Project Proposal

“We Like Food”

Alex Maiorella, Lily Li, Leoson Hoay, Nancy Gong

Datasets:

- **Yelp Academic Dataset (6.5GB)**
    - 174k businesses
    - 5.2 million reviews
        - Dated 2004 - 2015
    - 1.3 million users
    - 11 metropolitan areas
        - Including Ontario, Arizona, Edinburgh, etc.
- **IPUMS American Community Survey Data (~100MB)**
    - ~33179 rows, ~10 columns
    - Includes various socioeconomic and demographic variables (ex. race, age, poverty level, population, etc.) for all zip codes in the US
    - The exact number of columns will depend upon the number of variables we want to include.

We plan to link these two datasets together by zip code.

Potential Investigations:

We have three main questions we plan to investigate:

1. **What predicts business success?**
    a. Do businesses perform better when they are near other businesses, or when they are far away?
    b. Are there common themes or keywords that reviews of successful businesses share? What are the most important factors commonly mentioned in reviews that highly relate to the success of a certain type of business? (for instance, service and ambience are the main topics in reviews of highly rated restaurants)
    c. Do the socioeconomic conditions of a business’s location affect its success? Are there any interesting correlations between socioeconomic conditions and business characteristics?
2. **Where do food deserts or other types of service deserts exist?**
3. **Do friends have similar tastes?**

For our first question, we plan to measure business success using a business's average rating and the number of reviews it has. For question 1a, we would like to explore if businesses do better when they're located near other businesses or far away from other businesses. Perhaps centralized locations of shops (ex. shopping malls) bring more clients to each store, or perhaps they draw away clients to a business's competitors. We will examine which one of these two competing theories dominate by computing the distances between stores using their latitudes and longitudes. It may also be that the types of stores surrounding a business affect its success. Perhaps a dentist office does well when next to a shopping mall, but does poorly when surrounded by other dentist offices. We will examine this secondary question by modifying the previous analysis to account for the types of the businesses.

We will answer question 1b and identify the important factors by using topic modeling techniques, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization(NMF), to analyze reviews.

We will answer question 1c by using the linked Yelp and IPUMS data to examine the socioeconomic conditions of each business's zip code. We will then run regressions to test the effects of these socioeconomic variables on a business's success.

We will answer question 2 by grouping zip codes and calculating the number of grocery stores, restaurants, doctor offices, clothing stores, etc. in each area to understand which locations are oversaturated and which areas qualify as food/doctor/clothing deserts (i.e. areas that do not have access to grocery stores or other types of services).

We will answer our third question by using Yelp's friends data to identify pairs or groups of friends, and then examine the types of restaurants they visited, their reviews, and their star ratings data to understand if friends tend to share similar tastes. This analysis would focus on restaurants. We will implement NLP algorithms, such as sentiment analysis and document-to-vector models, to analyze their reviews.