1 Introduction

1.1 Background and interest of the project

For individuals who want to move elsewhere, it is sometimes important for them to get a place that is similar, or drastically different, from the one they are leaving. The criteria may vary for each individual, but essentially, the goal remains the same: to have insights regarding similar locations to a pre-defined one, in order to compare them and make a more informed choice.

1.2 Description of the problem

Let us imagine that someone named Jim is living in Manhattan, and is planning to moving out elsewhere. He considers several locations, such as other New York's boroughs, but also other global cities. To this end, he selects three European cities (Berlin, Rome, Paris) that he likes, as well as two Asian (Tokyo and Shanghai) and one South-American (São Paulo) city. One thing for sure, Jim loves Manhattan and its *population density*, and *all the most popular venues* that compose his borough. Thus, would it be possible to have a recommender system that suggests boroughs located in those cities that share, or at least are similar, to a precise borough? In Jim's case, boroughs similar to his Manhattan?

2 Data collection

To address the problem, we will collect data pertaining to the considered cities. As such, data will be scraped from the cities' Wikipedia pages. Pages do not necessarily need to be in English: for instance, the English São Paulo page is not displaying the data we need (population of the subprefectures), but the Portuguese page does. Data will then be cleaned, formatted similarly, and will then be regrouped under a single dataset.

To complement this data, we will use the GeoPy API to fetch the coordinates (Latitude and Longitude) of each borough of the dataset. Those coordinates are needed, as we will use them jointly with the FourSquare API to fetch the most popular venues for each borough.