Vacation Planner

Vivek Bharti
NYU Tandon School Of Engineering
Computer Science
Brooklyn, New York
Email: vb1275@nyu.edu

Roshan Sridhar
NYU Tandon School Of Engineering
Computer Engineering
Brooklyn, New York
Email: rs5788@nyu.edu

Abstract

The Vacation Planner is a system which leverages Big Data technologies and Machine Learning techniques to generate data-driven recommendations hotels/resorts and places to users based upon other users with similar tastes. The system uses Spark for data handling and transformations. Since the system uses Big Data technologies it can handle large volumes of data. To make the system scalable and robust, we have used NYU's HPC which is a distributed environment to handle data. The system uses User-based Collaborative Filtering techniques to get recommendations. The system shows it's top 10 recommendations to the users and also queries Airfare Reports to provide the average airfare for all those destinations to make the the user make his final decision. We also use R and Tableau to generate impactful visualizations.

Introduction

Vacation Planner will help the user plan a vacation that would be most fit as per their likes and dislikes. The system uses these ratings on hotels/resorts and places of users already in our data to understand the tastes of a neighborhood of users who want to plan their vacation. Using this collective information data, the system generates recommendations and provides it's top 10 to the user. These recommendations are the predictions that the user might love to visit and has never visited before, based on their previous visits and likes and dislikes of similar users.

Motivation

We would like to help the user obtain suggestions of resorts/places to visit which they otherwise couldn't get just by searching on a search engine or asking people, as those are anecdotal and not reliable. We wanted to suggest some new places to the user which is unvisited but they would probably like. We wanted to leverage Big Data and Machine Learning techniques to tackle this trivial but helpful issue. This project quite challenging because there was not many projects available on the internet and we saw that as an opportunity to learn.

Objectives

- To suggest new hotels/resorts and places to visit for a user based upon their previous ratings and using ratings data from other users to generate recommendations.
- Process data using Spark and to make the system scalable to any extent
- Generate recommendations using user-based collaborative filtering technique
- Provide the top 10 recommendations to the users

Data

Primarily we have 4 datasets used in this system. The data files obtained are all comma separated values:

- Ratings: This Data has the ratings given to the hotels/resorts by the users. Ratings are from 1 to 5. 5 being the most liked. There are roughly 220k ratings in this data which has been used to train our user-based collaborative filtering model to generate recommendations.

- Customer Master: This is the demographic data of the users. There are roughly 10k users in the system who have provided their ratings for the hotels/resorts.

- Hotel Master: The master table for hotels/resorts. This table has data of roughly 260 hotels. The data has hotel address, classification based on types of hotels etc. Below image shows the hotels on the US map.

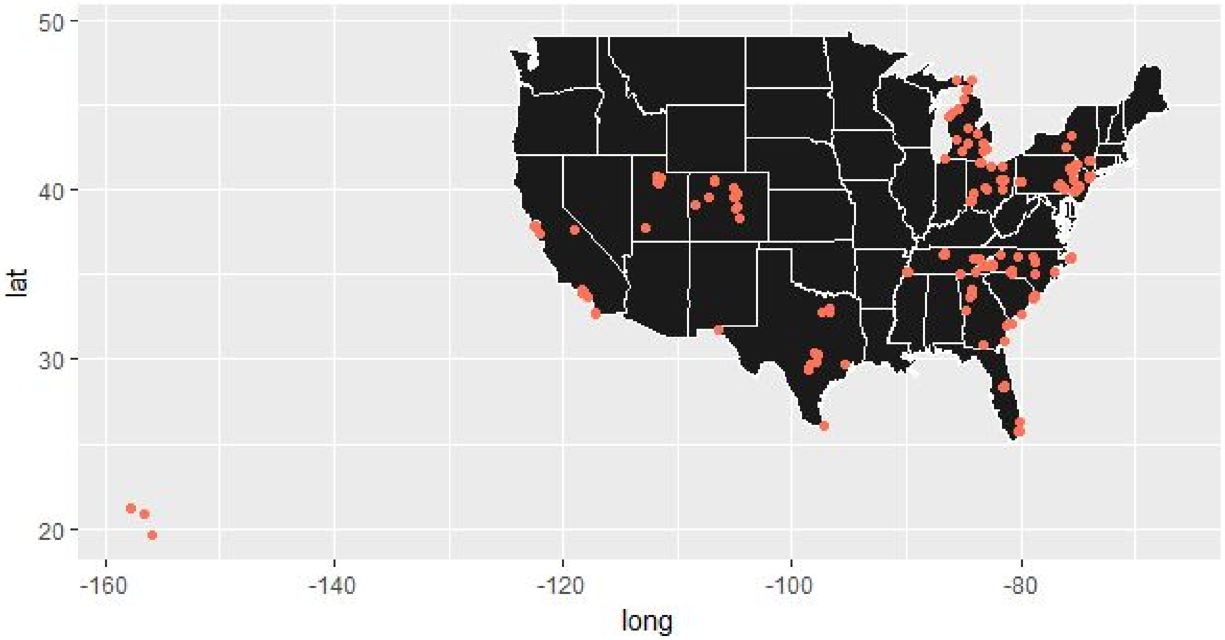


Figure 1: Location of the hotels in the schema

- Airfare Data: Consumer airfare report data. Contains distance average market airfare. Data collection from 1996-2017 is present but for our system, we have relatively recent data - 2012 onwards

Technologies Used

- Distributed Computing: The data was stored in the HDFS of the NYU Hadoop cluster - Dumbo. Spark on the cluster read files from the HDFS. We chose to use this resource to make our implementation scalable.

- Apache Spark: Spark 2.1.0 has been used to transform and manipulate data and the system is designed in such a way that it is scalable and can very elegantly handle large volumes of data. This system is developed in such a fashion that in future as the user base and the hotel data will keep on increasing the system can still handle the data.

- R/Tableau: R and Tableau have been used to generate visualizations.

- Spark ML: Machine learning part of the project is handled by Spark ML. ALS model present in spark MLlib library has been used to perform user-based collaborative filtering.

Architecture Design

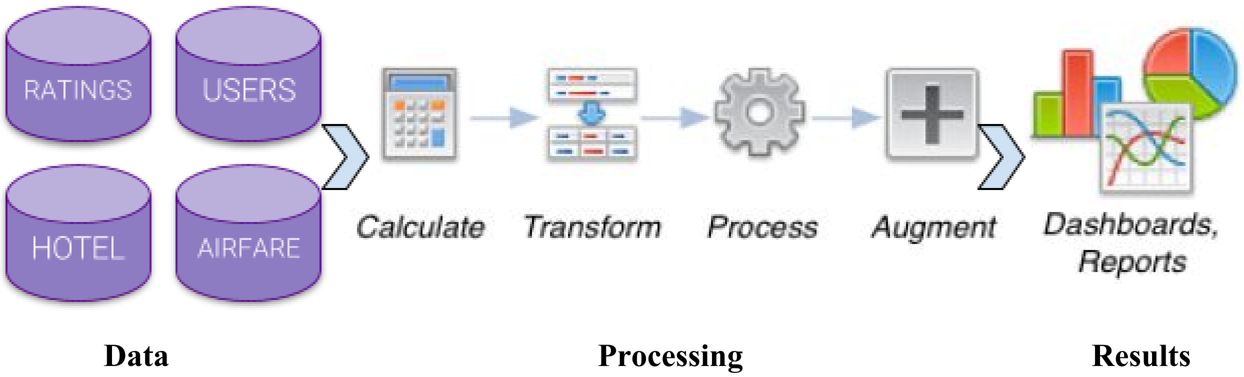


Figure 2: Pipeline of the system

Setup

Data is read into Spark DataFrames. Spark DataFrames turned out to be useful to prepare, manipulate and process the data. Spark has a good collection of functions to perform various transformations. The transformed data is passed into the scalable machine learning model that produced a prediction output. These predictions are joined with other datasets that contain more detailed information about hotels, distance from the hotel to the user and airfares.

The model used is called Collaborative Filtering. Collaborative Filtering is commonly used in recommendation systems. In Spark, the machine learning package uses the Alternating Least Squares(ALS) algorithm. This model is data intensive and computationally heavy. We use the Spark Machine Learning library to help tackle this. This model takes in User IDs, Item IDs and the user's ratings for that item to create a set of latent factors that can predict recommendations for new users. The model creates multiple neighborhoods of users that have similar likes and dislikes. This neighborhood of users consists of groups of like-minded users. When the model encounters an item that the user has not rated, it uses this neighborhood data to provide an estimate using other user's ratings on the same item.

We first prepared the rating data such that we can pass it through this ALS model. Once the model trains on this data, it can output a predict a calculated rating provided the User ID and an Hotel ID that the user has not rated. This predicted rating is calculated based on ratings of other users from the neighborhood of the selected user for the same hotel.

To complete the system, we provided a method to feed in new user data. This data included the user information and their hotel preferences. These hotel preferences of the new user are merged with the ratings and fed into the model to assign a neighborhood to this new user. We chose to obtain top 10 predictions of hotels for this new user for which we created a manual function. This function involves passing a row of two elements, the User ID and an unvisited Hotel ID for **every** hotel the User has not rated. This is such that the model can provide a predicted rating for the unrated hotels. This output containing the new predicted ratings are sorted in descending order on the rating values and the top 10 records are chosen to be returned.

Once the predicted Hotel IDs are returned, we join the prediction IDs with two datasets 'Consumer Airfare Report Data' and 'Hotel Schema'. This was performed with the help of Spark. This step involves aggregation of all four datasets. This helps us provide the following information with the output namely, the user we are catering to, the details of the hotels we are suggesting this user, airfares and distances between the location of the user and the all the location of the hotel suggestions respectively.

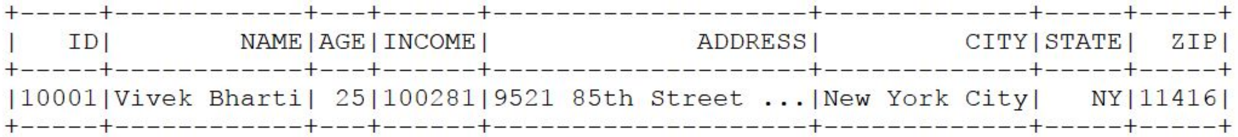
Code

The code involves working with data in PySpark. This involved data manipulation eg. slicing and dicing of various datasets mentioned above, using PySpark. Further, using the Collaborative Filtering Model using ALS present in spark MLlib library to get the final recommendations. The code is provided in an Apache Zeppelin notebook for ease of use and viewing format. It is attached with this report and has been uploaded on NYU Classes.

Visualizations, Results, and Evaluation

- **Test sample #1**

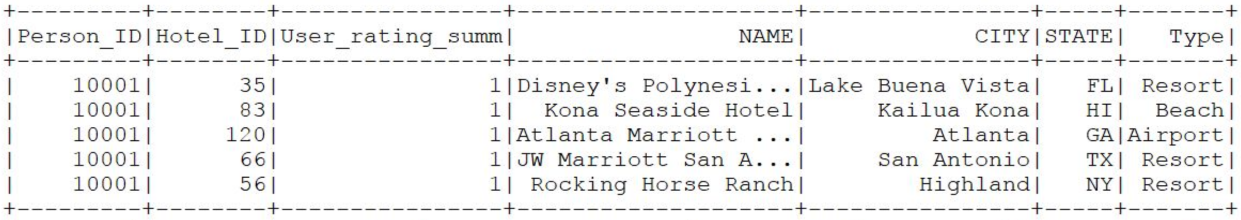
We first add a new user 10001 along with ratings for a few hotels.



```
+-----+------------+---+------+--------------------+-------------+-----+-----+
|   ID|        NAME|AGE|INCOME|             ADDRESS|         CITY|STATE|  ZIP|
+-----+------------+---+------+--------------------+-------------+-----+-----+
|10001|Vivek Bharti| 25|100281|9521 85th Street ...|New York City|   NY|11416|
+-----+------------+---+------+--------------------+-------------+-----+-----+
```

Table 1: New customer entry for evaluation.

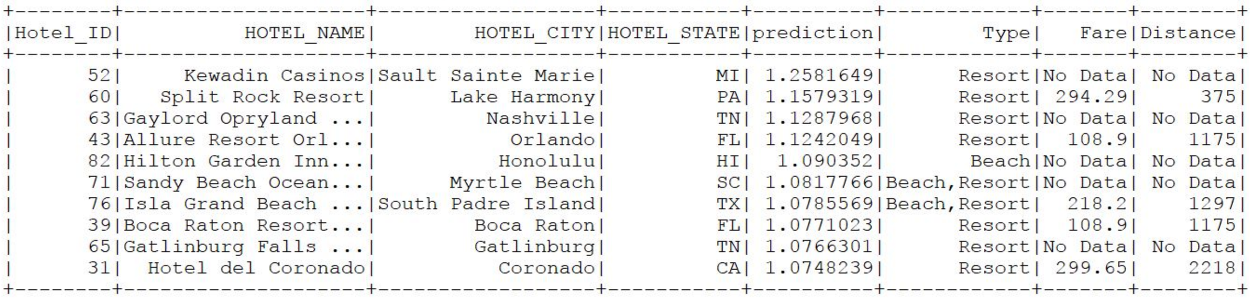
We assume that the user primarily likes resorts, as shown below. This data is provided to the rating data over which the model now retrains.



```
+---------+--------+----------------+--------------------+----------------+-----+-------+
|Person_ID|Hotel_ID|User_rating_summ|                NAME|            CITY|STATE|   Type|
+---------+--------+----------------+--------------------+----------------+-----+-------+
|    10001|      35|               1|Disney's Polynesi...|Lake Buena Vista|   FL| Resort|
|    10001|      83|               1|  Kona Seaside Hotel|     Kailua Kona|   HI|  Beach|
|    10001|     120|               1|Atlanta Marriott ...|         Atlanta|   GA|Airport|
|    10001|      66|               1|JW Marriott San A...|     San Antonio|   TX| Resort|
|    10001|      56|               1| Rocking Horse Ranch|        Highland|   NY| Resort|
+---------+--------+----------------+--------------------+----------------+-----+-------+
```

Table 2: Ratings provided by this customer which serves as training data.

After training over the data, the engine has now computer a neighborhood



```
+--------+--------------------+------------------+-----------+----------+------------+-------+--------+
|Hotel_ID|          HOTEL_NAME|        HOTEL_CITY|HOTEL_STATE|prediction|        Type|   Fare|Distance|
+--------+--------------------+------------------+-----------+----------+------------+-------+--------+
|      52|     Kewadin Casinos|Sault Sainte Marie|         MI| 1.2581649|      Resort|No Data| No Data|
|      60|   Split Rock Resort|      Lake Harmony|         PA| 1.1579319|      Resort| 294.29|     375|
|      63|Gaylord Opryland ...|         Nashville|         TN| 1.1287968|      Resort|No Data| No Data|
|      43|Allure Resort Orl...|           Orlando|         FL| 1.1242049|      Resort|  108.9|    1175|
|      82|Hilton Garden Inn...|          Honolulu|         HI|  1.090352|       Beach|No Data| No Data|
|      71|Sandy Beach Ocean...|      Myrtle Beach|         SC| 1.0817766|Beach,Resort|No Data| No Data|
|      76|Isla Grand Beach ...|South Padre Island|         TX| 1.0785569|Beach,Resort|  218.2|    1297|
|      39|Boca Raton Resort...|        Boca Raton|         FL| 1.0771023|      Resort|  108.9|    1175|
|      65|Gatlinburg Falls ...|        Gatlinburg|         TN| 1.0766301|      Resort|No Data| No Data|
|      31|  Hotel del Coronado|          Coronado|         CA| 1.0748239|      Resort| 299.65|    2218|
+--------+--------------------+------------------+-----------+----------+------------+-------+--------+
```

Table 3: Top 10 predictions provided by the model.

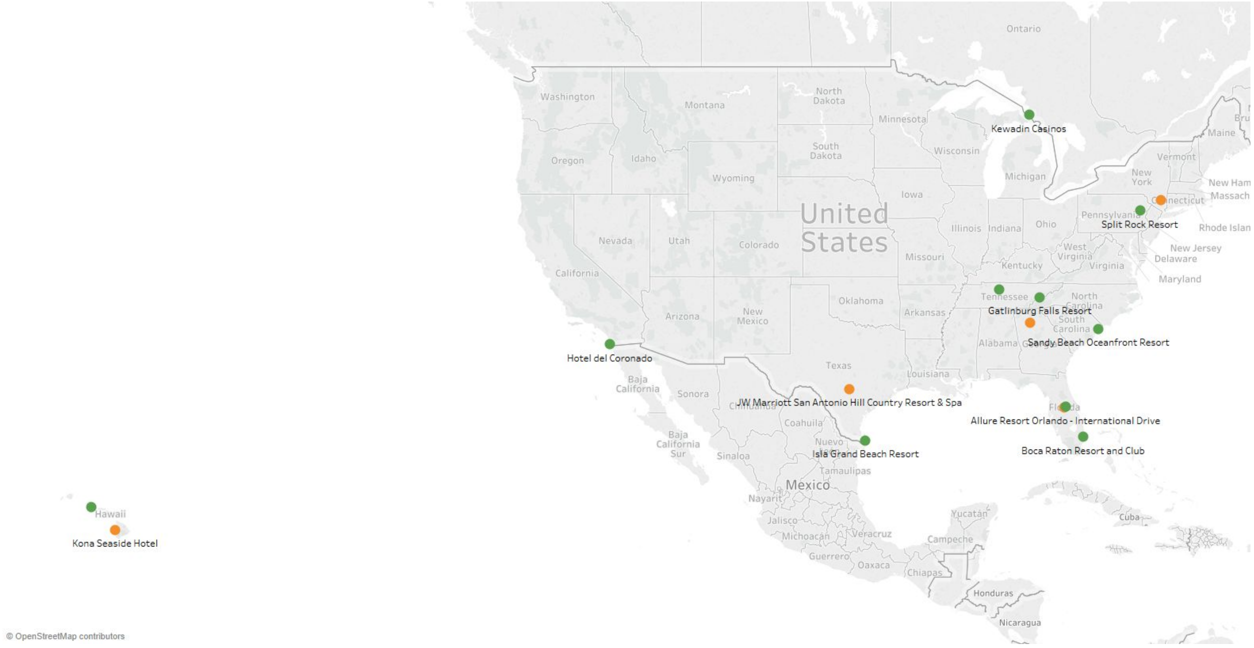
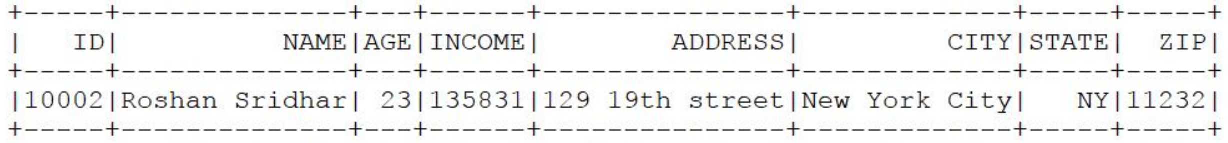

Kewadin Casinos
Split Rock Resort
Gatlinburg Falls Resort
Sandy Beach Oceanfront Resort
Hotel del Coronado
JW Marriott San Antonio Hill Country Resort & Spa
Allure Resort Orlando - International Drive
Isla Grand Beach Resort
Boca Raton Resort and Club
Kona Seaside Hotel
United States
Mexico
© OpenStreetMap contributors

Figure 3: Visualization of the geolocations of user's hotels and predictions

- **Test sample #2**

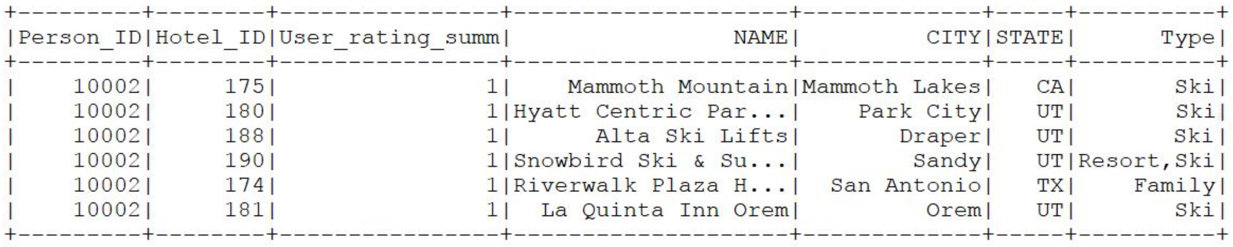
We add a new user 10002 along with ratings for a few hotels.



```
+-----+--------------+---+------+---------------+-------------+-----+-----+
|   ID|          NAME|AGE|INCOME|        ADDRESS|         CITY|STATE|  ZIP|
+-----+--------------+---+------+---------------+-------------+-----+-----+
|10002|Roshan Sridhar| 23|135831|129 19th street|New York City|   NY|11232|
+-----+--------------+---+------+---------------+-------------+-----+-----+
```

Table 4: New customer entry for evaluation.

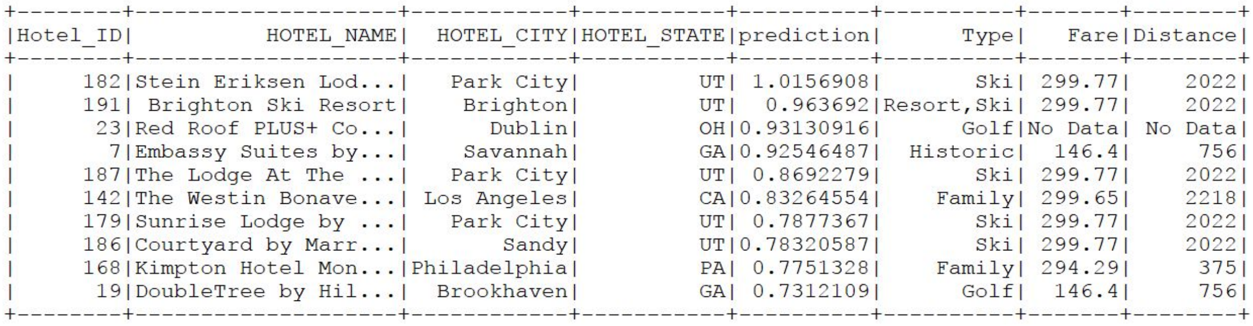
We assume that the user primarily likes resorts, as shown below. This data is provided to the rating data over which the model now retrains.



```
+---------+--------+----------------+--------------------+-------------+-----+----------+
|Person_ID|Hotel_ID|User_rating_summ|                NAME|         CITY|STATE|      Type|
+---------+--------+----------------+--------------------+-------------+-----+----------+
|    10002|     175|               1|    Mammoth Mountain|Mammoth Lakes|   CA|       Ski|
|    10002|     180|               1|Hyatt Centric Par...|    Park City|   UT|       Ski|
|    10002|     188|               1|      Alta Ski Lifts|       Draper|   UT|       Ski|
|    10002|     190|               1|Snowbird Ski & Su...|        Sandy|   UT|Resort,Ski|
|    10002|     174|               1|Riverwalk Plaza H...|  San Antonio|   TX|    Family|
|    10002|     181|               1|  La Quinta Inn Orem|         Orem|   UT|       Ski|
+---------+--------+----------------+--------------------+-------------+-----+----------+
```

Table 5: Ratings provided by this customer which serves as training data.

After training over the data, the engine has now computer a neighborhood



```
+--------+--------------------+------------+-----------+----------+----------+-------+--------+
|Hotel_ID|          HOTEL_NAME|  HOTEL_CITY|HOTEL_STATE|prediction|      Type|   Fare|Distance|
+--------+--------------------+------------+-----------+----------+----------+-------+--------+
|     182|Stein Eriksen Lod...|   Park City|         UT| 1.0156908|       Ski| 299.77|    2022|
|     191| Brighton Ski Resort|    Brighton|         UT|  0.963692|Resort,Ski| 299.77|    2022|
|      23|Red Roof PLUS+ Co...|      Dublin|         OH|0.93130916|      Golf|No Data| No Data|
|       7|Embassy Suites by...|    Savannah|         GA|0.92546487|  Historic|  146.4|     756|
|     187|The Lodge At The ...|   Park City|         UT| 0.8692279|       Ski| 299.77|    2022|
|     142|The Westin Bonave...| Los Angeles|         CA|0.83264554|    Family| 299.65|    2218|
|     179|Sunrise Lodge by ...|   Park City|         UT| 0.7877367|       Ski| 299.77|    2022|
|     186|Courtyard by Marr...|       Sandy|         UT|0.78320587|       Ski| 299.77|    2022|
|     168|Kimpton Hotel Mon...|Philadelphia|         PA| 0.7751328|    Family| 294.29|     375|
|      19|DoubleTree by Hil...|  Brookhaven|         GA| 0.7312109|      Golf|  146.4|     756|
+--------+--------------------+------------+-----------+----------+----------+-------+--------+
```

Table 6: Top 10 predictions provided by the model.

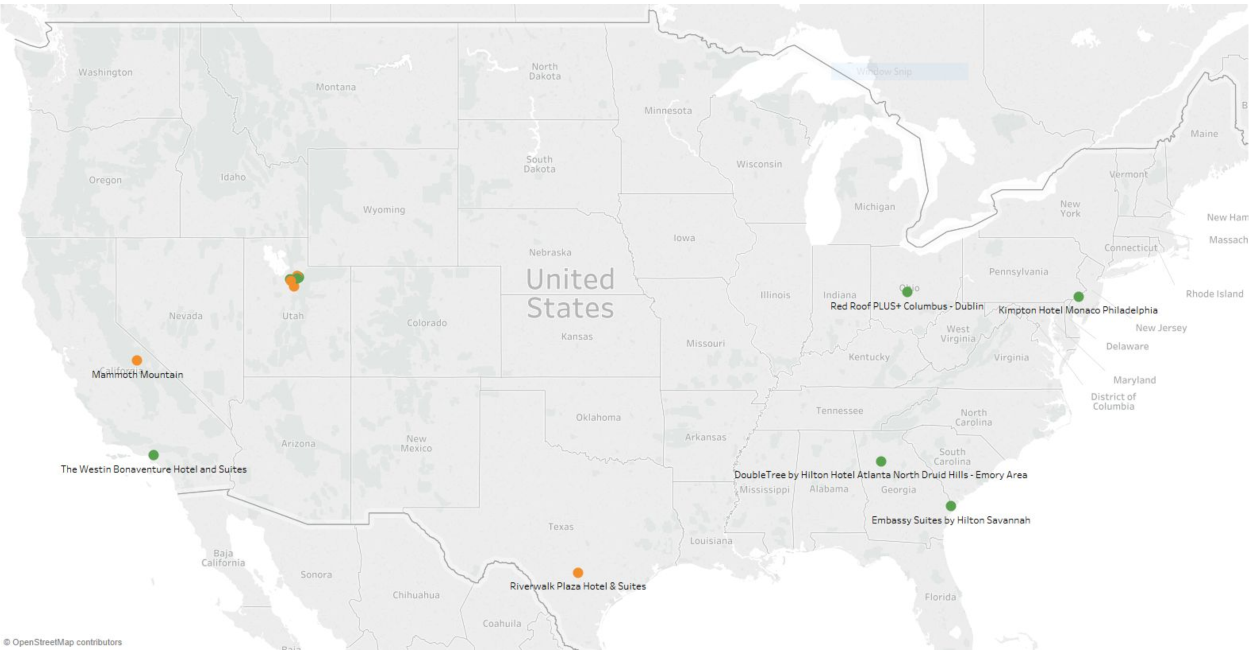


Figure 4: Visualization of the geolocations of user's hotels and predictions

Conclusion and Future Work

The system provides good predictions and suggestions to users. Apache Spark helps us scale the system to ingest and output larger data. User-based collaborative filtering helps to provide great recommendations provided the data is good. The predictions suffer when there is less data or due to cold-start i.e. if there are no ratings by the user due to which it cannot find similar users to group this user with. In the future, other consumer data like the age and income can be utilized to create more sophisticated neighborhoods i.e. like-minded groups of users and thus, provide better recommendations.

References

- Spark documentation: https://spark.apache.org/docs/
- Domestic Airline Consumer Airfare Report: https://www.transportation.gov/policy/aviation-policy/domestic-airline-consumer-airfare-report
- Factual Data - US Hotels (public sample): https://www.factual.com/data/t/hotels-us