Find your new home

Olesya Tsapenko

June 07, 2020

1. Introduction

1.1 Background

Every year thousands and thousands international students graduate from European Universities. Fair share of graduates want to stay and work in the EU. A lot of them might face the challenge of not knowing the local language. They start to search for cities/countries where they can find a job in English preferably in the EU. There are at least three well-known EU capitals for it - Amsterdam, Paris and Berlin.

1.2 Problem

To find a job purely in English - it is a big problem but to find a new home may be even harder.
On one hand - each city is a unique experience. But on the other hand there are many neighborhoods which have pretty similar venues around. At this work I will compare 3 EU capitals' neighborhoods and cluster them to identify some similar places for living based on surrounding venues. In the second part I will add my current neighborhood and show a comparison between it and possible places for living. The script is flexible and assumes that any user can provide their address and look at the personal recommendations.

1.3 Interest

Freshly graduated international students from EU universities would be interested to find a new home for themself which is similar to their living place. Comparison of neighborhoods/cities/countries can also help to find a place which is absolutely different with the current life situation if this is the goal of the user.
Others who are interested in this comparison such as educated english-speaking people may also want to choose a place to move in the EU.

2. Data acquisition and cleaning

2.1 Data sources

The first challenge which I met is to find a list of all neighborhoods of given cities. For this, I parsed Wikipedia web-pages using the beautiful-soup python framework. The next step is to

receive coordinates (latitude and longitude) of each neighborhood which I did with the help of a geopy library (also this library helps me to find coordinates of my current living place). When we already have latitude and longitude of each neighborhood it is easy to find venues in the given radius (500 meters) using API requests to Foursquare service.

2.2 Data cleaning

Unfortunately, I could not receive coordinates of absolutely all neighborhoods using geopy so these places were excluded from future analysis. Also I dropped neighborhoods which have less than 5 venues in the radius of 500 meters. On top of that, I renamed French, German and Dutch restaurants categories to one category “Local cuisine restaurant” to make clustering more unbiased.

As a next step of data cleaning, I calculated all unique venue categories and encoded each of them using one-hot-encoding. Later I group all venues by neighborhood and by taking the mean of the frequency of occurrence of each category. It became a feature for my future clustering and recommendation. I want to mention that I didn’t use any geographical-related features such as latitude, longitude, city or country names, neighborhood names to make the clustering more objective for geographically-separated places.