Yelp Restaurant Reviews and Ratings Analysis

Project proposal

CSDA1050 – Advanced Analytics Capstone

By Steven Too Heng Kwee

ID: 304449

Introduction

Yelp is an American multinational corporation founded in 2004 which aimed at helping people locate local business based on social networking functionally and reviews. Yelp also has a star rating system that lets users easily see what the general opinion about a particular establishment is without having to read all the reviews for that particular business.

Millions of people use Yelp restaurant reviews in their food choice decision-making. Empirical data research demonstrated that a one-star increase led to 59% increase in revenue of independent restaurants (Lucas, 2011).

Therefore, we see great potential of Yelp dataset as a valuable insights repository either for customers in their food hunting quest or for businesses to optimize their operations and align themselves with their market.

Research Question

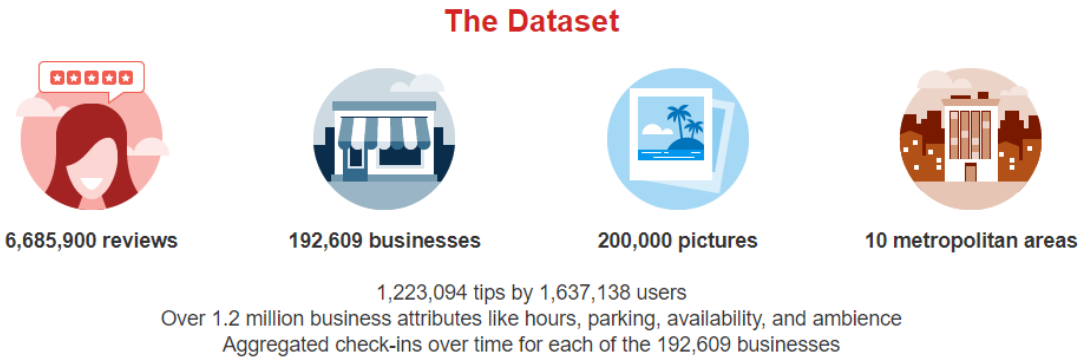
Our research will be 2 parts:

- Exploring the restaurant scene through the reviews. Identifying the positive and negative attributes
- Analysing the reviews to the star rating. Can we detect the inconsistencies and provide a more accurate rating?

Literature Review/Background

Data and Description

The Yelp dataset is available at https://www.yelp.ca/dataset.



We downloaded a 5.6 GB TAR file. This TAR file contained second TAR file that we extracted to get a series of JSON files: business, checkin, photos, review, tip, and user. Total real size is 8.05 GB.

We will be focusing on the following files:

Business.json: 131 Mb. Contains business data including location data, attributes, categories and average star rating

Review.json: 4.97 Gb. Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

The files will be read directly into a dataframe or converted to csv.

Due to the size of the files, local computer processing might be an issue. If it happens, we will consider the following alternatives:

- Extracting a subset
- Looking for a smaller dataset
- Using a Yelp fusion API extraction
- Cloud storage

Proposed Methodology

- Data will be restricted to Ontario restaurants or just Toronto
- Data will be process as is on local PC. Should there be limitations, consideration will be made to convert into SQL in order to partition the dataset.
- Python Notebook will be used for codebase and analytics
- Possible textual data clean that might be required: lower text, tokenize, remove punctuation and stop words, lemmatize.

- Natural Language Processing (NLP) techniques that consists in extracting emotions related to some raw texts will be mainly used for exploration and insights. The aim is to perform Sentiment Analysis to determine word/feature drivers(Word Cloud), segmenting opinions and setting polarity score or check at NLTK
- For the rating analysis and prediction, we explore several machine learning methods including logistic regression, Naive Bayes, Neural Network, and Support Vector Machine (SVM) to make relevant predictions. While logistic regression performs better than the others, predictions from all the methods are far from perfect.

Project deliverables timeline:

- Project Proposal – July 15, 2019
- Sprint #1 – Data Collection and exploration – July 29, 2019
- Sprint # 2 – codebase, report (brief), analysis plan - August 12, 2019
- Presentation review – August 20, 2019
- Final Project Submission – Final report, GitHub Repo, codes/analysis/results - August 27, 2019

References

https://github.com/Yelp/dataset-examples

https://www.geeksforgeeks.org/python-nlp-analysis-of-restaurant-reviews/