Coursera Capstone Project: A place for videogames Developers

@fserrey

Introduction and business problem

Toronto is an international centre for business and finance. Generally considered the financial capital of Canada. The city is an important centre for the media, publishing, telecommunication, information technology and film production industries; Although much of the region's manufacturing activities take place outside the city limits, Toronto continues to be a wholesale and distribution point for the industrial sector. The city's strategic position along the Quebec City–Windsor Corridor and its road and rail connections help support the nearby production of motor vehicles, iron, steel, food, machinery, chemicals and paper. The completion of the Saint Lawrence Seaway in 1959 gave ships access to the Great Lakes from the Atlantic Ocean.

There has recently been a substantial amount of interest in the emergence of video game development as an industry in Canada and its impact on the economy, the creative industries, the role studios play in specific city ecosystems and how video games affect physically and mentally. A recent study was done at McMaster University studying how playing video games improves the eyesight of those who suffer from vision problems.Toronto, Montreal, Quebec is a particularly popular subject of study due to the maturity of the gaming industry and its overall urban ecology.

Therefore, finding space for enough people to work with and/or start on the industry requres a selection of places where to share and build a network. As finding new places might be overwhelming, I decided to move to a office / coworking place locator in order to get to the right place.

Methodology

We will use K-mean clustering to segment and cluster Toronto neighborhoods to understand their similarity. With that understanding, we will be able to recommend a suitable place. Such locations would be near universities, media agencies, tech startups and coworkings spaces.

- List of Toronto boroughs and neighborhoods which can be found at https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M to explore, segment, and cluster.
- Toronto's sociodemographic data which can be found at https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods.

- Information on venues in Toronto extracted from Foursquare.com

	Postcode	Borough	Neighbourhood\n
0	M1B	Scarborough	Rouge , Malvern
1	M9B	Etobicoke	Cloverdale , Islington , Martin Grove , Prince...
2	M5S	Downtown Toronto	Harbord , University of Toronto
3	M3H	North York	Bathurst Manor , Downsview North , Wilson Heig...
4	M2N	North York	Willowdale South
5	M6J	West Toronto	Little Portugal , Trinity
6	M5T	Downtown Toronto	Chinatown , Grange Park , Kensington Market
7	M5C	Downtown Toronto	St. James Town
8	M4P	Central Toronto	Davisville North
9	M6R	West Toronto	Parkdale , Roncesvalles

In this project we will direct our efforts on detecting areas of Toronto that have high coworking spaces density. We will limit our analysis to area ~10km around city center.

In the first step we have collected the required **data: location and type (category) of every space within 10km from Toronto center**. We have **identified the type of spaces** (according to Foursquare categorization).

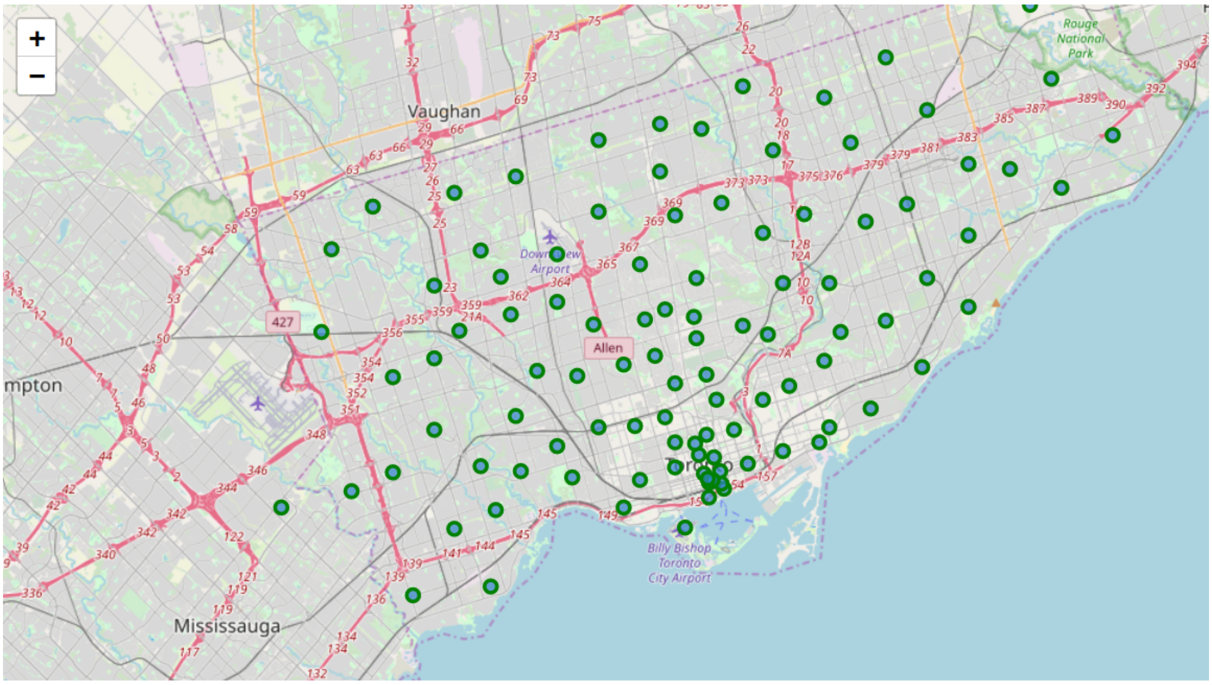
Second step in our analysis will be calculation and exploration of '**offices density**' across different areas of Toronto - we will use **heatmaps** to identify a few promising areas close to the center with a low number of spaces in general and focus our attention on those areas.

In the third and final step we will focus on the most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion with stakeholders. We will present maps of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

```python
map_geo = folium.Map(location=[latitude, longitude], zoom_start=11)

# add markers to map
for lat, lng, label in zip(df_geo['Latitude'], df_geo['Longitude'], df_geo['Neighbourhood\n']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='green',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_geo)

map_geo
```



Here we make a seleccion of the codes necessary to find closed-minded places where to find our place in Toronto. We have checked on Foursquare website which codes could fit in our search:

```python
list_of_interest = {
    'coworking_code':'4bf58dd8d48988d174941735',
    'tech_startup':'4bf58dd8d48988d125941735',
    'corporate_coffee_shop':'5665c7b9498e7d8a4f2c0f06',
    'recruiting_agency':'52f2ab2ebcbc57f1066b8b57',
    'college_technology':'4bf58dd8d48988d19f941735',
    'design_studio':'4bf58dd8d48988d1f4941735'
}
```

With more than 400 options located, we create a dataframe based on Foursquared API:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	venue_id	Venue Name	Venue Category	Venue Latitude	Venue Longitude
0	Rouge , Malvern	43.806686	-79.194353	50c5f88fe4b0eaecd9bec902	Imminent Concepts	[{'id': '4bf58dd8d48988d174941735', 'name': 'C...	43.804597	-79.199744
1	Harbord , University of Toronto	43.662696	-79.400049	5086c6ede4b0c33d74e5c691	Carolyn's Office	[{'id': '4bf58dd8d48988d174941735', 'name': 'C...	43.662975	-79.399147
2	Harbord , University of Toronto	43.662696	-79.400049	560a2a14498e624fdeecac0a	Grape Capital Office	[{'id': '4bf58dd8d48988d174941735', 'name': 'C...	43.662628	-79.403021
3	Harbord , University of Toronto	43.662696	-79.400049	4adf49b8f964a5201a7921e3	Health Strategy Innovation Cell	[{'id': '4bf58dd8d48988d174941735', 'name': 'C...	43.664691	-79.397242
4	Willowdale South	43.770120	-79.408493	51b0eeb7011c0a4b4d080de3	somolopro.com	[{'id': '4bf58dd8d48988d174941735', 'name': 'C...	43.769718	-79.411798

Analysis

Let's perform some basic exploratory data analysis and derive some additional info from our raw data. This can be achieved by clustering the neighborhoods on the basis of the office data we have acquired. Clustering is a predominant algorithm of unsupervised Machine Learning. It is used to segregate data entries in clusters depending on the similarity of their attributes, calculated by using the simple formula of euclidean distance.

We can then analyze these clusters separately and use those clusters that show high trends in our dataset

Normalization of the data for clustering

```
toronto_onehot = pd.get_dummies(tor_df[['Venue Category']], prefix="", prefix_sep="")

toronto_onehot['Neighbourhood'] = tor_df['Neighbourhood']

fixed_columns = [toronto_onehot.columns[-1]] + list(toronto_onehot.columns[:-1])
toronto_onehot = toronto_onehot[fixed_columns]

toronto_onehot.head()
```

	Neighbourhood	Bank	Coworking Space	Design Studio	Office	Recruiting Agency	Tech Startup
0	Rouge , Malvern	0	1	0	0	0	0
1	Harbord , University of Toronto	0	1	0	0	0	0
2	Harbord , University of Toronto	0	1	0	0	0	0
3	Harbord , University of Toronto	0	1	0	0	0	0
4	Willowdale South	0	1	0	0	0	0

With the following function, we get the most common venues in our DataFrame. This way, we can create columns according to the number of top venues:

```
num_class_venues = 6
indicators = ['st', 'nd', 'rd']

# Columns as number of class venues
columns = ['Neighbourhood']
for ind in np.arange(num_class_venues):
    columns.append(f'{ind+1} Most-common Type Venue')

# Create a new dataframe
venues_sorted = pd.DataFrame(columns=columns)
venues_sorted['Neighbourhood'] = toronto_grouped['Neighbourhood']

for ind in np.arange(toronto_grouped.shape[0]):
    venues_sorted.iloc[ind, 1:] = common_venues(toronto_grouped.iloc[ind, :], num_class_venues)

venues_sorted.head()
```

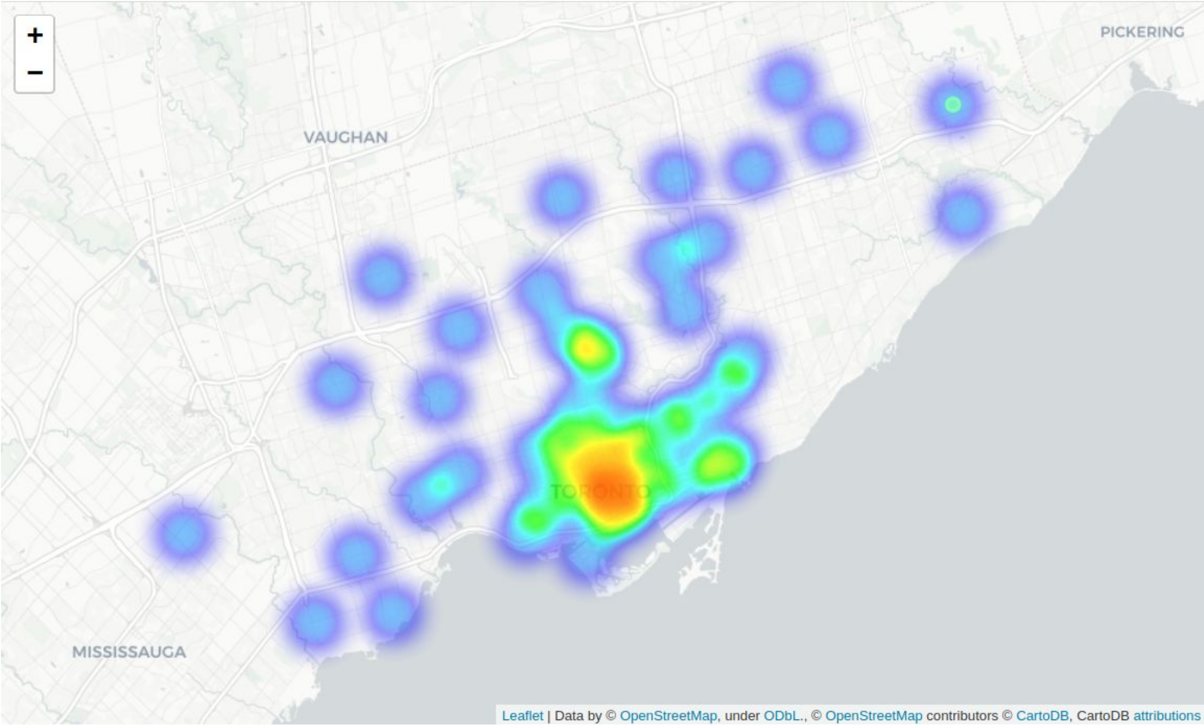
	Neighbourhood	1 Most-common Type Venue	2 Most-common Type Venue	3 Most-common Type Venue	4 Most-common Type Venue	5 Most-common Type Venue	6 Most-common Type Venue
0	Adelaide , King , Richmond	Tech Startup	Coworking Space	Recruiting Agency	Office	Design Studio	Bank
1	Agincourt	Tech Startup	Recruiting Agency	Office	Design Studio	Coworking Space	Bank
2	Agincourt North , L'Amoreaux East , Milliken ,...	Tech Startup	Coworking Space	Recruiting Agency	Office	Design Studio	Bank
3	Alderwood , Long Branch	Coworking Space	Tech Startup	Recruiting Agency	Office	Design Studio	Bank
4	Bedford Park , Lawrence Manor East	Tech Startup	Recruiting Agency	Office	Design Studio	Coworking Space	Bank

Modelization: K-Means

	Postcode	Borough	Neighbourhood	Postal Code	Latitude	Longitude	K-Labels	1 Most-common Type Venue	2 Most-common Type Venue	3 Most-common Type Venue	4 Most-common Type Venue	5 Most-common Type Venue	6 Most-common Type Venue
0	M1B	Scarborough	Rouge , Malvern	M1B	43.806686	-79.194353	3	Coworking Space	Tech Startup	Recruiting Agency	Office	Design Studio	Bank
2	M5S	Downtown Toronto	Harbord , University of Toronto	M5S	43.662696	-79.400049	0	Coworking Space	Tech Startup	Recruiting Agency	Office	Design Studio	Bank
4	M2N	North York	Willowdale South	M2N	43.770120	-79.408493	1	Coworking Space	Recruiting Agency	Tech Startup	Office	Design Studio	Bank
5	M6J	West Toronto	Little Portugal , Trinity	M6J	43.647927	-79.419750	1	Coworking Space	Tech Startup	Recruiting Agency	Office	Design Studio	Bank
6	M5T	Downtown Toronto	Chinatown , Grange Park , Kensington Market	M5T	43.653206	-79.400049	0	Tech Startup	Coworking Space	Design Studio	Recruiting Agency	Office	Bank

The K-Nearest Neighbors algorithm is a classification algorithm that takes a bunch of labeled points and uses them to learn how to label other points. This algorithm classifies cases based on their similarity to other cases. In K-Nearest Neighbors, data points that are near each other are said to be neighbors. K-Nearest Neighbors is based on this paradigm. Similar cases with the same class labels are near each other. Thus, the distance between two cases is a measure of their dissimilarity.

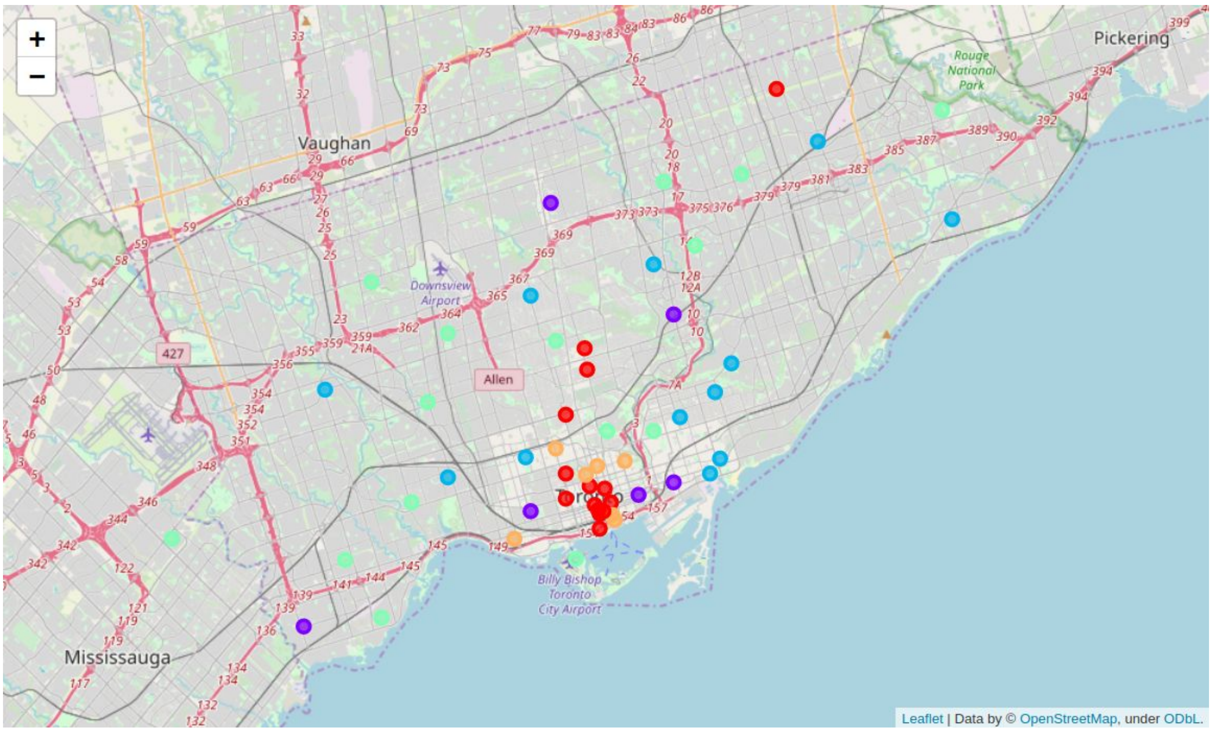
K in KNN, is the number of nearest neighbors to examine. The general solution is to reserve a part of your data for testing the accuracy of the model. Then choose k =1, use the training part for modeling, and calculate the accuracy of prediction using all samples in your test set. We can calculate the accuracy of KNN for different Ks.



Not bad - our clusters represent groupings of most of the candidate locations and cluster centers are placed nicely in the middle of the Financial District with location candidates.

Addresses of those cluster centers will be a good starting point for exploring the neighborhoods to find the best possible location based on neighborhood specifics.

Let's see those zones on a city map without heatmap, using shaded areas to indicate our clusters:



This concludes our analysis. We have created several addresses representing centers of zones containing locations with high density of techie offices, all zones being fairly close to city center (close to the Financial Center).

Most of the zones are located in that area, which we have identified as interesting due to being popular with companies related to those fields that videogames are involved, fairly close to the city center and well connected by public transport.

Results and discussion

Our analysis shows that although there is a great number of offices in Toronto (~400 in our initial area of interest and related to our industry), there are pockets of other interesting locations close to the city center and around the city. Highest concentration of these places was detected close to the Financial District, corresponding to boroughs Garden District, Harbourfront East, Toronto Islands and Union Station .

After directing our attention to this more narrow area of interest (covering approx. 5x5km south from Toronto) we first created a dense grid of location candidates; those locations were then filtered.

Those location candidates were then clustered to create zones of interest which contain the greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is zones containing the largest number of potential new officers locations based on number of and distance to existing venues - both design studios, technological companies, etc. This, of course, does not imply that those zones are actually optimal locations for a new videogame studio! Purpose of this analysis was to only provide info on areas close to Toronto center but not crowded with existing places. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

Conclusion

Purpose of this project was to identify Toronto areas close to the center with a high number of working areas suitable for videogame studios in order to aid stakeholders in narrowing down the search for optimal location for a new (and cool) place to create and develop new content for the videogame industry. By calculating offices and creative places density distribution from Foursquare data we have first identified general boroughs that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby companies. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal office location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to green areas, public transport, etc), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.