Improvements on Airbnb Search and Recommendation System

Yongfei Lu

December 2022

Abstract

In recent years, a new form of economy, named 'sharing economy', unrolls swiftly via platforms like Airbnb. Its search and recommendation engines provide travelers with an excellent experience. However, in order to obtain a list of accommodation listings, customers have to specify the address of their destinations, and also other requirements on amenities, including number of bedrooms, Wifi, air-conditioner, etc. In other words, users face strict query constraints. We believe that for travelers new to their destinations, owning the freedom to type in free text queries, such as “places with great lake views” or “places to have a great Swedish culture experience in Chicago”, brings more benefits and convenience. Moreover, we expect more socioeconomic information to be provided for users to gain general knowledge of the conditions of the neighborhoods they’re going to stay. Such information, including neighborhood average income, crime rate, and population statistics, can serve as a precious foundation to help users make decisions. To achieve such goals, we take advantage of the Airbnb listing and review datasets, create index on thousands of listing descriptions, build recommendation systems based on sentiment analysis on numerous reviews, and conduct geospatial data analysis on the Airbnb datasets. We hope that better user experience can be delivered with the above features.

Key words: Airbnb, search engine, recommendation system, geospatial analysis.

1. Introduction – Problems to Solve

Entering the Airbnb website, users will be forced to type the destination address, check in and out dates, and number of guests. While this ensures a well-formed search query, facilitates the search engine to better understand users' information needs and deliver good results, it does impose some inconvenience on the user side. Intuitively, it makes more sense for someone, who is new to the city, to ask the search engine "what's the optimal places with great lake views in the city of Chicago". To deliver highly relevant results for such free text queries, we build use ElasticSearch to create index on the listing description data provided by the hosts. To comply with the convention, we also build a recommendation system using the Collaborative Filtering method to predict user's preference and recommend other housings they probably also like.

On the other hand, while the application provides detailed information about the listing itself, it ignores users' need to gain more knowledge of the neighborhood they're going to stay, including the neighborhood crime rate, average income, and other demographic features.

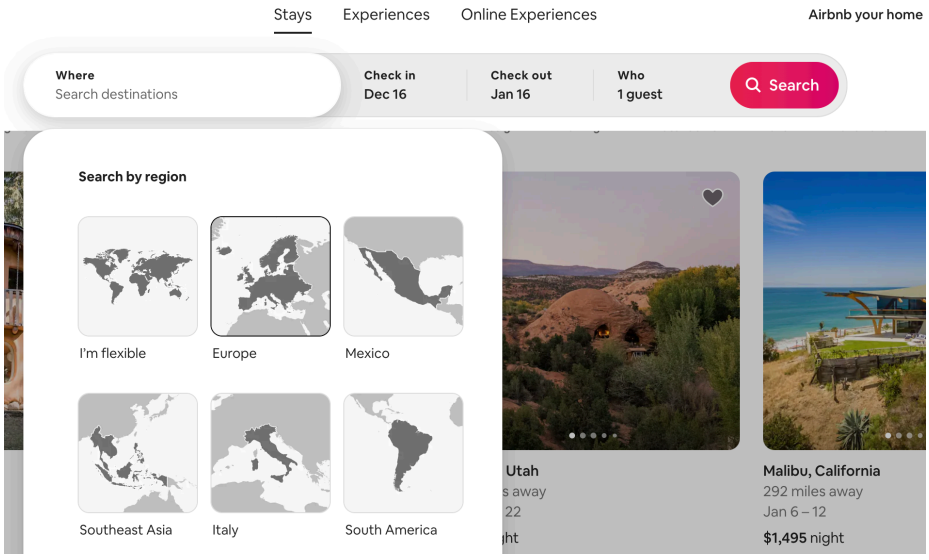

Figure 1: Airbnb User Interface

Therefore, via the datasets from the Chicago Data Portal, we conduct geospatial data analysis on city neighborhoods to give users such information. We believe that with such socioeconomic information, customers can make better housing decisions.

Due to limited time, we constrain the project scope to the city of Chicago, which contains 77 neighborhoods and can well present the gist of the project. The project is fully scalable and can be adapted to all major cities in the United States with more data included, which is also available in both the insideairbnb.com and the cities' data portal.

2. Data

2.1 Data Source

1) Detailed listing and review data: insideairbnb.com[1].

2) Neighborhood Boundaries Community Areas (current)[2]: boundary file for the Community, Areas in Chicago, Chicago Data Portal.

3) CSV table: Chicago Community Area (CCA) CDS (population) raw data.[3]

inputs for July 2020 Community Data Snapshots, presented at the CCA geographic level.

4) Crime data: Crimes - 2001 to Present, Chicago Data Portal[4].

2.2 Data Wrangling

For this project, we mainly use Python packages, pandas and geopandas, to wrangle all the datasets. We clean and select the main variables from listings.csv and form the listings_doc.txt, which is then used to build index for the search engine. We clean and merge the listings.csv, reviews.csv, CommAreas.csv, Crimes.csv, and chicago_population.txt, and export the final full dataset to neighborhoods.csv, which we use to conduct the geospatial data analysis work and

[1] http://insideairbnb.com/get-the-data
[2] https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6
[3] https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data/resource/8c4e096e-c90c-4bef-9cf1-9028d094296e
[4] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

present the socioeconomic information to users in our application. All the data wrangling code is in the Data_Wrangling.ipynb file.

2.3 Data Analysis

After all data is wrangled, we then plot the graphs to deliver the socioeconomic information in a straightforward way.

Figure 2 shows the distribution of sentiment polarity on Airbnb listings in Chicago. Polarity of 1 represents the attitude behind the review is very positive towards the listing, 0 means neutral and -1 stands for negative reviews. From the graph, we can see that most reviews are positive and some positive reviews (with polarity of 0.5 or less) can contain critics of listing flaws. Polarity. From Figure 3, we can see that the most positively reviewed listings tend to be located in the northern and southwestern part of the city. Figure 4 shows that the neighborhoods with highest level of crime rate tend to be the ones around the Chicago downtown area and the south ones. Another interesting pattern can be found when analyzing Figure 5 to 7: neighborhoods with higher white population ratio and lower black population ratio tend to have higher average income and lower crime rate. However, we cannot find any pattern indicating that crime rate is correlated with polarity levels in the scope of neighborhoods.

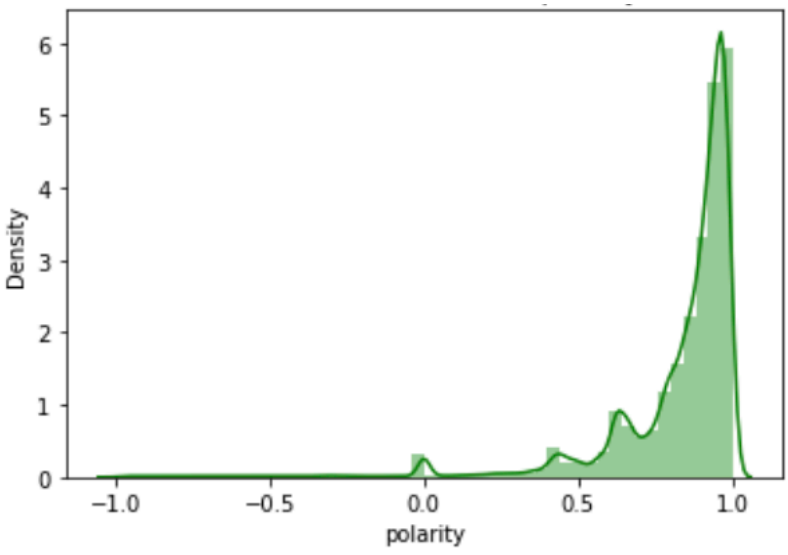

Figure 2: Distribution of Polarity

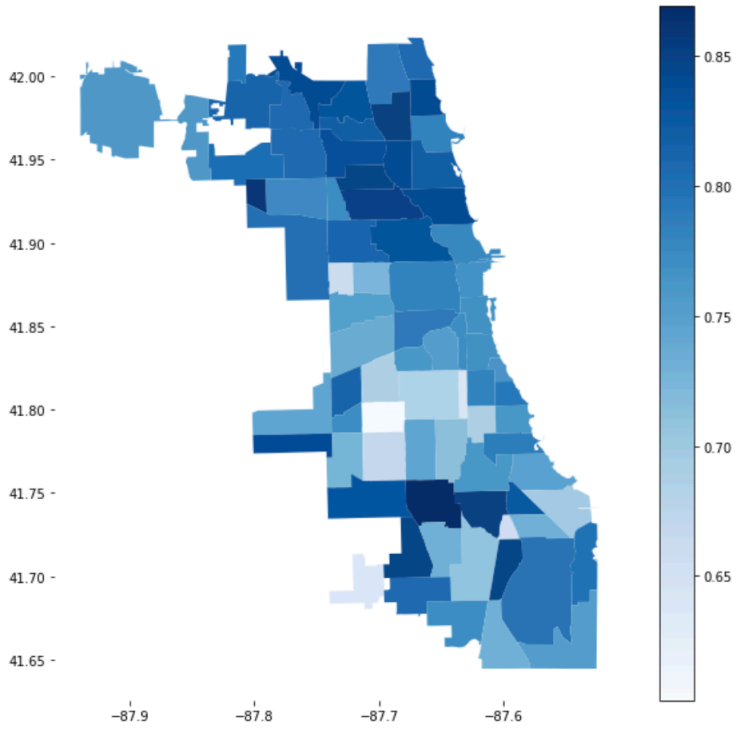

Figure 3: Average Polarity per Neighborhood

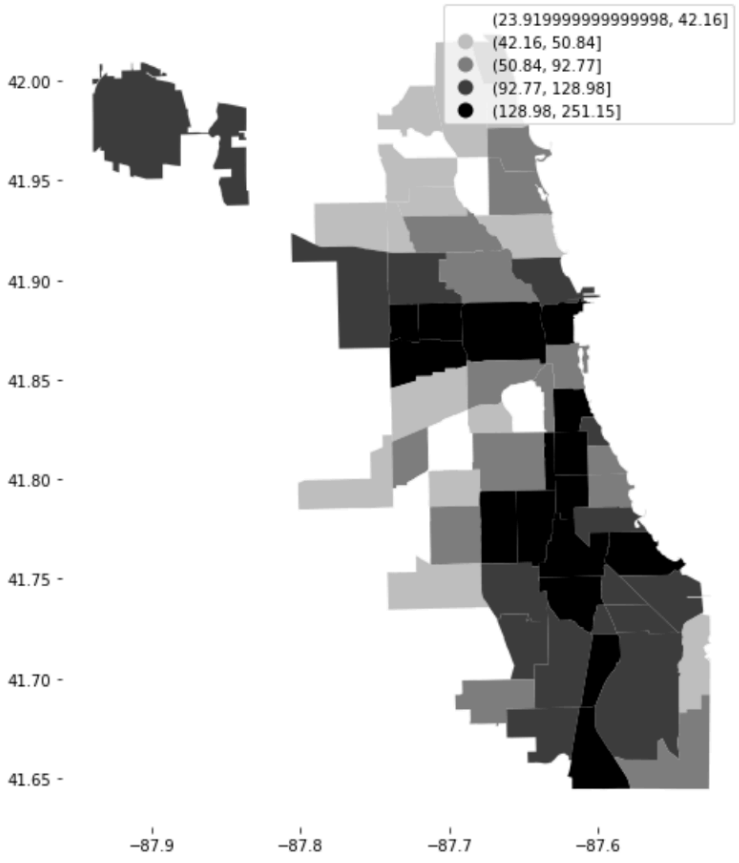

Figure 4: Crime Rate per Neighborhood

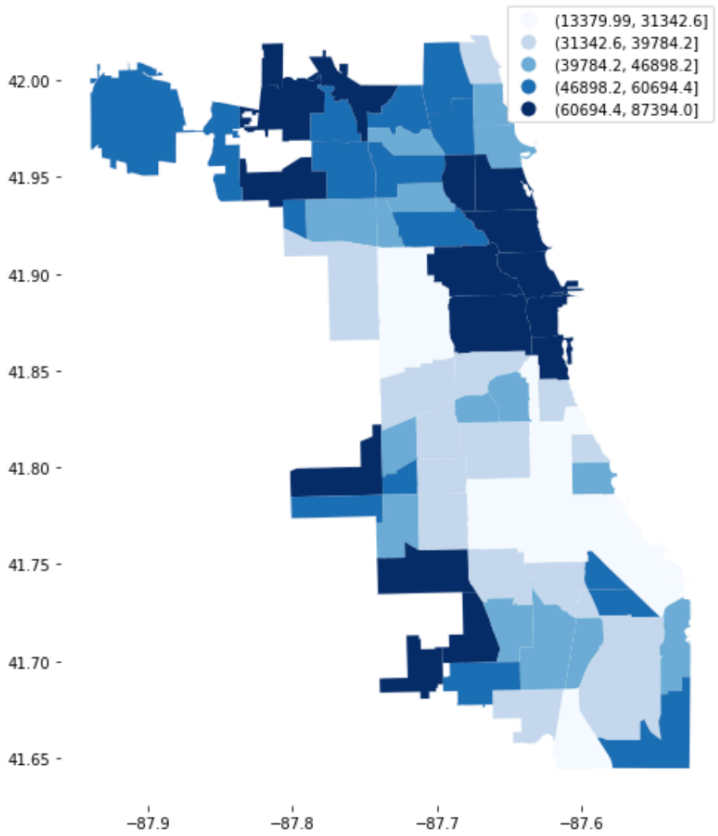

Figure 5: Average Income per Neighborhood

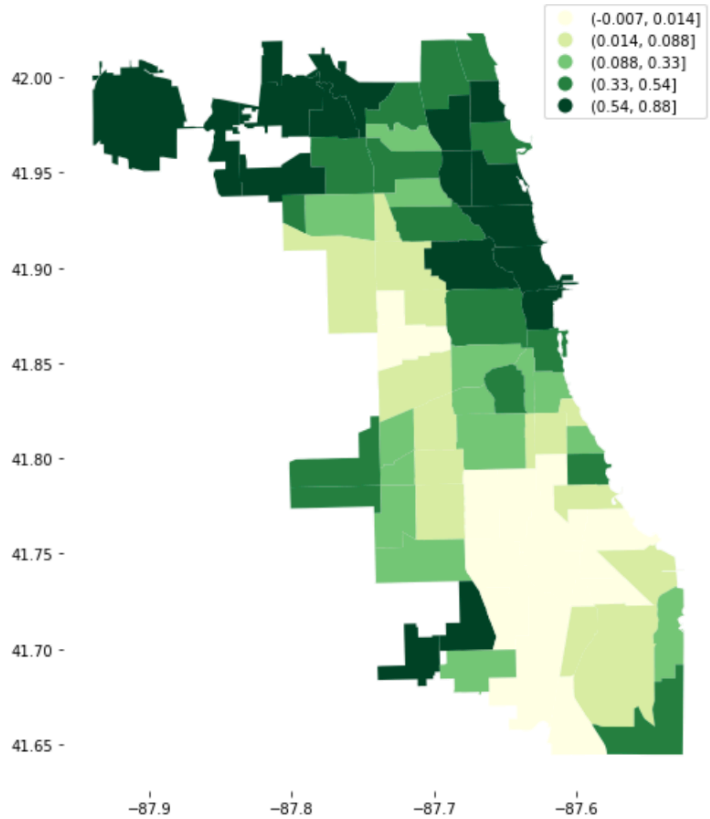

Figure 6: White Population Ratio per Neighborhood

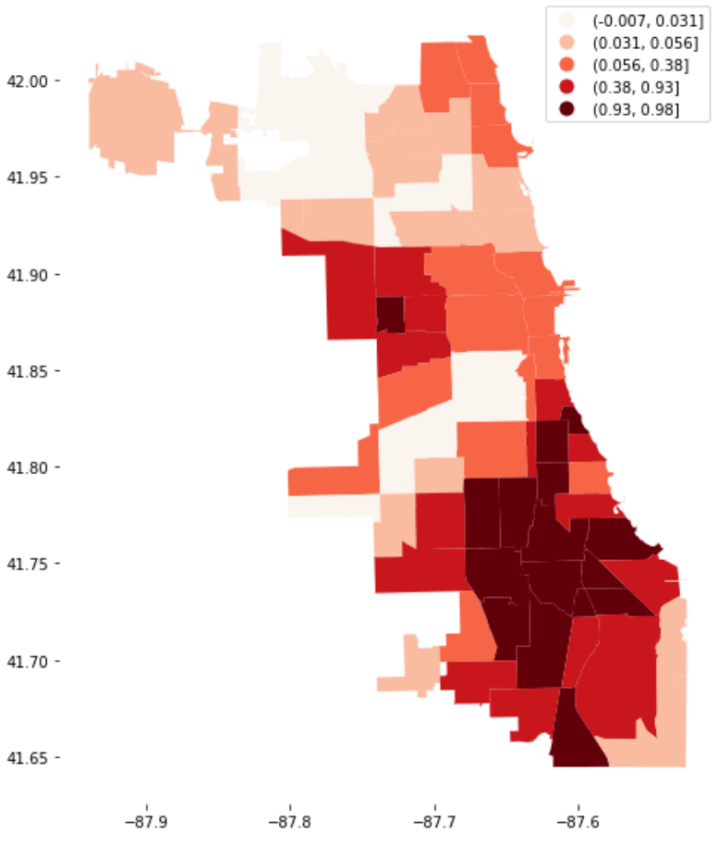

Figure 7: Black Population Ratio per Neighborhood

3. Methods and Implementation

3.1 Search Engine

To build the search engine, we create index on the description of the listing dataset, which is cleaned beforehand. The description contains the overall information about the housing. By using the index created by ElasticSearch, we successfully deliver the free text query functionality.

3.2 Recommendation System

To build the recommendation system using the collaborative filtering method, we first conduct the sentiment analysis on the reviews on Airbnb listings using the NLTK ' s Valence Aware Dictionary and sEntiment Reasoner (Vader) tool, which has great power in dealing with the work mainly by adding valence to words that are capitalized, have degree

modifiers or emotional punctuation, and can even detect the sentiment shift in one single sentence. After this, we use the dataset to train the Singular Value Decomposition (SVD) model, which is a grate matrix factorization tool to predict users' preference. By using this model, we predict and rank users' preference over the listings and return the ranked top listings.

3.3 Geospatial Data Analysis

By using the community boundary data and the neighborhood-level crime, income, population data combined with the listing dataset, we use python geopandas package to plot the Chicago neighborhood maps with relevant statistics (Figure 3 - 7). With such information, we aim to increase customers' knowledge of the neighborhoods where they plan to live, and eventually facilitate their decision-making process.

3.4 User Interface Construction

Finally, we build the simplified user interface of our application using React and Python Flask, through which we can conveniently search for relevant listings and get the recommendations automatically.