Benjamin Arancibia

12/20/2015

CUNY IS 622 – Machine Learning and Big Data

Building a Recommendation System for Restaurant Reviews

**Introduction**

The objective of this project is to build a recommendation system for restaurants using collaborative filtering. The process of identifying similar users and recommending what similar users like is called collaborative filtering. This project is structured in the following manner:

1. Exploratory data analysis
2. Neighborhood-based CF recommender
3. Predict a rating that user gives a restaurant that the user has not encountered before
4. Error analysis

**Background**

Collaborative filtering (CF) is a technique that is used by recommender systems. CF is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, and data sources. CF is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly. These predictions are specific to the user, but use information gleaned from many users. This project uses the Yhat example of building a beer recommendation system found [here](http://nbviewer.ipython.org/gist/glamp/20a18d52c539b87de2af).

**Methodology**

*Exploratory Data Analysis*

The first step in the analysis is collect the data from Yelp on restaurant reviews and then joined information about users and businesses. Fields in this data are:

['user\_id',

'business\_id',

'date',

'review\_id',

'stars',

'usefulvotes\_review',

'user\_name',

'categories',

'biz\_name',

'latitude',

'longitude',

'business\_avg',

'business\_review\_count',

'user\_avg',

'user\_review\_count']

After preparing the data, an exploratory analysis was performed. The exploratory analysis that was performed were to plot histograms of the review count grouped by the user\_id and business\_id. There are around 34,789 users and 4,503 business that with review counts. Afterwards then the average rating of reviews in the dataset and a histogram of all the ratings in the dataset. The next step in the exploratory data analysis is to plot histograms of the average user rating in the smaller dataset, and the average business rating in the dataset. The last part of the exploratory data analysis is to look at the common user support (the number of common reviewers) of each pair of restaurants on the dataset.

*Similarity*

The next step in the analysis is to calculate similarity and then create a database of similarities. By calculating similarities, the k-nearest restaurants to a given restaurant based on the database of similarities are calculated. The issue with this methodology is that there might be a small number or large number of common reviewers. This is resolved by normalizing the the common reviewers by:

After normalizing the data, the recommendation was made based on the idea of finding restaurants that users might also like based on the nearest neighbor to the original restaurant. After performing the nearest neighbor recommendation, the top recommendations for a user were then calculated.

*User Based Recommendation*

The next step in the analysis and creation of the recommendation project was to create a user based recommender with predicted ratings. The first part of this analysis define the predicted rating and then predict the rating for a user and an item. Then compute the predicted rating and compare it with the average rating over all users available.

*Error Analysis*

The next step in the project is to take a set of actual ratings, and a set of predicted ratings, and plots the latter against the former. This graph allows for a user to see how accurate the predictions are for the recommended restaurants. For selecting the nearest neighbor parameter of k, users are selecting a k that avoids outliers and this can be achieved by plotting graphs to see which k gives the most structured predictions with minimum outliers. Also by plotting the graphics, the user can see how close the mean lies to the slope of the predictions.

**Results**

Figure

The results of the overall project were good. Graphics from specific parts of each of the sections will be presented. Also the error percentage will be displayed and the graphics of different neighbor clusters and their error will be displayed. Figure 1 below shows the average rating of reviews and a histogram of the ratings in the dataset.

The results of the top matches for the Lobbys Beef Burgers Dogs, which has a business id of eIxSLxzIlfExI6vgAbn2JA are as follows:

0 La Condesa Gourmet Taco Shop | Sim 0.598714448434 | Support 6

1 Citizen Public House | Sim 0.571428571429 | Support 4

2 FnB | Sim 0.527129890943 | Support 5

3 Defalco's Italian Grocery | Sim 0.519456555658 | Support 6

4 Republic Ramen + Noodles | Sim 0.519140146937 | Support 5

5 unPhogettable | Sim 0.5 | Support 3

6 Haus Murphy's | Sim 0.467637235308 | Support 3

As described in the methodology, the ratings for a specific user, which is Vern, is as follows:

Vern the top recommendations are:

Rokerij | Average Rating | 4.37931034483

Wildfish Seafood Grille | Average Rating | 4.29411764706

Cornish Pasty Company | Average Rating | 4.20689655172

Pappadeaux Seafood Kitchen | Average Rating | 4.18518518519

Four Peaks Brewing Co | Average Rating | 4.16666666667

The Fry Bread House | Average Rating | 4.11538461538

Yasu Sushi Bistro | Average Rating | 4.07692307692

Pho Thanh | Average Rating | 4.04761904762

Carolina's Mexican Food | Average Rating | 3.91176470588

Pita Jungle | Average Rating | 3.91176470588

Malee's Thai Bistro | Average Rating | 3.875

Modern Steak | Average Rating | 3.84848484848

Arcadia Farms Cafe | Average Rating | 3.79310344828

Delux | Average Rating | 3.77611940299

CherryBlossom Noodle Cafe | Average Rating | 3.75

Hula's Modern Tiki | Average Rating | 3.74647887324

SanTan Brewing Company | Average Rating | 3.73076923077

Republic Ramen + Noodles | Average Rating | 3.69230769231

Oregano's Pizza Bistro | Average Rating | 3.68421052632

TEXAZ Grill | Average Rating | 3.65517241379

Pita Jungle | Average Rating | 3.63636363636

Yupha's Thai Kitchen | Average Rating | 3.625

Culinary Dropout | Average Rating | 3.62264150943

Rice Paper | Average Rating | 3.29411764706

The Breakfast Club | Average Rating | 3.23684210526

Stingray Sushi | Average Rating | 3.10714285714

Kabuki Japanese Restaurant | Average Rating | 3.08823529412

Teharu Sushi | Average Rating | 2.86666666667

This shows the top recommendations and the average rating of the restaurant from all the other ratings in the dataset. As a next step in the process, a predicted rating for a user will be displayed with the mean rating of the restaurant. Keeping Vern as the user, the predicted score of the restaurant and the mean of the restaurant are displayed below:

Rokerij | 4.71714023074 | Average 4.37931034483

Wildfish Seafood Grille | 4.27594504172 | Average 4.29411764706

Cornish Pasty Company | 4.62810510121 | Average 4.20689655172

Pappadeaux Seafood Kitchen | 4.08845573953 | Average 4.18518518519

Four Peaks Brewing Co | 4.26174734161 | Average 4.16666666667

The Fry Bread House | 4.2296159108 | Average 4.11538461538

Yasu Sushi Bistro | 4.61103444018 | Average 4.07692307692

Pho Thanh | 4.10317798035 | Average 4.04761904762

Carolina's Mexican Food | 4.31700962152 | Average 3.91176470588

Pita Jungle | 4.40378787384 | Average 3.91176470588

Malee's Thai Bistro | 4.39994642565 | Average 3.875

Modern Steak | 3.7629553634 | Average 3.84848484848

Arcadia Farms Cafe | 3.14091506468 | Average 3.79310344828

Delux | 3.91669620869 | Average 3.77611940299

CherryBlossom Noodle Cafe | 3.95053856287 | Average 3.75

Hula's Modern Tiki | 3.81932008942 | Average 3.74647887324

SanTan Brewing Company | 3.31082382422 | Average 3.73076923077

Republic Ramen + Noodles | 3.50907720603 | Average 3.69230769231

Oregano's Pizza Bistro | 3.91500219386 | Average 3.68421052632

TEXAZ Grill | 3.89033612648 | Average 3.65517241379

Pita Jungle | 3.73144334211 | Average 3.63636363636

Yupha's Thai Kitchen | 3.01145099571 | Average 3.625

Culinary Dropout | 3.50206479813 | Average 3.62264150943

Rice Paper | 3.25235100123 | Average 3.29411764706

The Breakfast Club | 3.04653217913 | Average 3.23684210526

Stingray Sushi | 3.02529455724 | Average 3.10714285714

Kabuki Japanese Restaurant | 3.18645444652 | Average 3.08823529412

Teharu Sushi | 2.43427557446 | Average 2.86666666667

Finally, for the error analysis the error was calculated can be seen by substracting the accuracy from one. This gives the percent error. The actual rating vs the predicting rating for 3 clusters is below in Figure 2.



Figure

**Conclusion**

There are limitations to the approach and how it was implemented. Next steps include to create a more in-depth recommendation model, which could be a Bayesian Model or other recommendation model. The Bayesian model for a recommendation system allows bring information into the model that are about similar users and about similar restaurants to create more accurate and robust ratings. Also, the current computation speed is quite slow. The next step would be use map reduce techniques, though the dataset is large it can still be calculated in memory, though it is computationally expensive.

# References

Employee, B. A. (n.d.). Tech Tank.

*Harvard School of Engineering and Applied Sciences*. (n.d.). Retrieved from CS109 Data Science: http://cs109.github.io/2015/index.html

Leskovec, J., Rajaraman, A., & Ullman, J. (2014). *Mining of Massive Datasets.*

Yhat. (n.d.). *Building a Recommendation System in Python*. Retrieved from Yhat: http://nbviewer.ipython.org/gist/glamp/20a18d52c539b87de2af