Capstone 2 Proposal- Yelp Recommender System

What problem do you want to solve?

Being able to recommend a restaurant to a user could be a very useful tool for Yelp. It could help them tailor advertisements to users and could improve their search algorithm. The more accurate restaurant recommendations become the more likely a user will be to return for another search. The goal is to utilize the dataset to develop a recommendation system for yelp users.

The Dataset

The dataset being used is available on Kaggle:
https://www.kaggle.com/yelp-dataset/yelp-dataset/version/6

Contains over 5,200,000 user reviews, 174,000 business and spans over 11 metropolitan areas. As a bonus this data can be enriched with data scraped from yelp, potentially images from the restaurants in the dataset.

Proposed Solution/Approach:

To solve this problem a combination of machine learning techniques will be used to build a hybrid recommendation system which will utilize both collaborative and content filtering.

For a given user the system can gather restaurant recommendations by analyzing restaurants visited by similar users (by computed user similarity). It can then qualify these recommendations based on a restaurant recommendation score. The restaurant recommendation score will be computed by taking business attributes, restaurant categories, and menu items and comparing these to a user's historical restaurant preferences. This will essentially score how likely it is a user will enjoy a restaurant given the restaurants she has enjoyed in the past and their attributes/profiles.

A sentiment analysis can be performed on the reviews to gain more information on user preferences as well as additional insights about specific businesses. This information can be fed into the user similarity and restaurant recommendation scores respectively.

Since restaurants are rated from 1-5 stars by users, this information can be used to predict the star rating a user will assign to an unrated restaurant. This would involve vectorizing the user review and business data into a form suitable for applying some regression algorithms. Note that this could also be treated as a classification problem or even a hybrid solution.

A network analysis is proposed for constructing the baseline recommendation system. This will treat users and restaurants as nodes in a bipartite graph with restaurant reviews as the edges linking the two sets of partitioned nodes. This will aid in analyzing user activity as well as in comparing user profiles. For efficiency purposes this will focus on a subset of the network, comparing users within a given city. This will also ensure that results are reasonable with regard to distance from the user.