Capstone Project – Report – Andrea Tamburri

1. Business Problem

1.1 Introduction

The total value of the global travel and tourism industry is one of the world's largest industries with a global economic contribution of over 7.6 trillion U.S. dollars.
This is a very complex and heterogeneous market, with tons of different businesses all around the world, as, for example:

- Aribnb: an online marketplace for arranging or offering lodging, homestays or tourism experience, with revenues of over 2.6 billion
- Booking.com: a metasearch engine for lodging reservation, with total revenues of over 14.5 billion
- Ryanair: a competitive cost airline, with total revenue amount of over 7.2 billion
- Yelp: a business directory service and crowd-sourced review forum, with total revenue amount of 1 billion

Those are just few examples of what the tourism market as a whole can offer, and how can be remunerative.

1.2 Business Problem

The main idea of the whole project that it's going to be developed, is to leverage the information obtained by Foursquare about venues to cluster some of the major cities all around the world in order to create a **content based recommender system:** for example, talking about Airbnb, **if a customer decide to rent an apartment in Paris, and he liked it, he is possibly willing to rent apartments in cities similar to Paris** (in our case similarity is "just" given by venues distribution) that will be offered in the main page. **This system improves the customer experience, increasing their purchases or, in the other way around, increasing the revenue of the company.**
In this project I'll focus on "selling" them the "cluster matrix" of the whole recommender system. Broadly speaking, in a real case scenario, I'd probably go for one of the two option:

- Selling to the companies a more complete cluster matrix, in order to have a broader point of view on each location and exploit in dept the advantages the reccommender systems: different clusters, based on, for example, venues distribution (the one that'll do), geolocation, life cost, etc.
- Selling, as a Saas, the whole recommender system: in this case, the cluster matrix as to be integrated with the information about customers behaviours (e.g., taking always Airbnb as example, the ratings about their experience in different localities) in order to create the "user profile" and, finally, in combination with the cluster matrix, the "recommendation matrix" that will be used to recommend to that customer the locations with the highest scores.

2. Data acquisition and cleaning

2.1 Data source

There are three major data sources used for this project:

- A Yelp businesses revue dataset, containing veues reviews from all around the world
- The "World Cities Table", in which you can find the geolocation of the most important cities from all around the world
- Foursquare, in which we can find more information about venues location, category, ratings

Yelp Dataset

A great repository of venues reviews, geolocation and users that actually made the reviews can be easily found in two Kaggle dataset that you can find here.
In our case this consists on two datasets:

- *Yelp_academic_dataset_business.json*
- *Yelp_academic_dataset_review.json*

The first one contains information about venues reviewed, as the address, the city in which is it located, the average rating etc.

```
RangeIndex: 192609 entries, 0 to 192608
Data columns (total 14 columns):
address         192609 non-null object
attributes      163773 non-null object
business_id     192609 non-null object
categories      192127 non-null object
city            192609 non-null object
hours           147779 non-null object
is_open         192609 non-null int64
latitude        192609 non-null float64
longitude       192609 non-null float64
name            192609 non-null object
postal_code     192609 non-null object
review_count    192609 non-null int64
stars           192609 non-null float64
state           192609 non-null object
```

Table1: Yelp_academic_dataset_business.json structure

The dataset has been cleaned of all the useless columns in order to have just the *business_id* and the *city* of the business. The purpose of the cleaning is because we only need to connect the information about the city of the venue with the review made for that venue.

The latter can be found in the second dataset, the *yelp_academic_dataset_review.json.*
This table contains the stored reviews of the previous businesses: the original dataset was bigger than 5GB, so it was impossible to upload entirely on IBM Watson Studio. In order to upload as much as possible, the dataset has been divided on chunks of 500K rows each. The total dataset uploaded contains 5M records of businesses reviews.

business_id	cool	date	funny	review_id	stars	text	useful	user_id
ujmEBvifdJM6h6RLv4wQIg	0	2013-05-07 04:34:36	1	Q1sbwvVQXV2734tPgoKj4Q	1	Total bill for this horrible service? Over $8G...	6	hG7b0MtEbXx5QzbzE6C_VA
NZnhc2sEQy3RmzKTZnqtwQ	0	2017-01-14 21:30:33	0	GJXCdrto3ASJOqKeVWPi6Q	5	I *adore* Travis at the Hard Rock's new Kelly ...	0	yXQM5uF2jS6es16SJzNHfg
WTqjgwHlXbSFevF32_DJVw	0	2016-11-09 20:09:03	0	2TzJjDVDEuAW6MR5Vuc1ug	5	I have to say that this office really has it t...	3	n6-Gk65cPZL6Uz8qRm3NYw
ikCg8xy5JIg_NGPx-MSIDA	0	2018-01-09 20:56:38	0	yi0R0Ugj_xUx_Nek0-_Qig	5	Went in for a lunch. Steak sandwich was delici...	0	dacAIZ6fTM6mqwW5uxkskg
b1b1eb3uo-w561D0ZfCEiQ	0	2018-01-30 23:07:38	0	11a8sVPMUFtaC7_ABRkmtw	1	Today was my second out of three sessions I ha...	7	ssoyf2_x0EQMed6fgHeMyQ

Table2: *Yelp_academic_dataset_reviews.json* raw table

The final table has been made joining the two Yelp datasets in order to put together all the necessary information to build our content-based recommender system (in particular this table we'll let us build the user's profile table: the one that explains how much the given user likes that cluster of cities).

user_id	city	stars_x
hG7b0MtEbXx5QzbzE6C_VA	Las Vegas	1
RBXSJA372ilErzNwz0jXvQ	Las Vegas	4
x3brMMbJrAW9PwW5A6YL5w	Las Vegas	1
Skzdl0sWhW88525a1vr59g	Las Vegas	1
3Y25VDfnQVcuc33T-U3Z6A	Las Vegas	5

Table3: final table extracted from Yelp dataset

The table contains just the information about the review's ratings made by the users in the city of the business. **That's because we are assuming that the city's rating of one particular user, is the mean of all the ratings made in that city.** The last step of preprocessing is to obtain for each user the mean rating of each city he reviewed.

user_id	city	stars_x
---1lKK3aKOuomHnwAkAow	Henderson	3.666667
---1lKK3aKOuomHnwAkAow	Las Vegas	3.934211
---1lKK3aKOuomHnwAkAow	North Las Vegas	3.666667
---94vtJ_5o_nikEs6hUjg	Phoenix	5.000000
---PLwSf5gKdIoVnyRHgBA	Phoenix	5.000000

Table4: Final table with mean calculated grouping by *user_id* and *city*

World Cities Table

Information about city geolocation can be easily found from the web: check the website at https://simplemaps.com/data/world-cities this interesting table about the biggest cities from all around the world.

As you can see below, this table, include:

- City name
- Latitude
- Longitude
- Country of the city
- Population of the city
- Etc.

city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
Tokyo	Tokyo	35.6850	139.7514	Japan	JP	JPN	Tōkyō	primary	35676000	1392685764
New York	New York	40.6943	-73.9249	United States	US	USA	New York		19354922.0	1840034016
Mexico City	Mexico City	19.4424	-99.1310	Mexico	MX	MEX	Ciudad de México	primary	19028000	1484247881
Mumbai	Mumbai	19.0170	72.8570	India	IN	IND	Mahārāshtra	admin	18978000	1356226629
São Paulo	Sao Paulo	-23.5587	-46.6250	Brazil	BR	BRA	São Paulo	admin	18845000	1076532519
Delhi	Delhi	28.6700	77.2300	India	IN	IND	Delhi	admin	15926000	1356872604
Shanghai	Shanghai	31.2165	121.4365	China	CN	CHN	Shanghai	admin	14987000	1156073548
Kolkata	Kolkata	22.4950	88.3247	India	IN	IND	West Bengal	admin	14787000	1356060520
Los Angeles	Los Angeles	34.1139	-118.4068	United States	US	USA	California		12815475.0	1840020491

Table5: World Cities Table

The purpose of this table is to obtain the geolocation of the cities, in order to call the Foursquare API.
For computational matters in cloud (that's a capstone project made in the lite Watson Studio, with limited computational capacity), we'll take the top 30 most populated cities in the U.S. that match with the yelp dataset in order to cluster them through their venue distribution.
This will give us the cluster matrix (a table containing information about the clusters in which a given city is) that will later be combined with the user's profile in order to build the recommendation matrix.

Foursquare

We'll do some regular call at the Foursquare API to "explore" the venues around all the selected cities from the previous table.
The purpose is to obtain a table with a record for each venues of the selected cities, containing information about:

- Venue name
- Venue latitude and longitude
- **Venue category** (e.g. coffee shop or bank, **that uniquely describe a venue**)

The next step is to do some data preprocessing, like one hot encoding (in fact, k-means clustering doesn't match with categorical variable) for the venue category variable, substituted with a plethora of dummy variables.
The final step consists on extracting the mean frequency distribution of each venue category type for each city: this can be obtained grouping by city the table extracted from Foursquare.

City	ATM	Accessories Store	Adult Boutique	American Restaurant	Amphitheater	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Weight Loss Center	Wine Bar	Wine Shop	Wings Joint	Women's Store
Akron	0.0	0.0	0.0	0.026667	0.0	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0	0.0	0.013333	0.00	0.00	0.0
Allentown	0.0	0.0	0.0	0.020000	0.0	0.0	0.0	0.0	0.0	...	0.00	0.020000	0.0	0.0	0.0	0.000000	0.00	0.00	0.0
Antioch	0.0	0.0	0.0	0.043478	0.0	0.0	0.0	0.0	0.0	...	0.00	0.043478	0.0	0.0	0.0	0.000000	0.00	0.00	0.0
Aurora	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0	0.0	0.000000	0.00	0.00	0.0
Brooklyn	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.01	0.000000	0.0	0.0	0.0	0.000000	0.02	0.01	0.0

Table6: Cities venue types distribution table

Thanks to the preprocessing we'll be able to cluster the cities, via k-means clustering, based on their venue frequency distributions and, finally, obtain the **cluster matrix** that we want to "sell" or to exploit in order to build our content-based recommender system to the tourism companies, as stated in the **Business Problem paragraph** of the report. Of course is important to notice that, in

order to have more sophisticated system, the cities should be clustered in many ways (e.g. life-cost or distance) in order to have a cluster matrix in which the cities are in more than just one cluster. In this case we are focusing on exploiting the Foursquare API and get information from the venue distribution of each city.