Segmentation and Clustering of Casablanca tram stations

Kenza Boutaleb Houssaïni Joutei

June 1, 2020

1. Introduction

1.1. Background

Casablanca is a city in Morocco which, in the last century, has seen a remarkable evolution in terms of area, population and activities. However, despite becoming an economic capital and one of the big cities in North Africa, public transport has not changed much. Indeed, apart from buses and taxis, its first modern tram line *(T1)* was not put into service until 2012. This line had been so sought-after by citizens that Casa Tramway – *the company in charge of the tram network in Casablanca* – opened a second line (*T2)* in 2019 in order to serve other areas. Despite the progress of the network, several neighborhoods are not served by the tram and have only a few Bus stations and taxis which can be very costly for residents in terms of time and money. Indeed, the city has about fifty neighborhoods grouped into 16 boroughs; while the two lines serve only 16 neighborhoods. The company will continue with two new lines *(T3)* and *(T4)* which are under construction, however it is advantageous to analyze the places in which stations are located or should be located to optimize routes and lines.

1.2. Business problem

Casablanca had a real problem of public transport. Neither bus nor tram station are enough for these 4 million residents. In order to optimize tram routes and lines, we will use this analysis to categorize stations by their surroundings and possibly by district.

1.3. Interest

This project could interest the community service in order to target the neighborhoods that require more public transport in general whether it's a tram or a bus. Also, public transport companies could use this analysis to optimize their stops and stations namely the case of Casa Tramway.

2. Data acquisition and cleaning

2.1. Data sources

For this project we will need:

- The list of Casa Tramway stations with their coordinates,
- The names of Boroughs and Neighborhoods in which belongs the stations,
- The venues surroundings each station.

Unfortunately we didn't find any page or document on the web which provides the coordinates of the stations, this is why we created the *geodata_stations* file which contains the name of the station, the line to which it belongs, the coordinates and the neighborhoods.

We may need the boroughs that we can find on the following pages: *https://fr.wikipedia.org/wiki/Ligne_1_du_tramway_de_Casablanca* for line 1, *https://fr.wikipedia.org/wiki/Ligne_2_du_tramway_de_Casablanca* for line 2.

Finally, we will use the **Foursquare API** to get the venues next to the stations.

2.2. Data cleaning and features selection

Since the file *geodata_stations* was created by us, no cleaning process was done. The dataset is already cleaned and well-structured for our analysis; we just loaded it and change the column names from French to English.

Next, we scraped the previous web pages to get the two tables corresponding to the two tram lines; and we removed columns that we deemed unreliable or unnecessary. We renamed the columns with the same format as previously and we combined the previous dataset with the two new datasets.

We retrieved the venues data with Foursquare API. There are several Macro-categories on Foursquare, the main ones are:

- **Art & Entertainment:** 4d4b7104d754a06370d81259
- **College & University:** 4d4b7105d754a06372d81259
- **Event:** 4d4b7105d754a06373d81259
- **Food:** 4d4b7105d754a06374d81259
- **Outdoors & Recreation:** 4d4b7105d754a06377d81259
- **Professional & other places:** 4d4b7105d754a06375d81259
- **Residence:** 4e67e38e036454776db1fb3a
- **Shop & services:** 4d4b7105d754a06378d81259
- **Travel & transport:** 4d4b7105d754a06379d81259
- **Train station:** 4bf58dd8d48988d129951735
- **Hotel:** 4bf58dd8d48988d1fa931735
- **Bus station:** 4bf58dd8d48988d1fe931735

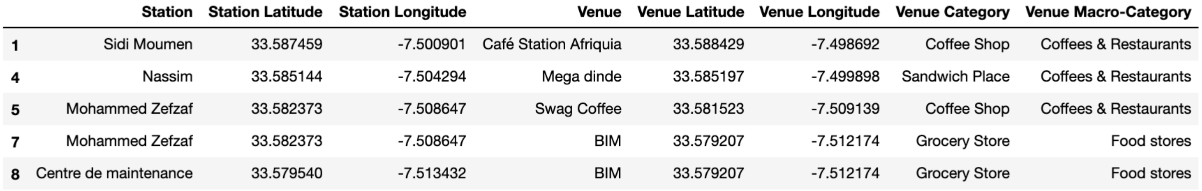
Unfortunately, the free version of this API didn't allow us to directly retrieve these categories from out dataset. Indeed, the loop exceeded the limit number of regular calls per day. So we imported the micro-categories specific to each place and then we grouped them into 9 macro-categories:

- Coffees & Restaurants
- Entertainments
- Arts & Cultures
- Sport Venues
- Shops
- Food stores
- Hotels & Bars
- Trains & Buses

Our dataframe contains now 428 rows which means 428 venues surrounding Tram Stations and it is shaped into 8 columns:

- **Station:** *name of the Tram station,*
- **Station Latitude:** *latitude coordinate of the station,*
- **Station Longitude:** *longitude coordinate of the station,*
- **Venue:** *name of the venue,*
- **Venue Latitude:** *latitude coordinate of the station,*
- **Venue Longitude:** *longitude coordinate of the station,*
- **Venue Category:** *Micro-category,*
- **Venue Macro-category:** *Macro-category.*

Bellow, a view of the final dataset:



	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Macro-Category
1	Sidi Moumen	33.587459	-7.500901	Café Station Afriquia	33.588429	-7.498692	Coffee Shop	Coffees & Restaurants
4	Nassim	33.585144	-7.504294	Mega dinde	33.585197	-7.499898	Sandwich Place	Coffees & Restaurants
5	Mohammed Zefzaf	33.582373	-7.508647	Swag Coffee	33.581523	-7.509139	Coffee Shop	Coffees & Restaurants
7	Mohammed Zefzaf	33.582373	-7.508647	BIM	33.579207	-7.512174	Grocery Store	Food stores
8	Centre de maintenance	33.579540	-7.513432	BIM	33.579207	-7.512174	Grocery Store	Food stores