Holiday Destination Recommender System

Introduction

Travel agencies make money by organizing travels for their clients and taking a fee on each travel organized. People who want to go on holiday do not always know precisely where they want to go. They might just have a broad idea of what the travel should look like. A recommender system could help the travel agent by providing ideas of trips to suggest to its clients. A recommender system could also be offered to clients on a website. In this case, I'm thinking of a recommender system where the client is asked to rate a few destinations that he knows and is then provided with a list of suggestions based on destinations he has liked in the past. Destinations would be characterized by the existence of certain categories of venues. The audience for this project would be the owners/managers of travel agencies looking to improve the services they offer to their clients and ultimately to improve their profits.

Data

To explore this idea, I used two sources of data: Wikipedia and Foursquare.

On Wikipedia, I found a list of cities ranked by the number of international visitors (https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors). These will represent the travel options that my travel agency can recommend.

For each of these cities, I used a geolocator to determine their latitude and longitude.

Based on the name, the latitude, and the longitude of the cities, I retrieved venues from Foursquare.

I used the broadest possible categorization of venues provided by Foursquare and obtained nine categories:

- arts_entertainment;
- food;
- nightlife;
- parks_outdoors ;
- shops;
- travel;
- building;
- education; and
- event.

For each city, I summed the occurrence of each category of venues.

I observed that Foursquare did not find many venues for some cities and that some categories included only very few venues. Pursuant to this observation, I dropped cities for which Foursquare did not retrieve 50 venues in a radius of 500 meters and I dropped the three categories for which I only had very few venues (building, education and event).

Then I transformed the venue category variables into quintiles (a city is in the top quintile for arts_entertainment if it is among the 20 percent of cities with the most venues of this type). This standardization is important for the recommender system because the value of the variables will affect the determination of the users' preferences. A normalization procedure (removing the mean and dividing by the standard deviation does not work here because the recommender system does not work well with negative values).

I will call the result of these transformations the **VenueTable**. The top five rows of this dataset are as follows:

	City	arts_entertainment	food	nightlife	parks_outdoors	shops	travel
0	Amman	1	5	4	2	1	1
1	Amsterdam	3	2	5	1	4	1
2	Athens	4	4	5	1	1	1
3	Auckland	5	4	1	2	3	2
4	Barcelona	2	3	4	5	1	3

Methodology

K-means

As a preliminary analysis, I will classify my cities into three clusters using K-means. The purpose of this analysis is to see if I am able to divide the cities into meaningful groups based my characterization of each city.

My conclusion from an earlier assignment in this module was that if we take into account very precise categories of venues (for example if we distinguish all possible sorts of restaurants) then we end up with clusters that are very difficult to interpret.

In this project, I have kept only six broad categories of venues to characterize each city and I think that it will be interesting to see if it enables a meaningful cluster analysis.

For the sake of completeness, the choice of a k-mean algorithm is dictated by the nature of the data and what I'm trying to achieve: I'm trying to classify unlabeled data. K-means seeks to minimize differences within groups and to maximize differences between groups.

Content-based recommender system

The main contribution of this project is to design a recommender system for holiday destinations.

This work is inspired from an earlier course on recommender systems where movies were recommended based on their genres.

In this case, holiday destinations will be recommended based on the venues that can be found in different cities.

First, I had to create an input user. I decided to give a rating between 1 and 5 to five cities that I have visited in the past:

	City	rating
0	Amsterdam	5
1	London	4
2	New York City	2
3	Paris	4
4	Toronto	1

To learn the users' preferences, I retrieved the data corresponding to those five cities (the quintile rankings for the different categories of venues):

	City	arts_entertainment	food	nightlife	parks_outdoors	shops	travel
1	Amsterdam	3	2	5	1	4	1
31	London	5	1	2	5	2	4
42	New York City	1	5	2	1	5	1
46	Paris	2	2	4	5	4	1
62	Toronto	4	4	1	3	2	1

Then, I defined the user profile by taking the dot product of the matrix above and the ratings assigned to each city. The user profile is a 6x1 array.

Then, I took the dot product of the **VenueTable** and the user profile (and scaled it by the sum of the elements in the user profile) to obtain an estimated rating for each destination.

The cities with the highest estimated ratings would be those that I would most likely appreciate based on my preferences (more precisely based on my ratins of five cities that I've already visited).

Results

K-means

After dividing my sample of cities into three clusters using the k-means algorithm, I calculated the centroid values by average the occurrence variables for each venue category and each cluster. The results were as follows:

	arts_entertainment	food	nightlife	parks_outdoors	shops	travel
Cluster Labels						
0	3.192308	31.192308	4.192308	2.384615	4.846154	3.192308
1	3.760000	22.440000	4.040000	2.360000	11.600000	4.280000
2	8.500000	20.777778	4.388889	5.111111	4.666667	4.388889

Cluster 0 is characterized by the highest number of food venues.

Cluster 1 is characterized by a high number of shopping venues.

Cluster 2 is characterized by a high number of arts & entertainment venues and by a high number of parks & outdoors venues.

Content-based recommender system

Based on my ratings of five cities and the venues in these cities, my user profile was as follows:

```
arts_entertainment    49
food                  36
nightlife             54
parks_outdoors        50
shops                 56
travel                28
```

This suggests that I give more importance to arts & entertainment, nightlife, parks & outdoors, and shops, but less importance to food and travel venues when comparing travel destinations.

Based on this user profile and the venues in my list of cities, my top 10 recommendations are as follows:

	City	arts_entertainment	food	nightlife	parks_outdoors	shops	travel	Estimated_Ratings
67	Washington D.C.	4	1	4	5	3	4	3.58242
51	Saint Petersburg	5	1	2	5	4	3	3.46886
39	Milan	2	1	4	4	5	4	3.45055
56	Sofia	5	1	5	5	2	1	3.44689
66	Warsaw	5	2	4	3	3	3	3.42491
46	Paris	2	2	4	5	4	1	3.25275
59	Taipei	1	2	4	3	5	4	3.21978
38	Mexico City	5	1	3	3	3	4	3.1978
15	Dublin	4	2	5	1	4	2	3.17949
31	London	5	1	2	5	2	4	3.16117

Discussion

K-means

The results of the clustering analysis are quite interesting. They provide distinct sets of destinations for travelers interested in food (cluster 0), shopping (cluster 1), and arts & entertainment and parks & outdoors (cluster 2).

Such analysis could be interesting for a travel agent seeking to make travel suggestions to potential clients after asking them their main centers of interest.

Content-based recommender system

The top 10 recommendations align well with the user profile. Washington DC is the top recommendation. This destination ranks high with respect to arts_entertainment, nightlife, parks_outdoors; and shops, which are the four attributes with the highest weights in my user profile. Food venues have a low rank for Washington DC which is consistent with low weight of this attribute in my user profile. Travel venues have a high rank for Washington DC despite the low weight of this attribute in my user profile. This is likely due to a positive correlation between the occurrence of travel venues and the occurrence of other venues which are given more weight in my user profile.

Conclusion

This report summarizes a preliminary work in view of developing a recommender system for holiday destinations. Such work could represent significant business value for travel agencies. However, more work would need to be done, especially on finding the best way to characterize a destination in terms of existing venues.