Title: Battle of the Neighborhoods

Name: Carl-Michael Edeling

1. Introduction

An Italian Restaurant is looking to open another branch due to their continued success. It is located in Sandton, Johannesburg, South Africa, and is looking to open the second branch in a neighborhood similar to Sandton.

This project would be for any kind of business (in any area) that is looking to either open a second location or needs to relocate their business. The features used for analysis (in this case) are:

- Four Square Venue Data
- Weather Data (including Temperature, Wind-speed, cloud cover and rain)

However, this is only due to the limited scope of this project. For restaurants, other features could include:

- Traffic Data
- Demographic Data
- Economic Data

2. Data

Since the problem involves comparing suburbs (neighborhoods), geospatial data will be used with the Four Square API to get venue data for each suburb in Johannesburg. This data includes:

- Suburb Name
- Suburb Latitude
- Suburb Longitude
- Venue Name
- Venue Latitude
- Venue Longitude
- Venue Category

Since the nature of the business is weather dependent, weather data will be included. The weather measures used:

- Average Temperature
- Maximum Temperature
- Minimum Temperature
- Maximum Wind Speed
- Average Cloud Cover
- Maximum Rainfall

Once the data is acquired, the average weather measure will be determined and combined with the data from Four Square for analysis. This is a very simplistic measure to use, but for the purposes of this project it is fine. However, should more accurate results be necessary, it is recommended to use much more specific analysis.

The weather data will be acquired using the weatherbit.io API. They have different tier API keys, and for this project, since it is unfunded, the free API key will be used. This is limited by:

- 500 calls per day
- History Range: 4 years – Current Date

3. Methodology

3.1 Data Acquisition

Data was acquired from 2 sources:

1) Four Square
2) Weatherbit.io

APIs were used to access the data.

3.2 Data Manipulation

The Four Square data was one-hot coded which was used to determine frequency of venues. This venue frequency was used to create a table with suburb names as the row index, and the columns were the frequency of each venue.

The weather data acquired from weatherbit.io was narrowed to include:

- Average Temperature
- Maximum Temperature
- Minimum Temperature
- Maximum Wind Speed
- Average Cloud Cover
- Maximum Rainfall

The average for each of these measures were combined with the Four Square venue data for analysis.

Then these values were normalized using the equation:

$$average_{normalized} = \frac{x_i - \mu}{\sigma} \qquad \text{(Equation 1)}$$

Where:

x_i = current value

μ = the mean value

σ = standard deviation

3.3 Data Analysis

Since the problem is a question of comparing suburbs, it was decided to use the unsupervised k-Means Clustering algorithm.

The optimal number of clusters to use was determined using inertia. This was plotted against number of clusters and the optimal value found at the “elbow-point”, representing the point where the diminishing value of inertia becomes significantly less.

2 runs of clustering was done:

1) Only on the venue data
2) On both the venue data and weather data

This was done to compare how the clustering is affected by the simplistic weather measure averages and normalization. Depending on how clustering changes could indicate potential problems with the weather data analysis.

Analyzing the venue data to test quality of the data was also performed. Something as simple as count showed that the data returned by Four Square was incomplete, and for more accurate results, it is recommended to improve the quality of the data to analyze.

4. Results

A total of 49 suburbs were analyzed.

The venue data returned by Four Square was shown to be incomplete. It indicated that for many of the suburbs, that there was only 1 Italian Restaurant. Where in actuality that are many more.

The k-Means algorithm was run twice. Once using only the venue data, the second using a combination of venue and weather data.

The results of the first run showed 3 suburbs similar to Sandton. This makes sense since Sandton is considered an upper-class neighborhood.

The suburbs similar to Sandton are:

1) Illovo
2) Sandhurst
3) Hyde Park

These are the recommended locations for opening of the second restaurant.

The results of the second run showed 20 suburbs similar to Sandton. This shows that the weather data dominated the classification, and that the results are not to be trusted. The significant increase in similar suburbs shows that the weather data analysis was inappropriate for the chosen method of clustering, or rather, that the chosen features needs to be updated.

5. Discussion

The fact that the second run (which included both venue and weather data) had a significant increase in number of similar suburbs shows that the inclusion of the weather data skewed the results. This could be because:

a. The venue data values are frequencies of venues which are absolute values ranging from 0 to 1. However, the weather data was normalized using Equation 1, resulting in values ranging from [-1; 1]. This means that the values are not comparable.
b. The weather data acquired was very limited. Ideally, daily data would have been used, but the API key has limited accessibility to data.
c. More statistically appropriate measures could yield better results that a simple average of the weather measures.

Running k-Means clustering on the venue data returned the best results. But that being said, alone, the venue data is insufficient to make the decision about whether the selected suburbs will be an adequate location to open the second restaurant. As it only includes the venue data, it is missing:

Economic Data – Average spending ability per household would indicate if people in Sandton have the ability to spend as much as people in similar suburbs.

Demographic Data – Social demographics would also be a useful indication about whether an Italian Restaurant would be popular in similar suburbs.

Traffic Data – Customers ability to get to the restaurant, and delivery drivers' ability to reach the customers would both be impacted by traffic data.

If these had been included, the results would yield far more reliable information. However, since they were not done, I cannot recommend any suburb as an appropriate substitute for Sandton.

6. Conclusion

Based on my knowledge of my hometown, Johannesburg, I can say that the k-Means algorithm correctly identified suburbs similar to Sandton based purely on venue data.

Inclusion of the weather data, skewed the results so much that it is recommended to redo the weather analysis completely.

I would not recommend any specific suburb as a facsimile for Sandton. But that is only due to the lack of data. If I had access to more complete data and I had more time, this analysis would be far more accurate and reliable. However, I do feel that this project showed me that I understand how to collect, clean, manipulate and analyze data using Python and the necessary libraries, which is the real essence of this capstone project.