

Mining Yelp reviews

Authors: Aditya Priyadarshi, Abhay Kasturia, Xingxing Liu, Varun Nandu and Gautam Vashisht

Problem Description: The objective of this project to analyze yelp reviews dataset [1] and build a recommendation system to suggest businesses to users based on their earlier reviews.

Summary of Data: The original dataset described in the Yelp Dataset Challenge 10 [1] has 4.7M reviews and 1M tips by 1.1M users for 156K businesses spread across 12 cities. The data is given in json format which include business.json, review.json, user.json, checkin.json and tip.json. Each business has name, address, star rating and textual reviews. Each individual review data consists of anonymized IDs for the business, user and review, star rating, review type, review text and votes on how useful, funny or cool the review is.

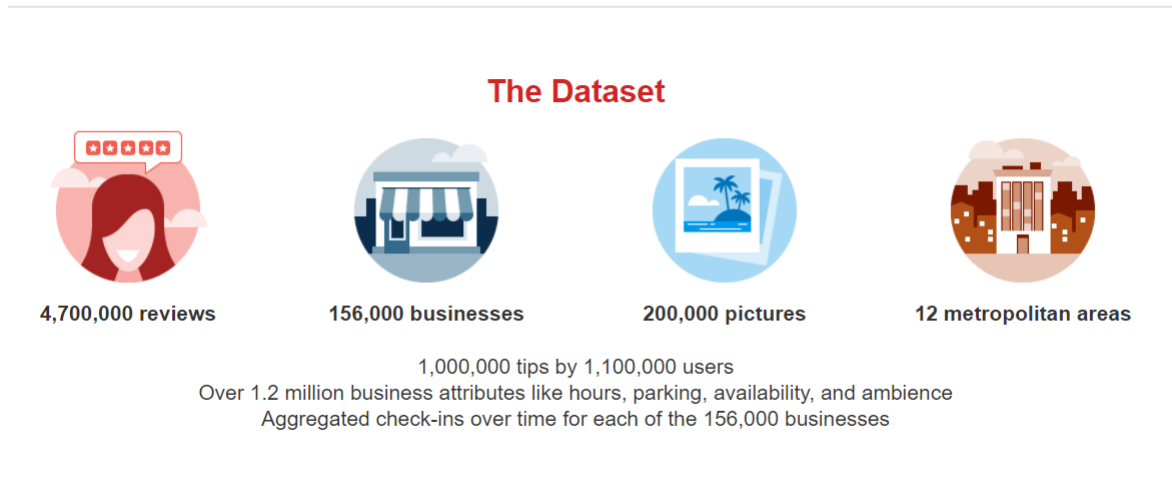


Figure 1: Dataset Details

Methods: We plan to use three different techniques for building a recommendation system and compare their performance on yelp dataset. The methods which we intend to use for building the model are Collaborative filtering[2], Clustering and deep learning [3].

Questions: Some initial questions which we plan to address. We may add more questions in future.

- Suggest a restaurant based to user based on his earlier review.
- Find most popular business in a city in different business categories.
- Find which type of cuisine are more related to each other based on user rating.

Division of Work: We initially want to work together on pre-processing and basis analysis of the dataset. Then each of us want to pick a technique and work on building a model using that method.

References

- [1] Yelp Dataset Challenge https://www.yelp.com/dataset_challenge
- [2] Collaborative filtering https://en.wikipedia.org/wiki/Collaborative_filtering
- [3] Paul Covington, Jay Adams, Emre Sargin *Deep Neural Networks for YouTube Recommendations*, ACM 2016.