwords. A sentence can also be converted to a token to delete some useless words or characters like punctuation, names etc.
- **Bidirectional Encoder Representations from Transformers (Bert)**
  One of the latest compatible NLP models for pre-training. At the end of 2019, Bert had dominated the language search area for Google (Roger 2019). Also, TensorFlow 2.3.0 provides the usage of Bert which makes it easy to handle. The bidirectional encoder technology may increase both the speed of fitting and the accuracy of predicting.
- **Recurrent Neural Network (RNN)**
  An RNN is a subclass of ANN which is closer to NLP. Long short-term memory (LSTM) networks are considered to be used in this project as it contains memory units for models to “remember” what has been seen which is better than the simple reinforcement learning algorithm in basic RNNs.
- **Random Forest (RF)**
  It is a supervised learning algorithm operated by constructing multiple decision trees and choosing the mode of classes or mean prediction. Being able to handle both classification and regression tasks as well as large datasets with higher dimensionality with good output results are the main advantages, which are very suitable for this project. However, as the number of trees increases, it would be ineffective due to longer prediction run-time. Although its ability in predicting is outstanding, the description of the relationship between attributes is poor (Donges 2019).
- **Stochastic Gradient Descent (SGD)**
  SGD is often used as an optimization method for machine learning algorithms. It is a simple yet efficient way to fit linear classifiers or regressors such as SVM. It is also suitable for large-scale NLP and text-classification problems. However, the optimization steps tend to be noisy due to the frequent updates as well as a potential increase in the computational expensiveness.

Reference

Trefis Team. (2019). *How Will Yelp Look In the Next 2 Years?* Forbes. https://www.forbes.com/sites/greatspeculations/2020/12/30/how-will-yelp-look-in-the-next-2-years/#5d6af4366641

Yelp, Inc. (2020). *Yelp official website*. https://www.yelp.com/

Yelp, Inc. (2020). *Yelp Open Dataset*. - 2020 [Data set]. Kaggle. https://www.kaggle.com/yelp-dataset/yelp-dataset

Southern. (2019). *Yelp Introduces Personalized Search Results*. Search Engine Journal. https://www.searchenginejournal.com/yelp-introduces-personalized-search-results/322727/#close

Donges. (2019). *A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM*. Builtin. https://builtin.com/data-science/random-forest-algorithm

Roger. (2019). *Google's BERT Rolls Out Worldwide*. Search Engine Journal. https://www.searchenginejournal.com/google-bert-rolls-out-worldwide/339359/