

A District Recommender

1. Introduction / Business Problem

Anyone having to move to a new city for work or studies, faces the problem of having to decide where to stay. If that person has not travelled to the new city in the past and does not have the time or budget to do an on-site exploration, he or she must base their decision on the information they can find on the Web.

The problem is that online travelers' guides of new destinations are destined to tourists, while forums for expats or students have scant information and carry subjective opinions based on small samples.

The new "District Recommender" service comes to fill that gap and support decisions with objective data.

1.1 Interest

Users interested in the new "District Recommender" service may be workers who received and accepted job offers that require their relocating to a new city, or students who have accepted enrolment offers and are about to start their first year in a University. The problem is particularly felt in big cities that present a large number of options.

The majority of users will presumably be young, so they will appreciate a Recommender service that takes their personal preferences into account.

The service will also be of interest to business partners, because it can draw a lot of advertising revenue. Users will declare their intention to move to a new city and their preferences in lifestyle, supplied in terms of areas in cities they have visited and enjoyed in the past. It will be possible to target advertisements even without collecting and storing other sensitive personal data.

1.2 Success

We will know the service is successful if it can produce neighborhood recommendations for new destinations that are compatible with a user's prior preferences. To understand if this is true, the new system will be tested by evaluating its recommendations on cities in which the author has first-hand living experience.

2. Data

The choice of data sources used to solve the problem depend heavily on the scope of the service. Our decision is to create a District Recommender service that supports mobility of

young workers or students in many countries. The system will be first tested on three popular destinations:

- London, England;
- Glasgow, Scotland;
- Copenhagen, Denmark.

Other types of data was considered, but to use it we would be obliged to restrict the scope. For example, happiness & wellbeing data or deprivation data were easily available only for London boroughs. Data on house prices per square meter could be found for England and Wales, but not for Scotland.

The following data will be used.

2.1 Districts

Our neighborhoods in each city will be defined by partially aggregated postcodes; popular, coarse-grained subdivisions that define entire boroughs. We will not use the detailed postcode used to define specific streets or even single buildings in the UK or Denmark. So, in the UK we will use the postal district 'EC2' (Bishopsgate), not the full postcode 'EC2M 4NR', of which there are 1.7 million in the UK.

Care must be taken to eliminate any districts that are not residential but are used for mail distribution purposes.

The easiest way to obtain the postcodes is to scrape them from Wikipedia, for example

- https://en.wikipedia.org/wiki/List_of_postal_codes_in_Denmark
- https://en.wikipedia.org/wiki/G_postcode_area
- https://en.wikipedia.org/wiki/London_postal_district#List_of_London_postal_districts

2.2 Geolocation data

For each of the previous districts, we will obtain geolocation data, either by an online geolocation service or downloading it from a Web site. To avoid too many API calls, for the UK we will use the detailed postcode data available from the Office for National Statistics – Open Geography Portal <http://geoportal.statistics.gov.uk/search?q=postcode>

2.3 Venue data

For each district, we will use the Foursquare API <https://developer.foursquare.com> to obtain a list of available venues in each location. The list of venues will be used to define the profile of each district. This is similar to the method employed in the last lab of this course relative to Toronto neighborhoods, but there will be some further processing as explained in the next paragraph.

2.4 Venue categories

A possible hindrance in using a venue category returned by the Foursquare API, is that it may be too detailed to allow for meaningful comparisons. For our purposes, it does not make much sense to make a very fine distinction between the sub-types of restaurants or the types

of outdoor recreational activities. The Foursquare API supports a hierarchy of categories <https://developer.foursquare.com/docs/build-with-foursquare/categories/> that we will experiment with, in order to understand if there are ways to improve venue classification and neighborhood comparison.

2.5 User preferences

The District Recommender service will ask users to submit their preferences in terms of a district they liked in a city they have visited in the past.

This location input by the user will be used to calculate the user's preferences in terms of district type and of the lifestyle they would like to enjoy. The user preference is essential to build an item-based recommender.