Northeastern University

Data Mining Project Report (Draft)

Recommendation Systems for Yelp Dataset

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

Supervised by
Dr. Nate Derbinsky

November 22, 2017

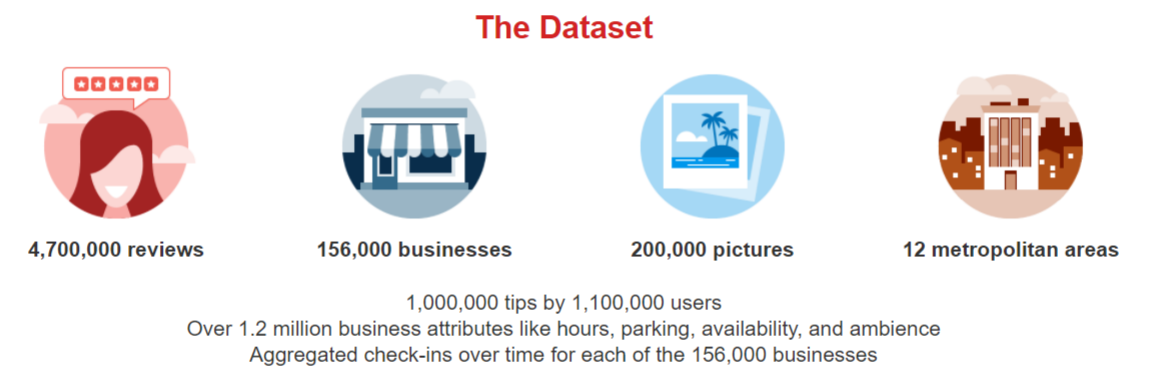
1 Introduction and Related Work

A recommender system or a recommendation system is a subclass of information filtering systems that seeks to predict the rating or preference that a user would give to an item [2]. Recommendations systems have become very relevant today given the presence of e-commerce website like Amazon and Netflix as well as other platforms like Facebook and Youtube. These are utilized in a variety of areas such as movies, music, videos, news, books, research articles, search queries and products in case of Amazon. Two most common methods to build a recommendation system are collaborative filtering and content-based filtering. Collaborative filtering methods use user's past behaviors and behaviors of similar users to find items which a user might like. Content-based methods use the features of the items liked by the user to suggest similar items. There are also hybrid recommendation system which combine both of these techniques.

2 Dataset and Analysis

2.1 Dataset

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.



The Dataset

4,700,000 reviews

156,000 businesses

200,000 pictures

12 metropolitan areas

1,000,000 tips by 1,100,000 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 156,000 businesses

Figure 1: Dataset Details

2.2 Analysis

We did an initial analysis of the dataset. Below sections present our analysis.

2.2.1 User data

There are 1183362 total users whose reviews are present in the dataset. We plotted a histogram to understand the distribution of user reviews. Looking at the histogram, we can observe that most of the user have very few reviews and some top users have significant number of reviews. Majority of user have 25 or less reviews which is also shown by a mean of 23..72 and standard deviation of 80.5. The maximum number of reviews given by any user is 11656.

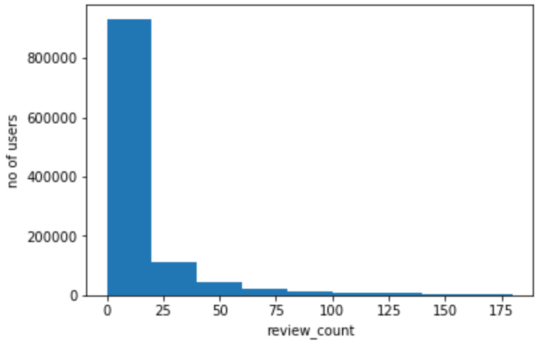


Figure 2: Review count per user

In addition to number of reviews, we also looked at distribution of star ratings given by a user. Looking at the histogram, we can observe that more users give higher rating which is shown by a median of 3.89 star rating. Mean and standard deviation for the same are 3.71 and 1.10 respectively. In order to group the reviews as positive, average and negative reviews, we have used the following method. We assume that if the rating lies in the range of (mean standard deviation, mean) which is 2.6 to 3.7, we will categorize it as average. Reviews lower than 2.6 will be considered as a negative review and anything greater than 3.7 will be considered as positive reviews with two extremes being 0 and 1.

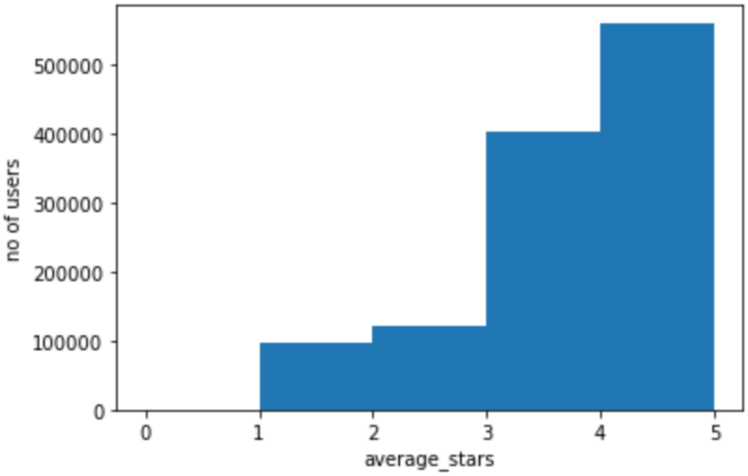


Figure 3: Rating per user

We also did some analysis to see the user growth on yelp. User growth has started declining after an increase in users joining from 2005 to 2014 .

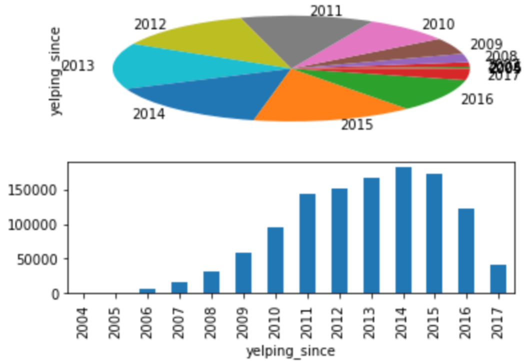


Figure 4: User Growth

2.2.2 Business Data

There are total 156639 business in the dataset. We grouped business according to city and business category to determine popular cities and categories. Below pie-charts give idea about popular cities and categories.

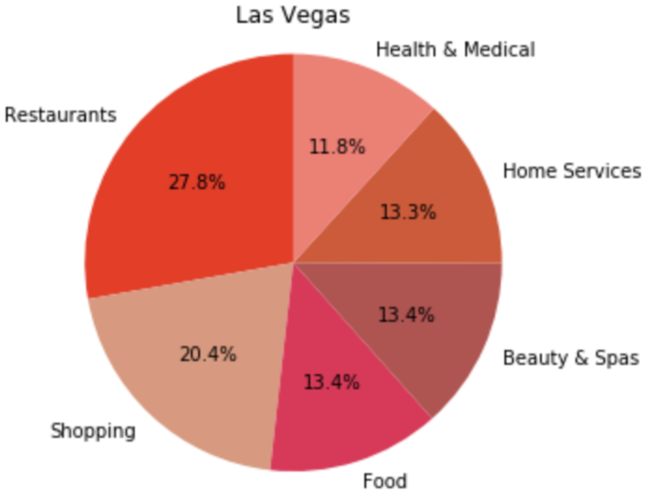


Figure 5: Top cities

We did more analysis into sub-categories of our most popular category i.e. resturants to map its distribution.

2.2.3 Checkin Data

Finally, we did analysis on use checkin data to find out popular timing in the top cities shown in our earlier analysis.

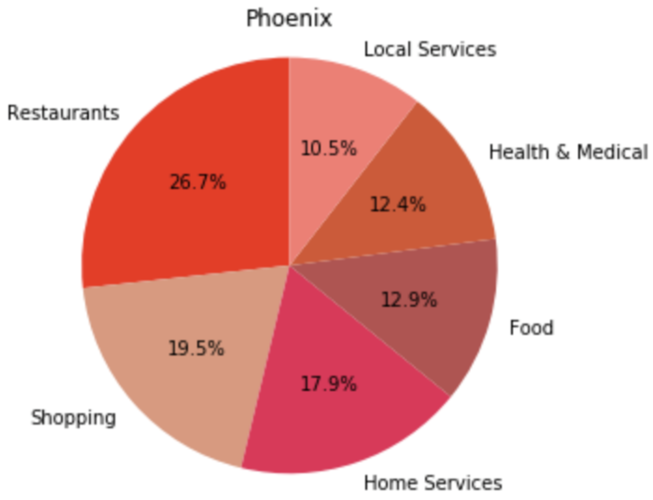


Figure 6: Top Business Categories

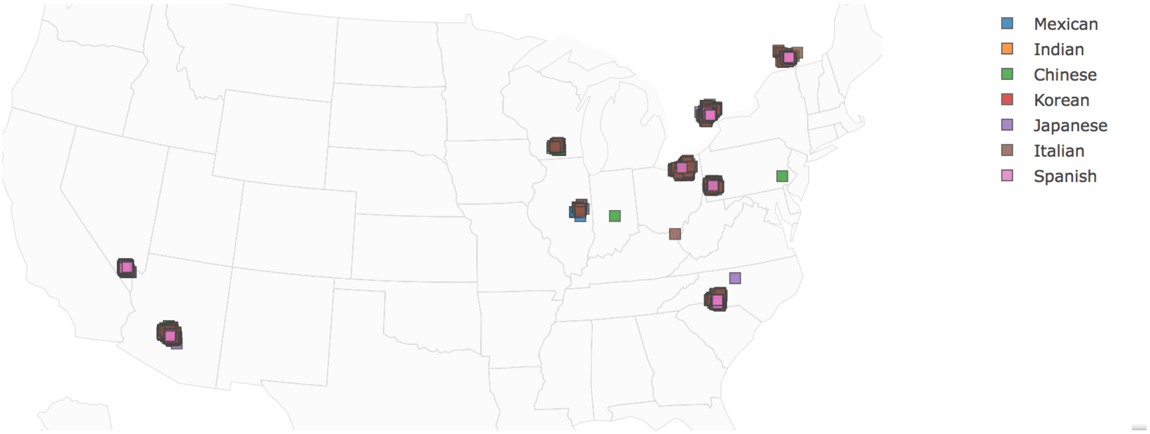


Figure 7: Resturant Sub-categories

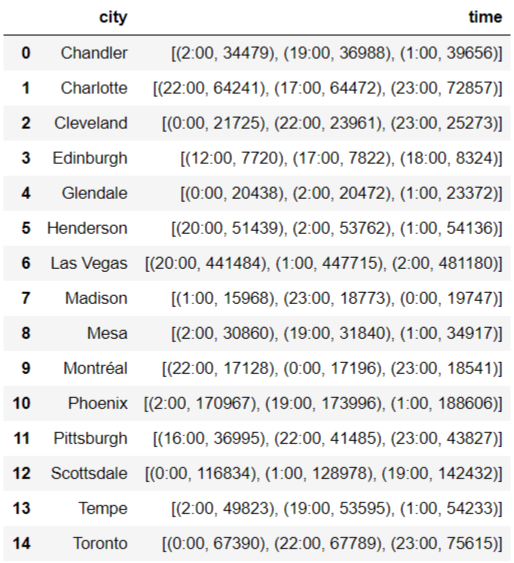
3 Methods

3.1 Clustering Based Approach

3.2 Collaborative Filtering

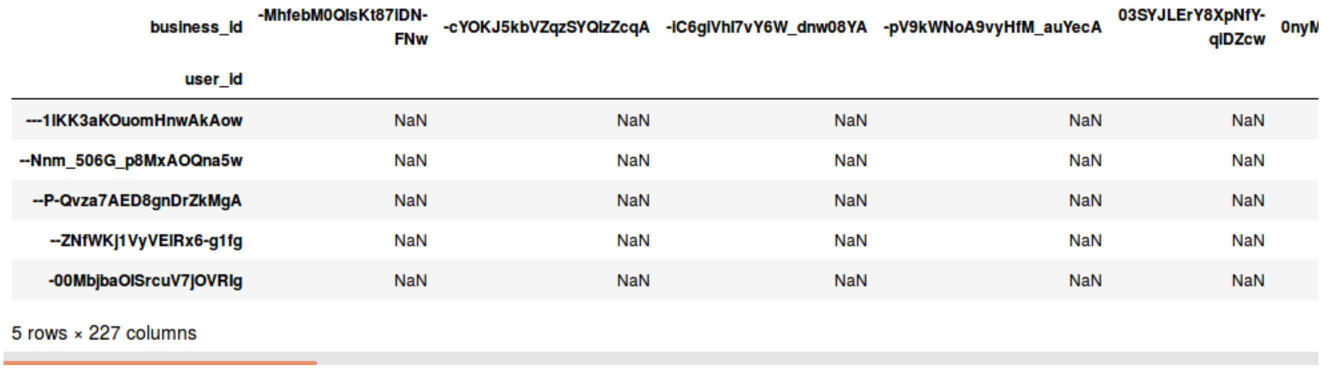
Collaborative Filtering is the method to implement recommedation system. It is the way to recommend item to user 'u1' by collaborating the choice of item of other users of similar type as 'u1'. We started with creating a matrix for users as rows and business as columns. Values in matrix contains the number of stars given by user to the particular business. Each row in matrix represent the vector of star rating given by user for all the businesses.

There are lot of missing values in matrix(for items which user has not given rating) and also the there is an issue of handling the rating given by soft users and hard users. Some users may rate the business they like with average star ratings or some may rate the business



	city	time
0	Chandler	[(2:00, 34479), (19:00, 36988), (1:00, 39656)]
1	Charlotte	[(22:00, 64241), (17:00, 64472), (23:00, 72857)]
2	Cleveland	[(0:00, 21725), (22:00, 23961), (23:00, 25273)]
3	Edinburgh	[(12:00, 7720), (17:00, 7822), (18:00, 8324)]
4	Glendale	[(0:00, 20438), (2:00, 20472), (1:00, 23372)]
5	Henderson	[(20:00, 51439), (2:00, 53762), (1:00, 54136)]
6	Las Vegas	[(20:00, 441484), (1:00, 447715), (2:00, 481180)]
7	Madison	[(1:00, 15968), (23:00, 18773), (0:00, 19747)]
8	Mesa	[(2:00, 30860), (19:00, 31840), (1:00, 34917)]
9	Montréal	[(22:00, 17128), (0:00, 17196), (23:00, 18541)]
10	Phoenix	[(2:00, 170967), (19:00, 173996), (1:00, 188606)]
11	Pittsburgh	[(16:00, 36995), (22:00, 41485), (23:00, 43827)]
12	Scottsdale	[(0:00, 116834), (1:00, 128978), (19:00, 142432)]
13	Tempe	[(2:00, 49823), (19:00, 53595), (1:00, 54233)]
14	Toronto	[(0:00, 67390), (22:00, 67789), (23:00, 75615)]

Figure 8: Checkin Times



business_id	-MhfebM0QIsKt87IDN-FNw	-cYOKJ5kbVZqzSYQIzZcqA	-IC6gIVhI7vY6W_dnw08YA	-pV9kWNoA9vyHfM_auYecA	03SYJLErY8XpNfY-qIDZcw	0nyM
user_id						
---1lKK3aKOuomHnwAkAow	NaN	NaN	NaN	NaN	NaN	
--Nnm_506G_p8MxAOQna5w	NaN	NaN	NaN	NaN	NaN	
--P-Qvza7AED8gnDrZkMgA	NaN	NaN	NaN	NaN	NaN	
--ZNfWKj1VyVEIRx6-g1fg	NaN	NaN	NaN	NaN	NaN	
-00MbjbaOISrcuV7jOVRIg	NaN	NaN	NaN	NaN	NaN	

5 rows × 227 columns

Figure 9: UserVsBusiness - Stars Rating Value

they don't like with good star ratings or so. We have used centered cosine similarity to group the similar users to handle the above mentioned issues. For centered cosine, we normalize the ratings by substracting row mean for each user and missing values are treated as zero average value.

To recommend the businesses to user 'u', we are finding the cosine similarity between all other users and 'u'. Businesses rated with positive average star rating, by users of having cosine similarity greater than 1, will be recommendeded to user 'u'.

3.3 Collaborative Deep Learing

4 Experiments and Results

4.1 User-User Collaborative Filtering

For the initial setup, we have worked on 100,000 rows of reviews.json file. We have created a matrix of 78276 users and 4224 businesses.We randomly choose one user to find the set of similar users of count 200. Based on positive average star ratings given by 200 users, 531 set of businesses are recommended to users.

5 Future work

References

[1] Yelp Dataset Challenge `https://www.yelp.com/dataset_challenge`

[2] Recommendation System Wiki `https://en.wikipedia.org/wiki/Recommender_system`

[3] Collaborative filtering `https://en.wikipedia.org/wiki/Collaborative_filtering`

[4] Collaborative filtering Techniques `https://www.youtube.com/watch?v=h9gpufJFF-0`

[5] Movie Recommenation System `https://beckernick.github.io/matrix-factorization-recommender/`

[6] Paul Covington, Jay Adams, Emre Sargin *Deep Neural Networks for YouTube Recommendations*, ACM 2016.

[7] Hao Wang, Naiyan Wang, Dit-Yan Yeung*Collaborative Deep Learning for Recommender Systems*, ACM 2015