CS123 Project Proposal

Prepare by Mitsue Iwata, Kyle Magida, and Scarlett Swerdlow.

Data set or type of data we want to work with

The foundation of our project will (hopefully) be the Yelp Dataset, which includes 1.6M reviews and 500K tips by 366K users for 61K businesses in ten cities. The Yelp Dataset on its own is roughly 1.5 GB. Thus, a first task will be to add to this data to meet the size requirement of this project. Our current thinking is we will do a cosine similarity analysis for either users or businesses, possibly in select cities. We estimate that if we analyze every pair of users, we will have as many as 1.8 billion rows of data.

Possible hypothesis / hypotheses

Do Yelp users who are similar based on cosine similarity:

- Share social networks?
- Go to the same types of places?
- Give similar ratings to businesses?
- Write similar reviews?

In identifying similar users, we will explore whether similarity is robust to different definitions (engagement with Yelp, quantity and quality of Yelp reviews, etc.)

People and their roles

- We will work together to implement cosine similarity
- We will break up exploratory data analysis tasks
    - One person will explore whether similar users write similar reviews
    - Two people will split the remaining four possible tasks
- We will break up analysis once we have a better idea of what we're doing

Timeline

- Meet with Prof. Wachs about cosine similarity (Week 4)
- Implement cosine similarity analysis to identify similar users (Week 4)
- Exploratory data analysis (Week 6)
- Decide on question to answer (Week 6)
- Do the analysis (Week 10)