Hotel Recommendation System

Pratik Parekh pparekh2@uncc.edu 801076521

Tannu Singh tsingh9@uncc.edu 801085297

Srishti Tiwari stiwari8@uncc.edu 801092498

Overview

The project aims to provide recommendations of the hotels to a particular user based on the past history of his/her reviews and past data of other reviews.

The goal of the project is to implement collaborative filtering and content based filtering algorithm with the help of Pyspark to build a recommendation system.

Context and Motivation

Recommendation system was introduced in online shopping to help the users to get personalized recommendations depending on their likes and dislikes. In addition to that, recommendation system also helps in increasing sales of the business along with satisfying the users with the recommendation.

In this project, we aim to provide content based recommendation as mentioned in Chen T, Han W (2007) [2], along with rating based recommendation.

Dataset Description

Dataset consists of 10000 rows and 25 columns provided by Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more.

Dataset link: https://data.world/datafiniti/hotel-reviews

Exploratory Analysis on dataset:

a. The dataset consists of more positive reviews than negative as mentioned in fig 1
b. The wordcloud shows that there are more positive words than negative word. As shown in fig 2(Wordcloud was created after removing stopwords)

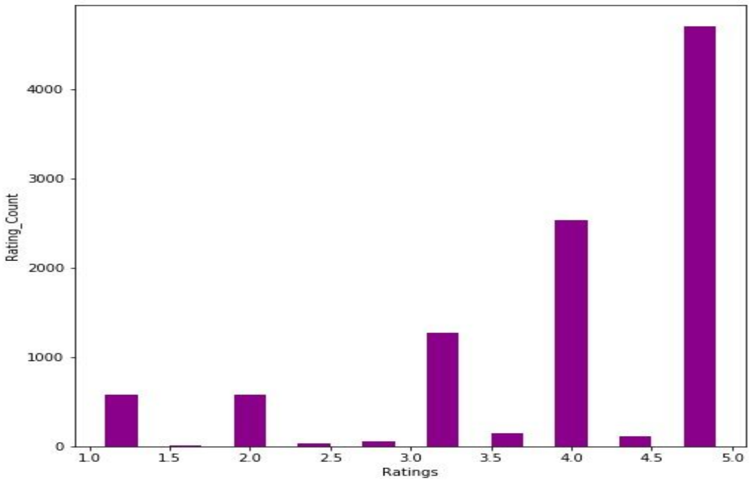


fig 1: Rating frequency for each rating from (1-5)

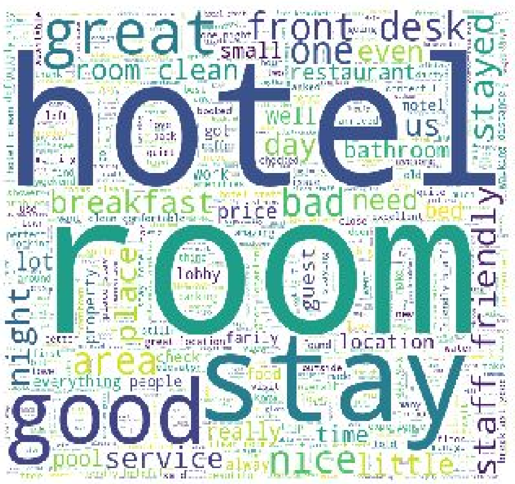


fig 2: Word cloud of reviews by all users.

Tasks Involved and Steps implemented

The project uses collaborative and content based recommendation to provide top 10 recommendations for a given user. The tasks involved in this are as follows:

1. Understanding and Implementing the Algorithm.
2. Setup up EMR Cluster Notebook.
3. Preprocessing of Data to bring in suitable form.
4. Implementing a pre processing pipeline to form word vectors.
5. Finding similarity score between hotels based on the reviews.
6. Implementing Content based recommendation for a given user very to to the recommendation system described in Chen T, Han W (2007)[2].
7. Providing results and details of recommended hotels.
8. Implementing Collaborative filtering to provide recommendation
9. Providing results and details of recommended hotels.
10. Documentation of the process.

Approach and Algorithm

1. ALS:

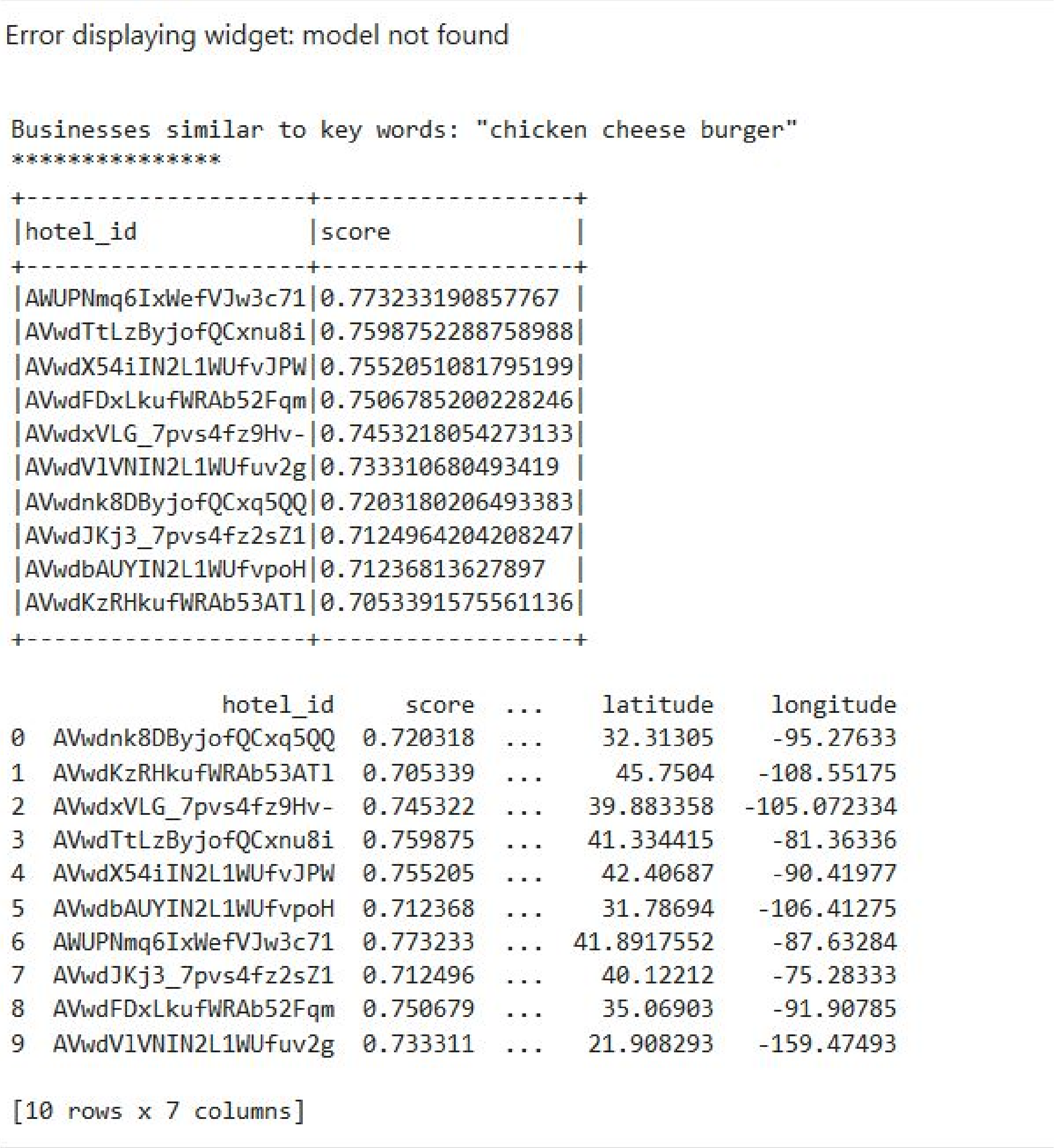
   Apache Spark ML implements alternating least squares (ALS) for collaborative filtering, a very popular algorithm for making recommendations.

   ALS recommender is a matrix factorization algorithm that uses Alternating Least Squares with Weighted-Lambda-Regularization (ALS-WR). It factors the user to item matrix *A* into the user-to-feature matrix *U* and the item-to-feature matrix *M*: It runs the ALS algorithm in a parallel fashion. The ALS algorithm should uncover the latent factors that explain the observed user to item ratings and tries to find optimal factor weights to minimize the least squares between predicted and actual ratings.

2. TFIDF : We used TFIDF to find the similarity of user reviews
3. Cosine Similarity: Cosine similarity is also another function we have made to find the similarity between reviews.

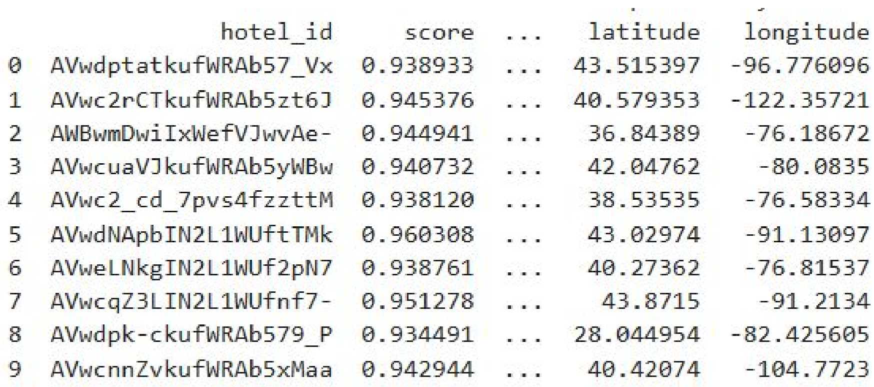
Sample Output and Result

1. Sample Output for Content based system where we are using the words common between user reviews and displaying the recommendation



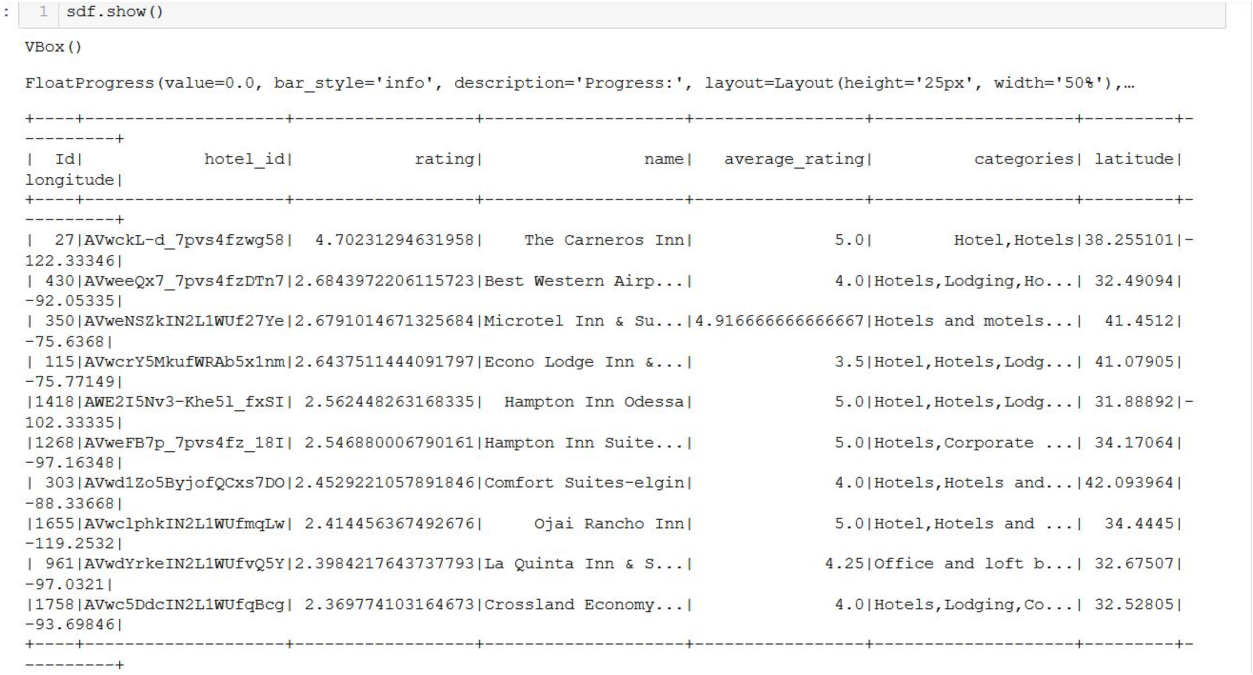
```
Error displaying widget: model not found


Businesses similar to key words: "chicken cheese burger"
***************
+--------------------+------------------+
|hotel_id            |score             |
+--------------------+------------------+
|AWUPNmq6IxWefVJw3c71|0.773233190857767 |
|AVwdTtLzByjofQCxnu8i|0.7598752288758988|
|AVwdX54iIN2L1WUfvJPW|0.7552051081795199|
|AVwdFDxLkufWRAb52Fqm|0.7506785200228246|
|AVwdxVLG_7pvs4fz9Hv-|0.7453218054273133|
|AVwdV1VNIN2L1WUfuv2g|0.733310680493419 |
|AVwdnk8DByjofQCxq5QQ|0.7203180206493383|
|AVwdJKj3_7pvs4fz2sZ1|0.7124964204208247|
|AVwdbAUYIN2L1WUfvpoH|0.71236813627897  |
|AVwdKzRHkufWRAb53AT1|0.7053391575561136|
+--------------------+------------------+

              hotel_id     score  ...   latitude   longitude
0  AVwdnk8DByjofQCxq5QQ  0.720318  ...   32.31305   -95.27633
1  AVwdKzRHkufWRAb53AT1  0.705339  ...    45.7504  -108.55175
2  AVwdxVLG_7pvs4fz9Hv-  0.745322  ...  39.883358 -105.072334
3  AVwdTtLzByjofQCxnu8i  0.759875  ...  41.334415   -81.36336
4  AVwdX54iIN2L1WUfvJPW  0.755205  ...   42.40687   -90.41977
5  AVwdbAUYIN2L1WUfvpoH  0.712368  ...   31.78694  -106.41275
6  AWUPNmq6IxWefVJw3c71  0.773233  ... 41.8917552   -87.63284
7  AVwdJKj3_7pvs4fz2sZ1  0.712496  ...   40.12212   -75.28333
8  AVwdFDxLkufWRAb52Fqm  0.750679  ...   35.06903   -91.90785
9  AVwdV1VNIN2L1WUfuv2g  0.733311  ...  21.908293  -159.47493

[10 rows x 7 columns]
```

2. Sample output for content based which considers the reviews made by user for the same category of hotels.



```
                 hotel_id     score  ...   latitude   longitude
0  AVwdptatkufWRAb57_Vx  0.938933  ...  43.515397  -96.776096
1  AVwc2rCTkufWRAb5zt6J  0.945376  ...  40.579353  -122.35721
2  AWBwmDwiIxWefVJwvAe-  0.944941  ...   36.84389   -76.18672
3  AVwcuaVJkufWRAb5yWBw  0.940732  ...   42.04762    -80.0835
4  AVwc2_cd_7pvs4fzzttM  0.938120  ...   38.53535   -76.58334
5  AVwdNApbIN2L1WUftTMk  0.960308  ...   43.02974   -91.13097
6  AVweLNkgIN2L1WUf2pN7  0.938761  ...   40.27362   -76.81537
7  AVwcqZ3LIN2L1WUfnf7-  0.951278  ...    43.8715    -91.2134
8  AVwdpk-ckufWRAb579_P  0.934491  ...  28.044954  -82.425605
9  AVwcnnZvkufWRAb5xMaa  0.942944  ...   40.42074   -104.7723
```

3. Sample output for collaborative filtering when we use user id from the dataset on ALS get recommendation function.



```
:  1  sdf.show()

VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

+----+--------------------+------------------+--------------------+-----------------+--------------------+---------+-
---------+
|  Id|            hotel_id|            rating|                name|   average_rating|          categories| latitude|
longitude|
+----+--------------------+------------------+--------------------+-----------------+--------------------+---------+-
---------+
|  27|AVwckL-d_7pvs4fzwg58|  4.70231294631958|    The Carneros Inn|              5.0|        Hotel,Hotels|38.255101|-
122.33346|
| 430|AVweeQx7_7pvs4fzDTn7|2.6843972206115723|Best Western Airp...|              4.0|Hotels,Lodging,Ho...| 32.49094|
-92.05335|
| 350|AVweNSZkIN2L1WUf27Ye|2.6791014671325684|Microtel Inn & Su...|4.916666666666667|Hotels and motels...|  41.4512|
-75.6368|
| 115|AVwcrY5MkufWRAb5x1nm|2.6437511444091797|Econo Lodge Inn &...|              3.5|Hotel,Hotels,Lodg...| 41.07905|
-75.77149|
|1418|AWE2I5Nv3-Khe5l_fxSI| 2.562448263168335|  Hampton Inn Odessa|              5.0|Hotel,Hotels,Lodg...| 31.88892|-
102.33335|
|1268|AVweFB7p_7pvs4fz_18I| 2.546880006790161|Hampton Inn Suite...|              5.0|Hotels,Corporate ...| 34.17064|
-97.16348|
| 303|AVwd1Zo5ByjofQCxs7DO|2.4529221057891846|Comfort Suites-elgin|              4.0|Hotels,Hotels and...|42.093964|
-88.33668|
|1655|AVwclphkIN2L1WUfmqLw| 2.414456367492676|     Ojai Rancho Inn|              5.0|Hotel,Hotels and ...|  34.4445|
-119.2532|
| 961|AVwdYrkeIN2L1WUfvQ5Y|2.3984217643737793|La Quinta Inn & S...|             4.25|Office and loft b...| 32.67507|
-97.0321|
|1758|AVwc5DdcIN2L1WUfqBcg| 2.369774103164673|Crossland Economy...|              4.0|Hotels,Lodging,Co...| 32.52805|
-93.69846|
+----+--------------------+------------------+--------------------+-----------------+--------------------+---------+-
---------+
```

Challenges

1. For the content based recommendation we have used user specific recommendation based on category of hotels and word specific recommendation. The major challenges we faced was how to devise or what features should we consider for the recommendation based on category of hotels.

2. The other challenge we faced was, how to optimize the time execution time when creating new schema from the existing corpus.
3.
4. Another challenge we had in citation network dataset was the improper data file and lots of missing values in abstract of the paper due to which we could not build a proper content based recommendation system based on textual data.

Tools and Technology

- Amazon EMR (Spark 2.3.2) (1 master 2 worker nodes) for running the program
- Amazon S3 (Storing data)
- Apache Spark -Pyspark
- Jupyter Notebook

Task Division

Project has been accomplished with the help of the inputs of all team members and team work.

Data Pre processing	Pratik Parekh, Srishti Tiwari
Applying SQL queries to make data suitable for the algorithms	Pratik Parekh, Tannu Singh
Forming Processing pipeline	Pratik Parekh, Srishti Tiwari
Implementing and finding similarity score amongst the hotels	Pratik Parekh, Tannu Singh
Visualization and data Exploration	Tannu Singh, Srishti Tiwari
Implementation of Collaborative Filtering Algorithm	Pratik Parekh, Tannu Singh, Srishti Tiwari

Result Evaluation and Providing recommendations for a user.	Tannu Singh, Srishti Tiwari
Implementation of content based filtering algorithm	Pratik Parekh, Tannu Singh, Srishti Tiwari
Implementation on AWS cluster and EMR notebook and UI	Srishti Tiwari
Result Evaluation and Providing recommendations for a use content based	Pratik Parekh

Future scope

1. The implementation can be extended to perform recommendation based on sentiment analysis and topic modeling of the reviews and by the positive reviews of the same topic which our model fails to address as implemented in Music Recommendation system by R. L. Rosa, D. Z. Rodríguez, G. Bressan [4]
2. With this dataset we can create a hybrid recommendation system where we can combine ratings and content of the reviews to get the personalized recommendation for users.
3. We can also use the geolocation information in the dataset to predict the rating of a hotel in a particular area as described in the methodology section of R. L. Rosa, D. Z.
4. Rodríguez, G. Bressan [3], of social network recommendation

References

[1]Content-Based Recommendation Systems, Michael J. Pazzani

Daniel Billsus

[2]Chen T, Han W (2007) Content recommendation system based on private dynamic user profile. Mach Learn 4: 2112–2118

[3] Betim Berjani , Thorsten Strufe, A recommendation system for spots in location-based online social networks, Proceedings of the 4th Workshop on Social Network Systems, p.1-6, April 10-13, 2011, Salzburg, Austria

[4] R. L. Rosa, D. Z. Rodríguez, G. Bressan, "Music recommendation system based on user's sentiments extracted from social networks", *IEEE Trans. Consumer Electron.*, vol. 61, no. 3, pp. 359-367, 2015.

[5] https://www.elenacuoco.com/2016/12/22/alternating-least-squares-als-spark-ml/