Recommending Restaurants for Female Students

Chang Liu

November 17, 2015

Introduction

With the increasing capability of collecting data for online user behavior and customers' preferences, analyzing big data for commercial use to target potential customers and provide personalized advertising and recommmendations have become a major challenge. "Recommendation system"" has become a buzzword since then. We are especially interested in knowing which products are to recommend to which group of people in order to grow company's revenue. In this document, we are going to explore a dataset to answer the following question.

Question Which restaurants in Arizona are recommended to female graduate students?

Why This question is of interest since it provides a framework for recommending services for a special group of people. In this case, the people of interest are females as well as graduate students. It is not only to answer this question but also to offer insights into how to explore the data using collective filtering (finding the answer with collective intelligence from large audience) and use them to serve as a basis for future recommendation system research.

Project Background This is a capstone project in Data Science course series from Coursera.org. We are given a set of data from Yelp, which is an online business company founded in 2004. The dataset is part of the Yelp Dataset Chanllenge and it needs a space of approximately 575MB in the computer. The dataset is in the format of JSON and consists of 5 data files including the information of business, checkin, review, tip and user. We will explore this dataset and answer the question that we proposed.

Task To answer the question, there are several questions we need to answer first.

- Who are the females in the dataset
- How to determine what they like
- How to model the graduate student feature
- How to find the restaurants which they would prefer

With all those questions in mind, we develop a filtering method as shown in the methods session.

Methods

In this session, we are going to explain the steps for creating recommendation systems. I will first give you an overview of three major steps that I used for prediction followed by a more detailed description for each step.

Overview of Prediction Algorithm

- Step 1, cluster the reviewers into females and males from the user dataset thus find all the female users
- Step 2, list all the restaurant features that females prefer from restaurant reviews
- Step 3, look for and sort the restaurants in AZ which match most with the features females pay attention to in step two

More Details First, we use the data from "yelp_academic_dataset_user.json" to find the female reviewers. To do this, we need to get the data from Social Security Administration (see https://www.ssa.gov/oact/babynames/limits.html) for all the most used female and male names and compare those with the names on Yelp user data profile. This will give us a set of female reviews ID.

The number of users in user file is

```
## [1] 366715
```

After filtering with the female and male names, we get a set of female IDs. The size of female IDs is

```
## [1] 152369
```

Second, we abstract all the important features/key words which females pay attention to. To do this, we first take out all the restaurant IDs from "yelp_academic_dataset_business.json" file. The number of restaurants in file is

```
## [1] 21892
```

Then from "yelp_academic_dataset_review.json" file, we find all the reviews for restaurants based on the IDs we obtained above. Then by further filtering the restaurant reviews for the female IDs and using one-stem histogram from text mining, we can obtain a set of key words which are from the **female** customers with its frequencies. We also draw it in a word cloud fashion as you see in the next page.

```
head(frequency)
```

```
##  food  good place  like  just great
##  4496  4134  3612  2856  2742  2490
```

We can again use the same method for finding the most featured words for **all** the restaurant reviews.

To find the features which female reviewers care more than the average reviewers, we sort the words by frequency and compare the rank of the words frequency from female reviews to the average reviews. Then we obtain the features which females prefer:

```
## [1] "friends" "soup"    "happy"   "hour"    "friend"  "dessert"
```

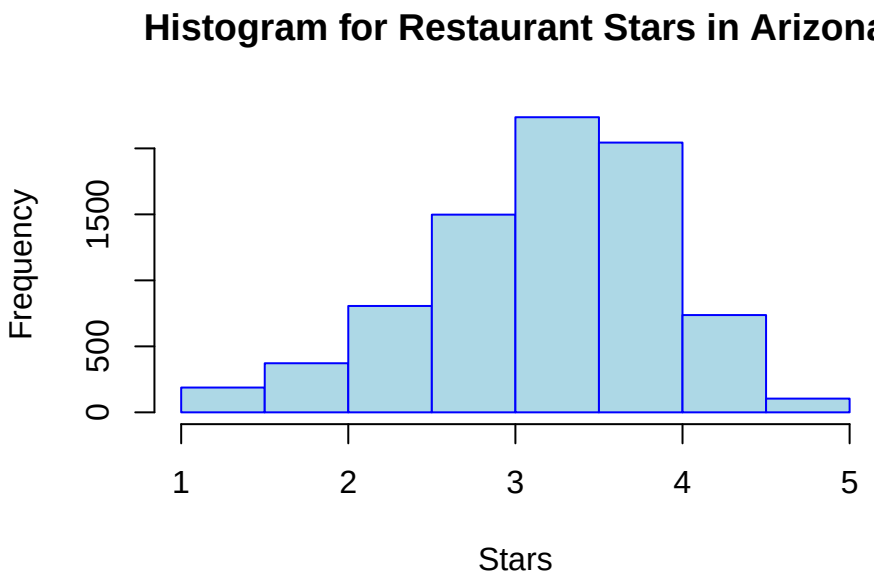
```
## Loading required package: RColorBrewer
```

Third, we find all the potential restaurants in Arizona State which have the desired characters for female gradaute students. From "yelp_academic_dataset_business.json" file, restaurant IDs in Arizona state can be drawn. Then we confine our results with the following filtering strategy:

1. The restaurants should have stars more than average score of the average in Arizona.
2. The restaurants should not be too pricy – to satisfy graduate students feature
3. The restaurants should have reviews which match most the key words we obtained from step 2 – to satisfy female feature

Results

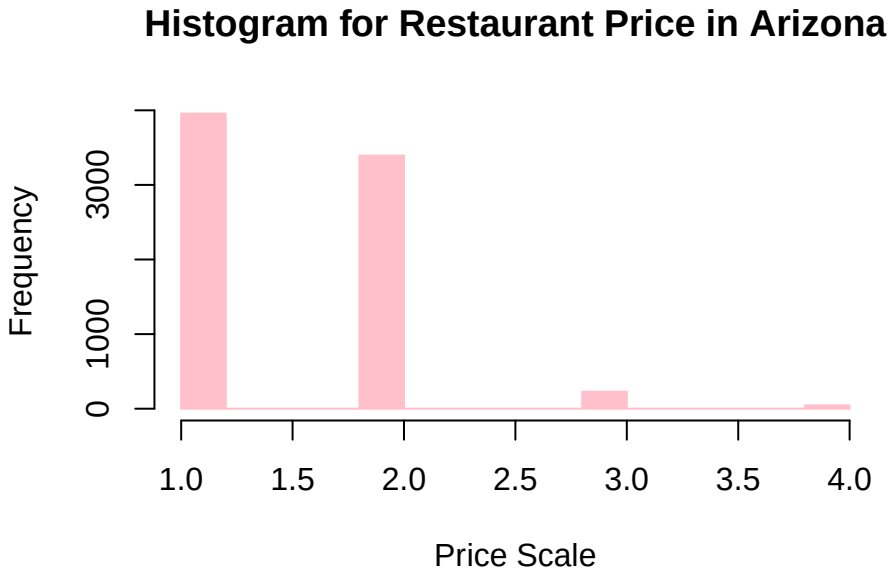
Here is the histogram of the restaurant stars in Arizona.



The mean value for the star is

```
## [1] 3.425798
```

Thus we need to remove the restaurants which has less than 3.4 stars. Similarly, we can draw a histogram for the restaurant prices in Arizona and remove those which are too pricy.



Here we assume that the price scale 4 is too pricy for students. Now a subset of the restaurant can be obtained based on the first two filtering rules. The number of remaining restaurants is

```
## [1] 788
```

Now we are looking up the reviews file to find all the potential restaurant reviews and filter them based on the matching between female prefered features and the reviews. The restaurants whose reviews have the most female preferred words will be sorted.

Answer The top restaurants female graduate students should go to are:

- "Crackers & Co Cafe"
- "Cafe Monarch"
- "Good Fellas Grill"
- "The Gladly"

Interpretation of the results Now we have found out top four restaurants which female graduate students would probably prefer to go. Taking a close look at the dataset, we realize that these four restaurants are not too pricy nor too low quality. And most importantly, they offer good desserts, happy-hour as well as are suitable for friends hanging out together. This is because when we were looking for the preferred features for females, we have found that they are interested in restaurants which emphasize on "friends", "soup", "happy hour", "dessert". And these four recommended restaurants apprear on top of the list because their reviews have the highest correlation with these words. With the high stars and reasonable prices along with the matching standards for female graduate students, I am sure these would be very good choices.

Discussion

This project is aiming at finding some restaurants to recommend to female graduate students. To achieve this goal, I have used clustering algorithms to divide the reviewers into females and males. With text mining techniques, I find the female-preferred features for restaurants according to their reviews compared with average reviews. Then I use those features to find the top matching restaurants in Arizona.

When analyzing the words which females care more about than males, I find it quite interesting to study the eating preferences between genders. For instance, females prefer more places where they have happy hours and provide space for friends to get together. I also did some research on the eating preferences. Some interesting insights can be found from a report by GrubHub. (See http://media.grubhub.com/files/doc_downloads/GrubHub-Inc-Men-vs-Women-Eating-Preferences-White-Paper_v001_b3cw14.pdf)

Limitations and future research The algorithm still has its limitations since although it looks at reviews which pays attention to the features females prefer but it fails to learn the positive or negative attitude for those features. So it is possible that the restaurants have reviews on its bad "happy hour" service. Thus in the future a sentiment analysis for the reviews when finding the recommended restaurants will be further explored.

This project is just a basic step to form a simple framework for recommendation systems. A more sophisticated algorithm can be done with neural networks and deep learning which are the other directions to go as well.