NYC Small Business Analytics

<Shi Fan, fanshi118, sf2632> 12/08/2015

Overview

In this extra credit project for PUI, I explored the potential of creating a recommender system of small businesses for different neighborhoods in New York City. The workflow is as follows:

- Select the neighborhoods to be analyzed
- Get the ratings for a list of businesses in each neighborhood
- Calculate the similarity score between one neighborhood and another for all the neighborhoods included in the analysis
- Find the type of business worth developing for each neighborhood by comparing its business scores with its most similar neighborhood

Data

Geo data of NYC neighborhoods

- This data is obtained directly from installing the ArcGIS software, the package of which comes from the book *GIS Tutorial 1: Basic Workbook*. Inside the NYC geodatabase, the Neighborhoods shapefiles contains the geo data for all the NYC neighborhoods. I selected all the neighborhoods in Manhattan plus some neighborhoods in Brooklyn. Also, I merged a few neighborhoods, just to be consistent with the Yelp location search, which uses the Google API. (Google's neighborhood classification is slightly different from DCP's.)

Business rating data from Yelp

- All the data are scraped from Yelp reviews. I looked into ten business types which I consider as the major components of urban lifestyle: restaurants, bars, coffee & tea, health & medical, arts & entertainment, fitness & instruction, grocery, education, haircut and boutique. I took review counts into consideration for each business, so that the score is determined by both the rating and the review count on Yelp. I also used different metrics for different types of businesses, because some businesses (restaurants) tend to have more reviews than others. Figure 1 is a box plot of the preprocessed data.

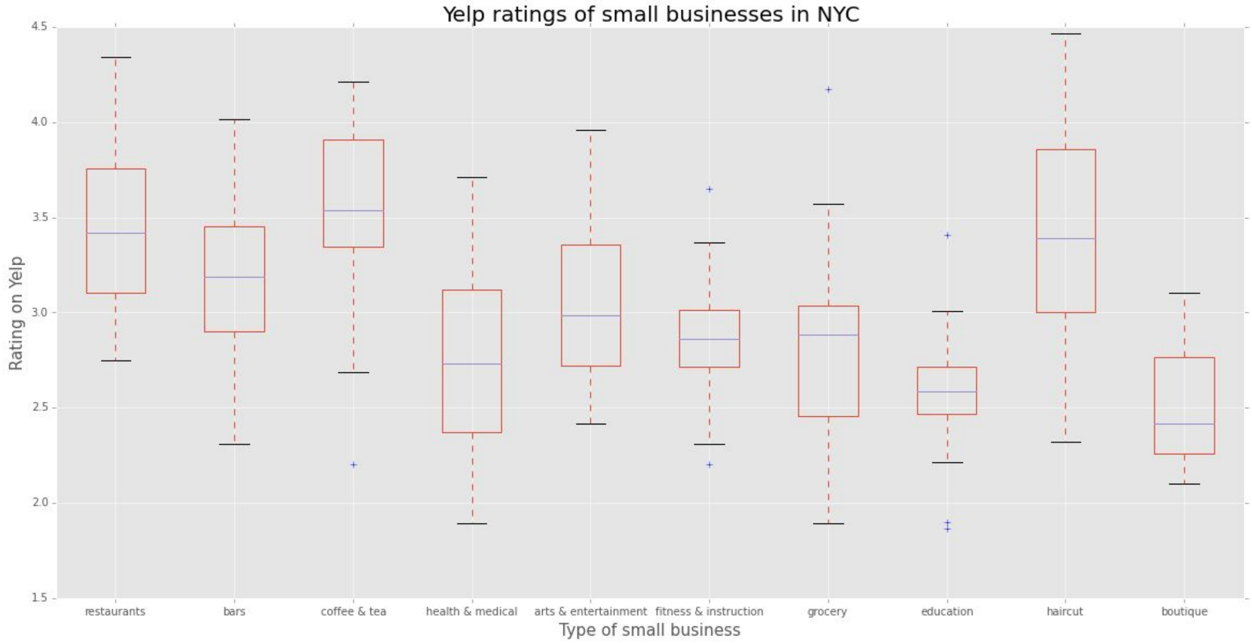

Fig. 1. Box plot of Yelp ratings collected for each type of small business.

Analysis

In the simple analysis, I directly calculated the cosine distance of each pair of neighborhoods using the preprocessed Yelp ratings, which act as our vectors. Then I subtracted the cosine distance from 1 in order to get the "similarity score," which is later normalized for each neighborhood. After getting all the similarity scores for each pair of neighborhoods, I compared the specific ratings of each business among each most similar pair. For each neighborhood, I chose the one type of business that is outperformed the most by its most similar neighborhood as the recommended option. The justification is that if such type of business has an obviously higher rating in the most similar context, it will most likely have the room to grow in this particular neighborhood, based on what the numerical value suggests.

Then I tried to base my recommending approach on a more sophisticated method. I used matrix factorization to discover latent features underlying the data. Afterwards, I calculated the similarity scores for the factorized data, following the same steps as above to get the recommended business for each neighborhood.

Deliverable

Figure 2 shows the results of the recommended type of business for each neighborhood. The simple method and the matrix factorization method present different results in each case. Education and grocery are popular recommendations in both cases.

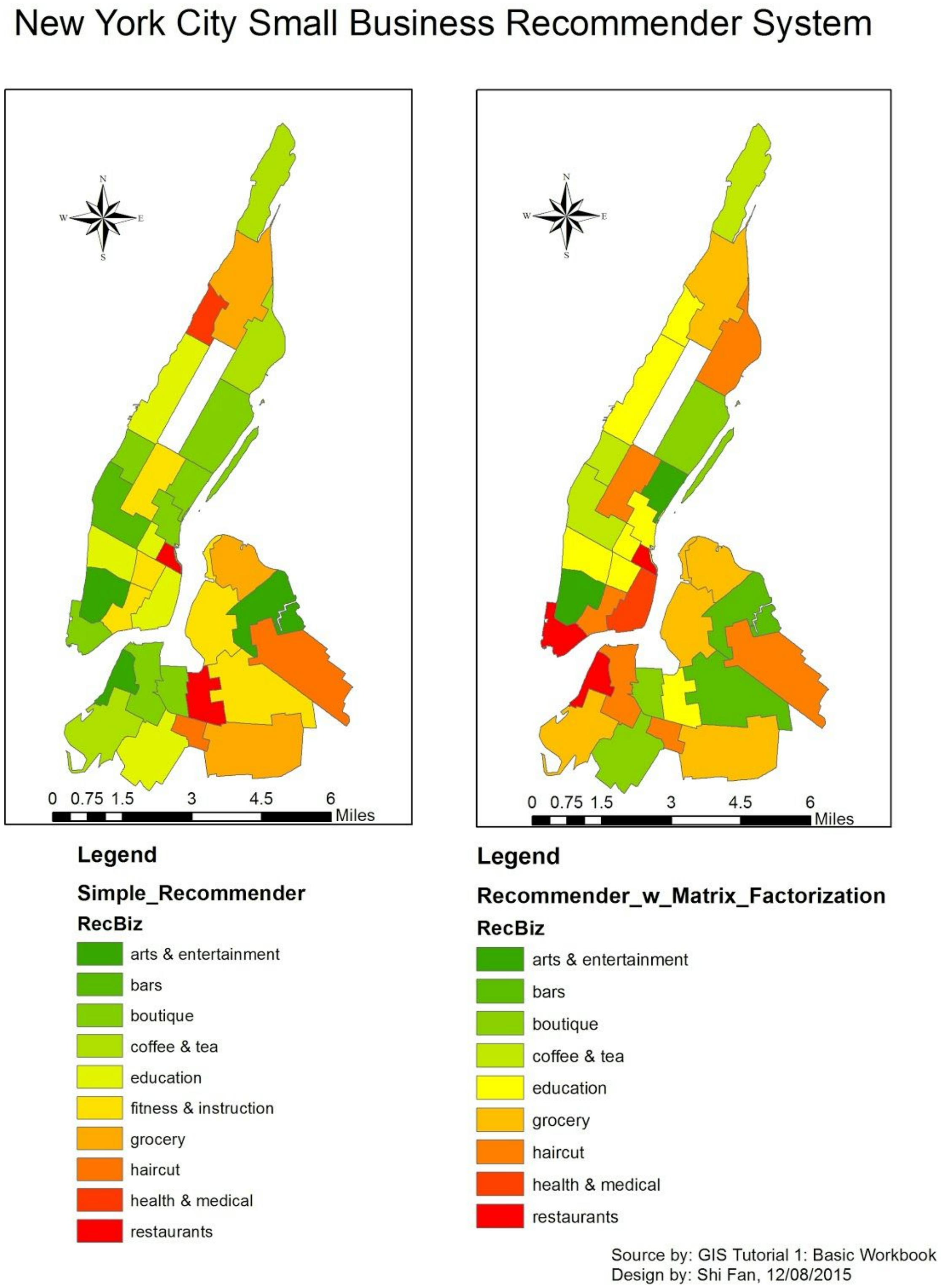

Fig. 2. Small business recommendation for neighborhoods in Manhattan and Brooklyn.

Discussion

The recommender system takes on a quite straightforward yet simple approach. In addition to the Yelp reviews, we may also need to take the total number of each type of businesses into account as well as the land use conditions. Also, it can be helpful to compare the business ratings of each neighborhood to those of its nearby neighborhoods, as well as the most similar neighborhood. This is a preliminary attempt from a computational standpoint; however, in reality the development of such metrics require domain knowledge from policy makers. The project mainly shows how social media data can be applied and potentially helpful to the field of city planning.

References

1. Yeung, Albert Au. “Matrix Factorization: A Simple Tutorial and Implementation in Python.” quuxlabs, 16 September, 2010. Web. Accessed on 8 December, 2015 URL: <http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/>