Capstone Project - Battle of Neighbourhoods

1. Introduction

Almost each company has a time period, when it want to extend their area and open a new filial in another city. As well as some people want to move or visit another city somewhere.

Habemus Immobilien GmbH & Co KG is in Wien, wants to extend their facilities renting real estate all over the world according to clients preferences.

The company wants to get an advantage of local companies by creating an automatic system to help their clients to find a good area, according their preferences.

Later, when the system works properly, the company will order an app, that could help people to find a perfect location all over the world.

2. Data

In our prototype we will use some data of Toronto to look at the algorithm of the project.

We use the following resources:

- ✓ Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) to get Postal Code, Borough and Neighbourhood in Toronto
- ✓ Geospatial data for Toronto (http://cocl.us/Geospatial_data) to get the geographical coordinates of each postal code
- ✓ Foursquare API to obtain more information about venues
- ✓ Random user data, with a random number (from 1 to 10) of preferences to check, how our system works.

3. Methodology

In this project we will create a data based on a recommendation system that allows a user to choose the neighbourhood, which fit the best to his interests.

In the prototype there are used Toronto's data and random generated for user data to test the system.

- **Preparing and Obtaining the Data.**

First, we scrap the information about boroughs and neighborhoods in Toronto from the following Wikipedia page: wiki= 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'. Pandas' command "read_HTML" allows us to read HTML tables into a list of DataFrame objects and remove cells with a borough that is "Not assigned". Then we substitute unnamed neighborhoods for the name of relevant borough. After that, we group the table by the postal code the result is shown in the **Table 1**

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

The next step is to add the geographical coordinates (Latitude and Longitude) of each postal code to our data. We use some prepared date from http://cocl.us/Geospatial_data, but it is possible to use geocoder as well. The result of merging of 2 tables is shown in the **Table 2**

Out[27]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	43.757410	-79.273304
11	M1R	Scarborough	Wexford, Maryvale	43.750072	-79.295849

In the prototype we use the Foursquare API to get the venues for each neighbourhood, after tat we limit the amount of venues per neighbourhood to 100 and the range from the centre of the neighbourhood to 500 m. With this API we get all the venues for each neighbourhood and group them for each neighbourhood. We get a new table with the neighbourhood as the index and percentage of each category available in that neighbourhood applying OneHotencode in the categories and the mean for the amount venues for each category

➢ Random user

To generate a random user, we get a list of all categories available in the city. After that we select a random number from 1 to 10 to represent the amount of categories selected by the user. Then, from the list of categories we will sample the same amount obtaining the list of categories that our user will have interest in. Now we create a table with the categories as the columns and one row, where the values are 1 if the user has that category in his list and 0 for vice versa. This will result in a user profile that will be used in the recommendation system.

➢ Recommendation system.

Now, to make a recommendation system. We compare our user profile to the table with the neighbourhoods and the mean of value for the amount of venues of each category in it. So we multiply both matrix and apply a sum for each row. As the result we get a new matrix with the neighbourhoods and the score for each one of them. The higher the score the better the neighbourhood matches the user's interests. If we merge this table with the Table 2. We will be able to print in a map where are the better neighbourhoods for our user.

4. Results

In this prototype our user chose the next categories (Figure 1):

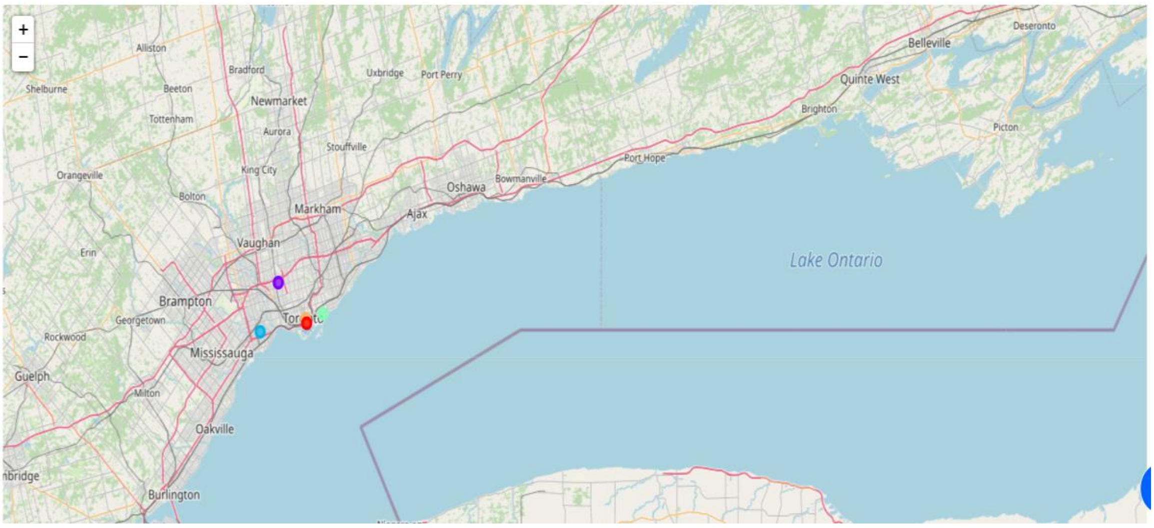
```
['Doner Restaurant', 'Taiwanese Restaurant']
```

With this user, we got the following score (Table 3) and a map (Figure 2) with the best 5 areas, which could fit to our user

Table 3

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Score
0	M4X	Downtown Toronto	St. James Town, Cabbagetown	43.667967	-79.367675	0.020833
1	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	0.013514
2	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.672710	-79.405678	0.000000
3	M6G	Downtown Toronto	Christie	43.669542	-79.422564	0.000000
4	M6E	York	Caledonia-Fairbanks	43.689026	-79.453512	0.000000

Map



5. Discussions

From this result, we can see that the 2 best neighbourhoods for our user are "North York Flemingdon Park, M4X :Downtown Toronto","M5T: Downtown Toronto",and "M5R :Central Toronto". From the Table 3 we can see that 2 areas, which have the same score, but the difference amount the 5 neighbourhoods is not big. A probable reason is that categories, which our user chose are more or less common, they don't include anything extraordinary as "Airport Food Court".

6.Conclusions

This is a sample content-based recommendation system that still need to be improved. The data and algorithm need more date and accuracy, especially for some small towns with a few venues. As well as a lot of work to collect the huge mount of data for other cities.