Constructing a Family Friendly Neighborhood Index

Public Project Summary

By Fasih Atif

Submitted to

Central European University

Department of Economics and Business

Master of Science in

Business Analytics

June 2021

Contents

1. Objective

Family-friendliness is a complex term with many dimensions and requirements. This project aims to summarize them by combining global and local data sources from various sources and develop quantitative and qualitative measures on the family-friendliness of urban areas.

The project's first phase consisted of collecting POI and other city element datasets relevant to family-friendliness to quantify it. In the second phase of the project, I collected POI user reviews and relevant tweets and conducted sentiment analysis to support the family-friendliness of the urban areas. As family-friendliness has many dimensions, the final index is a composite measure, which combines with a certain computational method both different POI-level information and other statistics. Based on this, each urban area was attached by a single parameter.

2. Project Initiation

The project started with a team meeting with the stakeholders to design a high-level project plan. An 8-week progress plan was designed with several milestones. We decided to work on–5 international cities from which 1 city was decided as our primary focus while the rest were the secondary focus. After that, we discussed which factors increase/decrease the family-friendliness of a neighborhood. I spent a week reading up on several research papers to brush up on my geospatial knowledge and get hands-on with the Geospatial libraries in Python. The data collection process began with surveying various sites and databases for data. The initial data included basic details such as the name and address for each entity under a specific POI. For example, a POI could be a marketplace and the entities would be all the marketplaces in a city. The goal was to find the most up-to-date and comprehensive data so that our outcome could be based on the most accurate information. I found 3-4 large (POI) resources (hereafter mentioned as data source) that were sufficient to get data for all target cities.

3. Data Collection

The first data source A required web scraping to extract the data. I learned Selenium (automation tool) and beautiful soup (web scraping library) and used the acquired skills to scrape data from several pages and manipulate the collected data into a tidy table. The data from Data source B was

acquired using a Python Package that let me download data from an online platform. This data was much easier to obtain but was much harder to clean due to incorrect and outdated details for some of the POIs. Data from Data source C was obtained through Application Programming Interface (API) and CSV files. Data Source D is an online platform that stores location data and reviews. This source has several good APIs that can be used according to our needs.

4. Data Cleaning

Once all the data was collected and cleaned, I moved on to cleaning and validating the data. There were several duplicate POIs across all sources of data but spelled with different numbers of words and letters. The duplicates posed a problem since there was no way to remove duplicates from the available data itself. On further research, I used Data Source D to geocode the POIs collected from all sources and extracted the relevant details from the source. The details included a unique ID for each distinct entity, so I was able to remove the duplicates using that ID. This process was repeated for all cities.

5. Sentiment Analysis

Once the entities were collected and finalized, I searched for online reviews for those entities for the calculation of sentiment scores. These scores would be incorporated into the final index. I came across a lot of review sites, but each site only possessed a small number of reviews for the top entities of a POI in that city. I again turned to Database D which stored a good number of reviews for a lot of the entities. Using its API service, I was able to retrieve the reviews. I conducted sentiment analysis on the reviews using 4 different sentiment packages namely Textblob, Vader, Afinn, and SentimentR. The results received varied between the different packages. Out of the 4, Vader and SentimentR provided the nearest true measure of the sentiment. I then compared the results between Vader and SentimentR. For easy comparison, I split the sentiment scores into Positive, Negative, and Neutral and then matched the respective packages. Vader and SentimentR provided the same sentiment category on approximately 80% of the data. I proceeded to use the numeric sentiment scores of Sentiment R since it proved to be a better predictor of the sentiment.

Twitter can be a very good source for gauging the sentiment of people towards different places. Using the Twitter API Standard Track, I obtained geotagged tweets by specifying coordinates of

the city center and defining a 25 miles radius around it. The number of tweets returned were low and irrelevant. I came across another python package called Twint. It scrapes Twitter for tweets and does not require an API. I was able to obtain a huge number of tweets but many of them were not relevant to family friendliness.

6. Limitations

1. I was forced to drop some of the entities due to incorrect results. The index tried to cover as many entities under each POI as possible, but it was not possible to obtain every single entity.
2. The index had been created at a sub-neighborhood level. Understandably, some POIs did not exist in some neighborhoods. So, that POI's weightage in the index was 0 which affected the overall ranking of the sub-neighborhood.
3. Lack of reviews for many of the neighborhoods led to null values which affected the final index. Hence, the incorporation of reviews was dropped for now.
4. It was a very complicated task to filter for tweets related to family-friendliness from a limited collection of tweets. Moreover, the tweets had location information at a city level instead of a sub-neighborhood level. Considering the limitations, the idea of tweet sentiment scores was dropped.

7. Constructing the Family Friendly Index

As a primary deliverable of the project, I constructed a Family Friendly index that captures whether a certain small urban area is family-friendly or not. As family-friendliness has many dimensions, the final index is a composite measure, which combines with a certain (NDA-protected) computational method different POI-level information, and other statistics. Based on this, each urban area was attached by a single parameter.

8. Recommendations

1. To get a better picture of the ground truth, we can try to obtain data for entities that were originally left out due to discrepancies in details.

2. To get Twitter data, we can apply for the Academic Researcher track that allows us to obtain tweets dating back to 2006. This will allow us to obtain a bigger sample of geotagged tweets.

3. To filter for family-friendliness-related tweets, we can use the Stanza NLP package developed by Stanford University Researchers. We can define keywords for the most important dependencies that we choose and let the grammatical relations help identify useful and meaningful tweets.

4. Since the number of reviews that could be extracted from Data Source D was limited, we can program selenium to scrape the reviews from the source.

5. We can use 2-factor analysis or conjoint analysis to determine the division of weights for the dimensions of the index.

6. We can use population parameters to filter out all irrelevant areas so that the deserving sub-neighborhoods get the family-friendly rating that it deserves.

9. Lessons Learned

- It is important to establish a clear research focus and small goals.
- We should research the tools we will use and read documentation beforehand. This can save hours of debugging code.
- Data collection and cleaning takes the most time in a project.
- Keeping in mind the growing population and demand for infrastructure, it is more important than ever to learn how to use spatial analysis to achieve different outcomes be it for personal use, business, or even humanitarian.
- Communication is key in making a project successful.
- It is important to create a Knowledge Database for every project which you can consult later if required.