SOEN 691 : BIG DATA ANALYTICS PROJECT

RESTAURANT RECOMMENDER SYSTEM

Presented By - Team 9

Members:

Dhaval Modi (40083289)
Sai Krishna (40087977)
Manushree (40082236)

Introduction

- Goals
  - To Implemented the recommender systems to recommend the restaurants using Content Based Filtering and Collaborative Filtering.
  - Compare the RMSE, MEA and MSE metrics of these algorithms.

Dataset

- Collected using the APIs.
- Data collected in CSV files
- Restaurants ~ 42k
- Reviews ~ 1400k
- Users ~ 200K

Restaurant.csv

- No of columns
  - 105 attributes of different restaurants
- Important fields
  - restaurant_id, city, categories
- Sample Data

restaurant_id	categories	TotalReviews	name
0k6wQvOkc7aioBcAv...	Car Stereo Instal...	15	Sound Xpression
m3O2EFVs3eo8s40MG...	Dine-out	69	Buono's Pizza
ZvXEc7adhC2N6zQXN...	Nightlife	7	Beautiful Nails b...
CdEmvGXqIsUS3e5dE...	Oil Change Statio...	6	Jiffy Lube
H-99nG_KfynWPXh84...	Pakistani;Indian;...	32	Copper Kettle-Sal...

Reviews.csv

- No of columns
  - 10 attributes of different restaurants
- Important fields
  - user_id, restaurant_id, user_rating
- Sample Data

user_id	restaurant_id	review_id	user_rating	text
Xqd0DzHaiyRqVH3WR...	vcNAWiLM4dR7D2nww...	15SdjuK7DmYqUAj6r...	5	dr. goldberg offe...
H1kH6QZV7Le4zqTRN...	vcNAWiLM4dR7D2nww...	RF6UnRTtG7tWMcrO2...	2	Unfortunately, th...
zvJCcrpm2yOZrxKff...	vcNAWiLM4dR7D2nww...	#NAME?	4	Dr. Goldberg has ...
KBLW4wJA_fwoWmMhi...	vcNAWiLM4dR7D2nww...	dNocEAyUucjT371NN...	4	Been going to Dr....
zvJCcrpm2yOZrxKff...	vcNAWiLM4dR7D2nww...	ebcN2aqmNUuYNoyvQ...	4	Got a letter in t...

Content Based Filtering

- Steps
  - Data Preprocessing
  - Normalize the data
  - Generate frequency–inverse values for different categories.
  - Generate user profiles
  - Generate the model predictions
  - Calculate the RMSE, MEA and MSE metric value
  - Generate the recommended restaurants

Data Processing

- Removed restaurants with null restaurant_id
- Removed users with null user_id
- Filtered out restaurant based on categories.
- Filtered restaurants based on city

Data Normalization

Normalized value = 1 / sqrt (n1 + n2 + n3 + ...)

restaurant_index	Afghan;Restaurants	Barbeque;Restaurants	norm_value
26	0	0	1.0
29	0	0	1.0
191	0	0	1.0
222	0	0	1.0
270	0	0	1.0

Algorithm

Frequency Inverse (TF-IDF)

Frequency Inverse of category= log10(x/y)

x = total number of restaurant

y = number of restaurant in category

```
Asian Fusion;Restaurants 2.7234556720351857
Breakfast & Brunch;Restaurants :  1.769213162595861
Peruvian;Cuban;Caribbean;Restaurants :  2.7234556720351857
Fast Food;Pizza;Restaurants :  2.7234556720351857
Seafood;Fish & Chips;Restaurants :  2.7234556720351857
```

User Profiles

Rating Threshold value : 3

Rating >= 3 => User value = 1

Rating <= 3 => User value = -1

user_index	sum(Burgers;Bars;Steakhouses;Nightlife;Restaurants)	Barbeque;Restaurants
26	0.0	1
300443	0.0	0
13723	0.0	0
178443	-1.0	0
60683	0.0	0

Model Predictions

$$\text{Prediction} = \frac{m - r1}{r2 - r1} * (t2 - t1) + t1$$

r1 = -1 t1 = 1

r2 = 1 t2= 5

m = cos(Θ) of User vector & Category vector

prediction	restaurant_index	uid	user_rating
5.0	26	1128	4
5.0	26	1252	5
5.0	26	854	5
3.3922322702763683	26	5	4
1.0	26	811	2

Metrics

RMSE = 1.406242291057611

MAE = 1.174760279634468

MSE = 1.9775173811589584

Recommended Restaurants

uid	restaurant_index	name	prediction	stars	categories
0	82	Thai Food Corner	3.659380473395787	3.5	Thai;Restaurants
0	4	Palee's Crown	3.659380473395787	4.0	Thai;Restaurants
0	272	Thai Patio	3.659380473395787	4.0	Thai;Restaurants
0	160	Benjarong Thai Re...	3.659380473395787	4.0	Thai;Restaurants
0	37	Swan Thai	3.659380473395787	3.5	Thai;Restaurants
0	224	Thai House Restau...	3.659380473395787	4.0	Thai;Restaurants
0	186	Royal Thai Grill	3.659380473395787	4.0	Thai;Restaurants
0	208	Mango Thai	3.659380473395787	3.0	Thai;Restaurants
0	109	Thai Basil	3.659380473395787	4.0	Thai;Restaurants
0	106	Culver's	3.5275043787166296	3.0	Food;Burgers;Fast...

Collaborative Based Filtering

- Steps
  - Data Preprocessing
  - Normalize the data
  - Split data into train, validation, test(Train-test- validation split)
  - Calculate RMSE
  - Make rating predictions on all restaurants for that user
  - Generate the recommended restaurants based on the ranking of restaurant rating predictions

Data Preprocessing

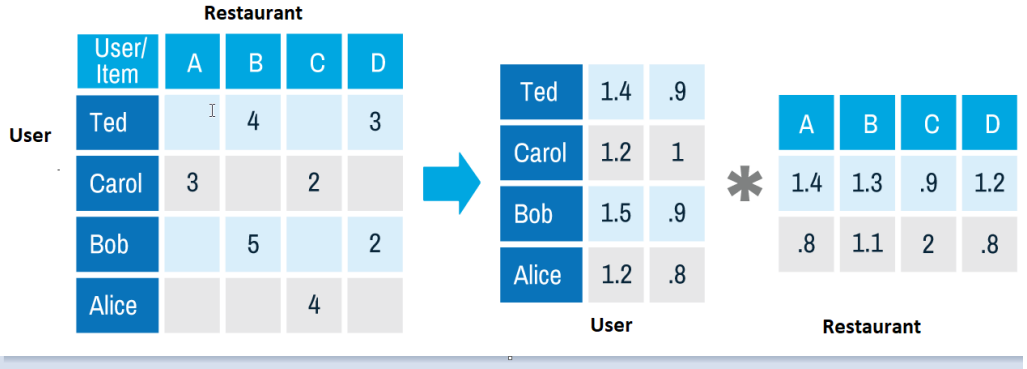
- Join the reviews file and restaurants file based on the restaurant ID

restaurant_id	user_id	user_rating
vcNAWiLM4dR7D2nww...	Xqd0DzHaiyRqVH3WR...	5
vcNAWiLM4dR7D2nww...	H1kH6QZV7Le4zqTRN...	2
vcNAWiLM4dR7D2nww...	zvJCcrpm2yOZrxKff...	4
vcNAWiLM4dR7D2nww...	KBLW4wJA_fwoWmMhi...	4
vcNAWiLM4dR7D2nww...	zvJCcrpm2yOZrxKff...	4
vcNAWiLM4dR7D2nww...	Qrs3EICADUKNFoUq2...	1
vcNAWiLM4dR7D2nww...	jE5xVugujSaskAoh2...	5
vcNAWiLM4dR7D2nww...	QnhQ8G51XbUpVEyWY...	5
JwUE5GmEO-sH1FuwJ...	zvNimI98mrmhgNOOr...	4
JwUE5GmEO-sH1FuwJ...	p4ySEi8PEli0auZGB...	4

only showing top 10 rows

ALS Algorithm

- A popular approach for this is matrix factorization, where we fix a relatively small number k (e.g. $k \approx 10$) and summarize each user u with a k dimensional vector x_u, and each item i with a k dimensional vector y_i . These vectors are referred to as factors. Then, to predict user u's rating for item i, we simply predict $rui \approx x^T_u y_i$



Hyperparams

- These are important in ALS, After optimizing we choose following values of different params
- maxIter : 10 (It is the maximum number of iterations to run)
- rank : 70 (The number of latent factors in the model)
- regParam : 0. 5 (The regularization parameter in ALS)

```
ALS (maxIter=5, regParam=0.5, rank=70,
userCol="userId", itemCol="movieId",
        ratingCol="user_item_interaction",
        coldStartStrategy="drop", seed=seed)
```

Metrics

- We have used the ALS plus biases and calculated predictions as follows
- **user_item interaction+user_mean+item_mean-global rating**
- RMSE : It measures numerical difference between all ground-truth ratings and actual ratings in test set.
- MAE: It measures numerical difference between all ground-truth ratings and actual ratings in test set

Metrics

$$MAE = \frac{1}{N}\sum |predicted - actual|$$

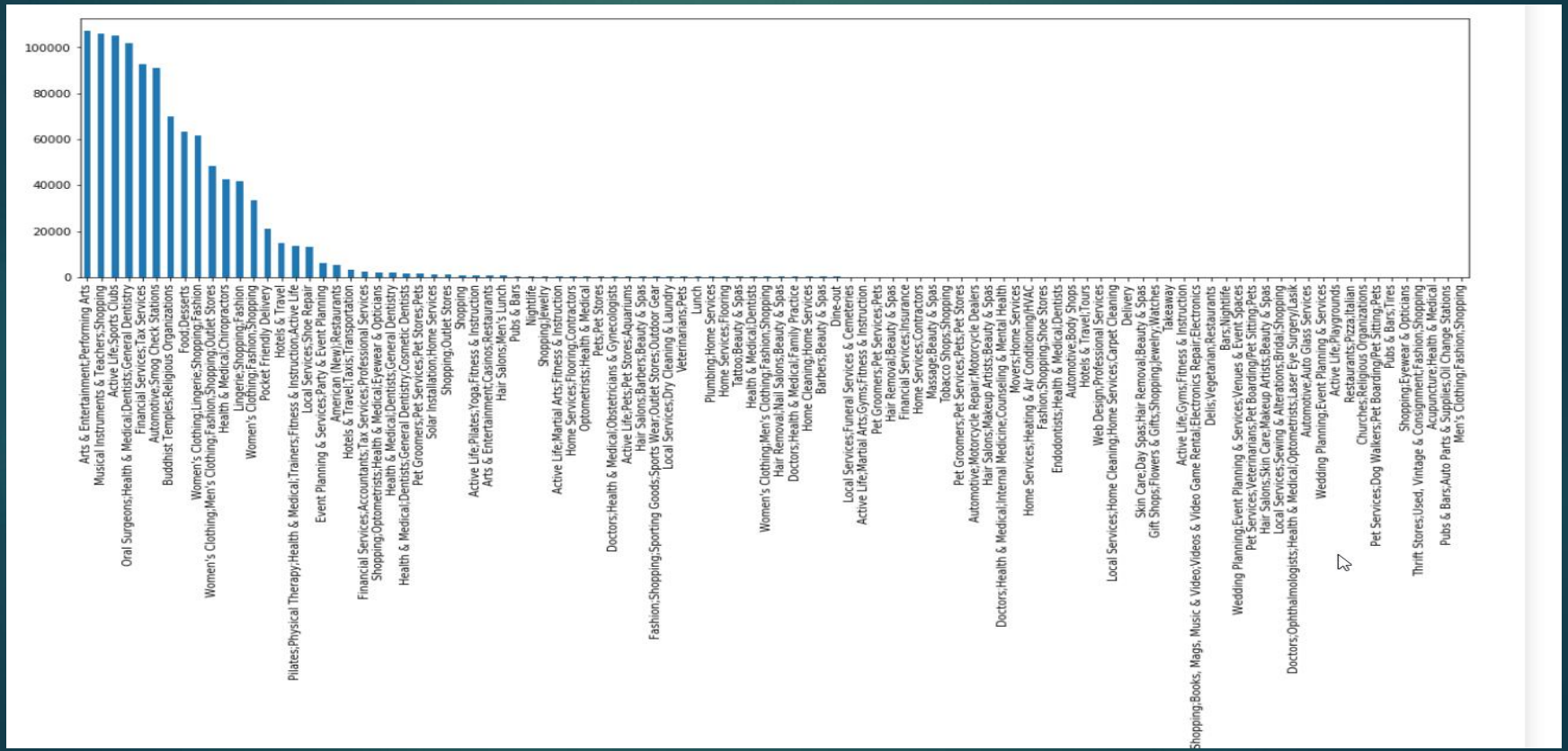
$$RMSE = \sqrt{\frac{1}{N}\sum (predicted - actual)^2}$$

RMSE	1.2542102156106667
MAE	0.9522392380468712
MSE	1.5730432649421549

	user_id	recommendations
0	--0ml_q_0D1CdU4P_holmQ	[(Gorilla Cafe, 5.184454441070557), (Island De...
1	--2QZsyXGz1OhiD4-0FQLQ	[(Jiffy Smog, 5.453772068023682), (Conrado & C...
2	--4fX3LBeXoE88gDTK6TKQ	[(SGI-USA Buddhist Center, 3.6345648765563965)...
3	--9HuvEtLhp21SelEEltnA	[(Desert Strings, 5.777367115020752), (Conrado...
4	--RM-_N1-5AiFSzdU5Gn1Q	[(SGI-USA Buddhist Center, 4.083629608154297),...
5	-02lt4LFEgradr6uK9KCfA	[(Smith's Food & Drug Centers Inc, 4.007578849...
6	-02LD_yedEFQKLuLDO-gWw	[(Gorilla Cafe, 5.564972400665283), (Gap Outle...
7	-05sPcciFbK8O_Npx8Dz2A	[(Kiki de Montparnasse, 4.699824333190918), (D...
8	-0AH2QoTlvKDf7CcjkJ39Q	[(Dior Boutique, 4.62614107131958), (Kiki de M...
9	-0ETqxKsxlBnmJklugoTrg	[(Rod Stewart: the Hits, 5.019903659820557), (...

Recommendations

Distribution of restaurants by category



Results Comparison

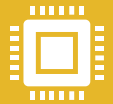
	Collaborative	Content based
RMSE	1.2542102156106667	1.406242291057611
MAE	0.9522392380468712	1.174760279634468
MSE	1.5730432649421549	1.9775173811589584
Runtime	768000 ms	930000 ms

Conclusions

- The RMSE of content based is higher than the collaborative.
- Restaurant ratings are diverse in content-based recommender.
- Content based recommended the restaurants that have high review count whereas collaborative recommendations tend to have higher ratings and lower review count,
- Collaborative Filtering algorithm has the limitations of Cold-start problem where a recommender does not have the adequate information about a user or an item to make relevant predictions. Data Sparsity Is the problem that occurs as a result of lack of enough information

Limitations

As API gives the large data around 400 gb, it is difficult to extract and process.



There are limited hardware resources to process and filter large data.

Future Work

Hybrid recommender systems - The best approach would be to use a combination of different approaches. Mix of collaborative and content-based filtering. Some of it will depend on preferences of the users and some on item features.

We could predict the rating on all cities instead of one cities which gives user preferences in various cities.

Implement recommender systems with various algorithms such as SVD++,Bayesian Personalized Ranking (BPR) and deep learning algorithms.

THANK YOU!

ANY QUESTIONS ?