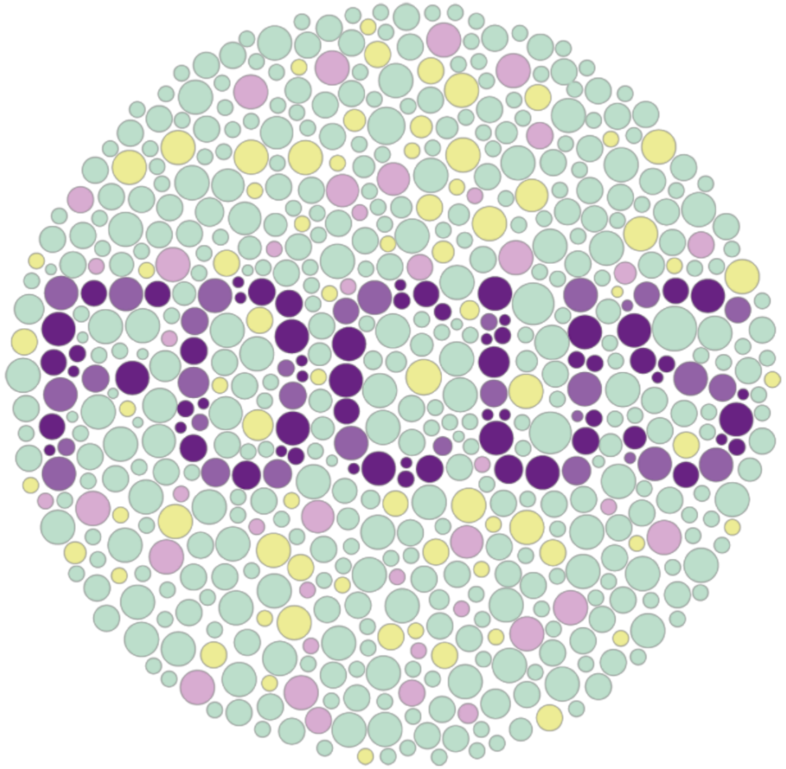


A Restaurant Recommendation System for Yelp

Location-based Collaborative Filtering and Frequent Itemset

Thanh Tung Nguyen (ID: 40042891) & Huy Nguyen (ID: 40023289) @ ***April 12, 2019***

Agenda

Project Overview	3
Dataset & Methods Used	4
Algorithm & App Result	5
Issue Explanations	7
Conclusion	8

Project Overview

Project Introduction & Related Works

PROJECT INTRODUCTION

1. **What we try to achieve:**

    - To help Yelp users make better choices of restaurants, we use techniques and principles of recommendation systems to create an application which makes predictions based on the user similarities

    - Develop an enhanced collaborative filtering using location (**postal codes**) as a key criterion for generating recommendations (Scope of Work: **Canada**)

2. **Methods we use:**

    - Collaborative Filtering
    - Frequent Itemset

3. **How we evaluate the results**

    - Use Root metrics Mean Squared Error (RMSE)
    - User Mean Absolute Error (MAE)

RELATED WORKS

1. **Using location for personalized POI recommendations in mobile environments:**

    - By *Tzvetan Horozov, Nitya Narasimhan, Venu Vasudevan*

    - Discussion of GeoWhiz, a real-world deployment of our restaurant recommender system for location-based points of interest (POI).

2. **Collaborative Filtering using Weighted BiPartite Graph Projection - A Recommendation System for Yelp**

    - By *Sumedh Sawant*

    - Recommendation system on the Yelp Dataset Challenge dataset using the network-based-inference collaborative filtering algorithm

    - Same Yelp dataset was used (2013 version)

Dataset & Methods Used

Yelp Dataset and Collaborative Filtering & Frequent Itemset

YELP DATASET

1. **Source of Data:**

- Yelp Dataset - Yelp's businesses - 4 GB - www.kaggle.com/yelp-dataset/yelp-dataset

- Canadian Postal Codes - Google Fusion Tables - 49 MB - https://fusiontables.google.com/

2. **Dataset overview:**

Original Dataset

Number of businesses	192,609
Number of review	6,685,900
Number of users	1,637,138

Canada

Number of Canadian businesses	50,644
Number of Canadian reviews.	1,063,142

Canadian Postal Codes

Number of postal codes	889,320

METHODS USED

1. **Collaborative Filtering :**
   - Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). Matrix factorization is a good solution for sparse data problem.

$$\tilde{r}_{ui} = \sum_{f=0}^{nfactors} H_{u,f} W_{f,i}$$

   - where H is user matrix, W is item matrix

2. **Frequent Itemset**
   - Find sets of items that appear together ‘frequently’ in baskets with a minimum support and confidence to be qualify as ‘frequent’
   - Association Rules

$$\begin{tikzcd}
 & Support = \dfrac{frq(X,Y)}{N} \\
Rule\colon\ X \Rightarrow Y \arrow[ur] \arrow[r] \arrow[dr] & Confidence = \dfrac{frq(X,Y)}{frq(X)} \\
 & Lift = \dfrac{Support}{Supp(X) \times Supp(Y)}
\end{tikzcd}$$

Project Overview

Dataset & Methods Used

Algorithm & App Result

Issue Explanations

Conclusion

Algorithms & App Results

Algorithms & Results (1/2)

SN	Algorithms	Result
Step 1	• In order to get a user with good history profile, we sort top 100 most review users in Canada, then take 5 random users and select one.	
Step 2	• To make sure the user's rating history and their restaurant choices are relevant, we find the base city AND top most reviewed postal codes of the user based on their rating history. • We then prompt the user input their location (select one of the top postal codes).	
Step 3	• Find all postal codes and their associated businesses located within a 3-km radius from **the chosen postal code**	

Step 1 result:

SN	Item Description
1.	65yB0ydGXOZ_-T6J_GbKfw
2.	jnB_saJqNfOmVoCWquhAzg
3.	iRQ_YKpCBdaCwvc2X8_3NQ
4.	tWBLn4k1M7PLBtAtwAg73g
5.	Wu0yySWcHQ5tZ_59HNiamg

Step 2 result:

SN	Item Description
1.	M8X 1E9
2.	M5A 2L2
3.	M6K 1L4
4.	M5T 2W6
5.	M5V 3M4

Step 3 result:

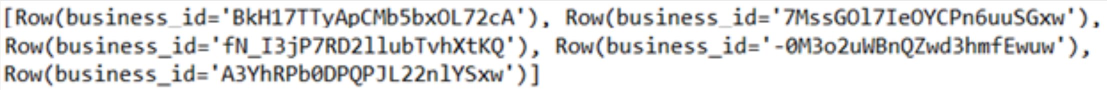
	3 km	5 km	10 km
Number of postal codes	2,835	6,731	21,050
Number of businesses	582	1,506	8,005

Algorithms & App Results

Algorithms & Results (2/2)

SN	Algorithms	Result
Step 3	• Use Pyspark MLlib ALS library to do basic ALS recommender, as well as global average recommender. Compute RMSE and MAE for both approaches and take recommendation from the better RMSE approach.	see below
Step 4	• Use FP-Growth Library in Pyspark to perform Frequent Itemset listing out top 5 most frequently chosen items. • Selected confidence and support values are 0.4 and 0.4 respectively • It resulted to an empty RDD	see below

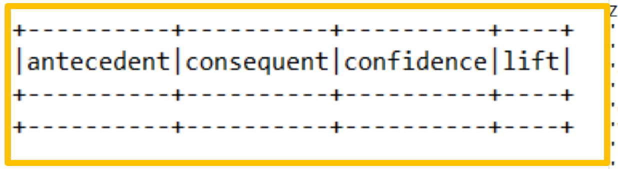
Step 3 – Result



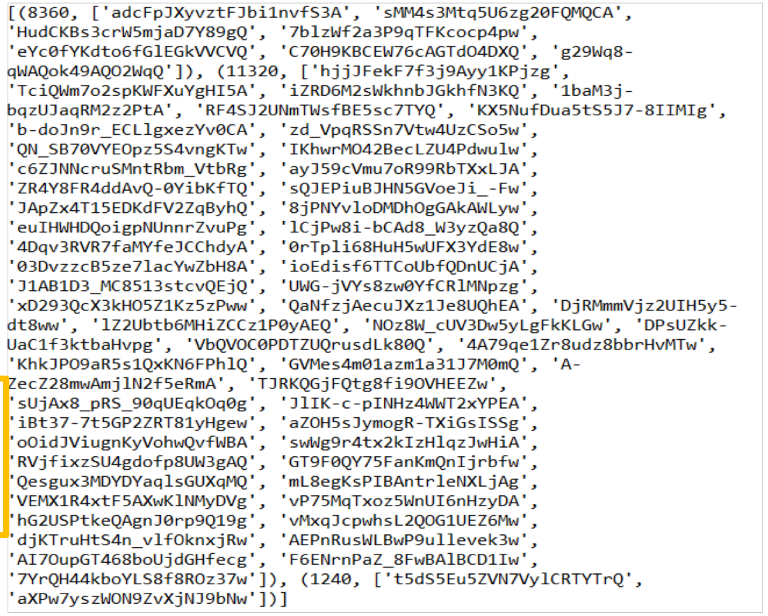
```
[Row(business_id='BkH17TTyApCMb5bxOL72cA'), Row(business_id='7MssGOl7IeOYCPn6uuSGxw'),
Row(business_id='fN_I3jP7RD2llubTvhXtKQ'), Row(business_id='-0M3o2uWBnQZwd3hmfEwuw'),
Row(business_id='A3YhRPb0DPQPJL22nlYSxw')]
```

	RMSE	MAE
Basic ALS Recommender	2.147847095	1.722178277
Global Average Recommender	1.435300565	1.242512298

Step 4 – Result



```
+----------+----------+----------+----+
|antecedent|consequent|confidence|lift|
+----------+----------+----------+----+
+----------+----------+----------+----+
```



```
[(8360, ['adcFpJXyvztFJbi1nvfS3A', 'sMM4s3Mtq5U6zg20FQMQCA',
'HudCKBs3crW5mjaD7Y89gQ', '7blzWf2a3P9qTFKcocp4pw',
'eYc0fYKdto6fGlEGkVVCVQ', 'C70H9KBCEW76cAGTdO4DXQ', 'g29Wq8-
qWAQok49AQO2WqQ']), (11320, ['hjjJFekF7f3j9Ayy1KPjzg',
'TciQWm7o2spKWFXuYgHI5A', 'iZRD6M2sWkhnbJGkhfN3KQ', '1baM3j-
bqzUJaqRM2z2PtA', 'RF4SJ2UNmTWsfBE5sc7TYQ', 'KX5NufDua5tS5J7-8IIMIg',
'b-doJn9r_ECLlgxezYv0CA', 'zd_VpqRSSn7Vtw4UzCSo5w',
'QN_SB70VYEOpz5S4vngKTw', 'IKhwrMO42BecLZU4Pdwulw',
'c6ZJNNcruSMntRbm_VtbRg', 'ayJ59cVmu7oR99RbTXxLJA',
'ZR4Y8FR4ddAvQ-0YibKfTQ', 'sQJEPiuBJHN5GVoeJi_-Fw',
'JApZx4T15EDKdFV2ZqByhQ', '8jPNYvloDMDhOgGAkAWLyw',
'euIHWHDQoigpNUnnrZvuPg', 'lCjPw8i-bCAd8_W3yzQa8Q',
'4Dqv3RVR7faMYfeJCChdyA', '0rTpli68HuH5wUFX3YdE8w',
'03DvzzcB5ze7lacYwZbH8A', 'ioEdisf6TTCoUbfQDnUCjA',
'J1AB1D3_MC8513stcvQEjQ', 'UWG-jVYs8zw0YfCRlMNpzg',
'xD293QcX3kHO5Z1Kz5zPww', 'QaNfzjAecuJXz1Je8UQhEA', 'DjRMmmVjz2UIH5y5-
dt8ww', 'lZ2Ubtb6MHiZCCz1P0yAEQ', 'NOz8W_cUV3Dw5yLgFkKLGw', 'DPsUZkk-
UaC1f3ktbaHvpg', 'VbQVOC0PDTZUQrusdLk80Q', '4A79qe1Zr8udz8bbrHvMTw',
'KhkJPO9aR5s1QxKN6FPhlQ', 'GVMes4m01azm1a31J7M0mQ', 'A-
ZecZ28mwAmjlN2f5eRmA', 'TJRKQGjFQtg8fi9OVHEEZw',
'sUjAx8_pRS_90qUEqkOq0g', 'JlIK-c-pINHz4WWT2xYPEA',
'iBt37-7t5GP2ZRT81yHgew', 'aZOH5sJymogR-TXiGsISSg',
'oOidJViugnKyVohwQvfWBA', 'swWg9r4tx2kIzHlqzJwHiA',
'RVjfixzSU4gdofp8UW3gAQ', 'GT9F0QY75FanKmQnIjrbfw',
'Qesgux3MDYDYaqlsGUXqMQ', 'mL8egKsPIBAntrleNXLjAg',
'VEMX1R4xtF5AXwKlNMyDVg', 'vP75MqTxoz5WnUI6nHzyDA',
'hG2USPtkeQAgnJ0rp9Q19g', 'vMxqJcpwhsL2QOG1UEZ6Mw',
'djKTruHtS4n_vlfOknxjRw', 'AEPnRusWLBwP9ullevek3w',
'AI7OupGT468boUjdGHfecg', 'F6ENrnPaZ_8FwBAlBCD1Iw',
'7YrQH44kboYLS8f8ROz37w']), (1240, ['t5dS5Eu5ZVN7VylCRTYTrQ',
'aXPw7yszWON9ZvXjNJ9bNw'])]
```

Project Overview

Dataset & Methods Used

Algorithm & App Result

Issue Explanations

Conclusion

Issue Explanations

Issues Encountered & Explanations

<table>
<tr><th>SN</th><th>Issue Description</th><th>Solution & Explanation</th></tr>
<tr><td><b>Issue 1</b></td><td>• Global Average RMSE is better than ALS RMSE

• Increase search radius and add biases to ALS do not make it better</td><td>• <b>Test 1: Increase search radius from 3 to 5 and to 10 km</b>
<table>
<tr><th></th><th>3 km (RMSE)</th><th>5 km (RMSE)</th><th>10 km (RMSE)</th></tr>
<tr><td>Basic ALS Recommender RMSE</td><td>2.127391175</td><td>2.147847095</td><td>1.852597978</td></tr>
<tr><td>Global Average Recommender RMSE</td><td>1.435300565</td><td>1.435300565</td><td>1.329057787</td></tr>
</table>
• <b>Test 2: Switch the metric from ALS to ALS + Biases</b>
<table>
<tr><th></th><th>RMSE</th><th>MAE</th></tr>
<tr><td>Bias ALS Recommender</td><td>3.042489076</td><td>2.554701563</td></tr>
<tr><td>Global Average Recommender</td><td>1.435300565</td><td>1.242512298</td></tr>
</table>
✓ Perhaps there are many latent factors that are not in the dataset itself to compute a more accurate matrix factorization.
✓ Perhaps there is not many “similar” users who rate one or more businesses like each other in our narrow distance.</td></tr>
<tr><td><b>Issue 2</b></td><td>• FP-Growth gives out empty result even when confidence and support value are are lowered down to 0.1</td><td>• <b>Test 1: Increase search radius from 3 to 5 and to 10 km</b>
▶ Still gives empty RDD
• <b>Test 2: Reduce confidence and support value from 0.4 to 0.1</b>
▶ Still gives empty RDD
✓ There is no similarity frequent itemset within small-distance localities</td></tr>
</table>

Project Overview

Dataset & Methods Used

Algorithm & App Result

Issue Explanations

Conclusion

Conclusion

Our Conclusion, Future Work, and Q&A

CONCLUSION

1. Yelp official open dataset is not suitable for small scale locality recommendation as there is a low possibility of similar-rating restaurant sets among users

2. This also means that frequent itemset method might not be applicable since there is low possibility of frequent patterns in a small scale.

FUTURE WORKS

1. Look for alternative way to do recommendation, such as Cluster Weighted BiPartite Projection or Multi-Step Random Walks. For example, the project performed by Sawant - “Collaborative Filtering using Weighted BiPartite GraphProjection - A Recommendation System for Yelp” shows remarkable improvement.

2. Additional data attributes and information from Yelp could be taken into account, such as type of restaurant and its price range to improve algorithm, in order to give a more precise result.

3. Find out the real correlation between the issues and actual restaurant business natures.

Questions

& Answers

Thank you.