CMSC 12300 Project Proposal
Lucy Chen, Daniel Cheng

Dataset
Our primary dataset will be Yelp's Academic dataset (https://www.yelp.com/academic_dataset), along with Yelp's much larger Phoenix-metropolitan database: (http://www.yelp.com/dataset_challenge). These databases contain information from over 15,000 businesses, 110,000 associated business attributes, 150,000-edged social graph, and 300,000 reviews. The dataset is in JSON format, with one JSON-object per line. Our final product may also webscrape data directly from Yelp's pages to test that our findings generalize outside of the two databases.

Data Exploration and Initial Prototype
Questions of Interest and Corresponding Techniques/Algorithms
Should Sichuan and Beijing cuisine have their own category, or should these restaurants all be lumped under the category of "Chinese restaurants"? We will create a program to analyze Yelp data and suggest new restaurant categories, thereby implying which existing labels are too redundant, and which new ones might be added.

Our program will ultimately rely on techniques of cluster analysis: We shall first determine ways of measuring similarity, then identify clusters of restaurants. First, determining two restaurants to be similar will involve testing the following sub-hypotheses. For example, restaurants could be said to be more similar for the following reasons:

1. They serve similar types of foods. Food similarity would be computed initially by simple frequency counts (and pulling the top k-similar items using algorithms discussed in class), and subsequently by higher level concept categories.
2. The same people (i.e. reviewers) tend to frequent them. Finding where groups of people prefer to dine is itself another cluster analysis task.
3. The restaurants have a similar "image" based on similarities in their self-descriptions. This would be computed as a natural language processing task (likely using either Stanford NLP Group's Parser or Python's NLTK package).

Cluster analysis will then be carried out in R, given its existing cluster analysis libraries (e.g. hierarchical clustering, and k-means/Partitioning Around Medoids). Python's NLTK cluster packages will also be helpful in analyzing textual data. In computing the clusters, we may also have to write a learning algorithm trained on extant categories to determine how these factors should be weighted.

Timeline
April 16 - April 23: (1) explore the smaller Yelp Academic Dataset (~250 businesses) and write simple R and Python code that calculates e.g. simple frequency of foods, distribution of reviewers, and repetition of words (corresponding roughly to #1-3 in the above section) (2) based on this data exploration, determine what cluster analysis and NLP techniques will work best for the Yelp Dataset
April 24 - May 1: apply these more sophisticated cluster analysis and NLP techniques to build on the code in (1)-->push a prototype showing restaurant categories in our smaller dataset, as well as discovered clusters of popular foods and of groups of reviewers

May 2 - May 16: scale analytical techniques to the entire Phoenix-metropolitan database by running k-means (or another clustering algorithm) on Hadoop
May 17 - May 24: debug program; possibly test findings on Yelp pages outside the two given databases
May 25 - June 3: transform data findings into easy-to-visualize network or graph

*Our proposal focuses specifically on restaurants because most people seem to use Yelp's food reviews, but our algorithms can be generalized to all Yelp businesses (i.e. we would carry out the same analysis within every Yelp business category).

Final Results

We aim to develop a program that can analyze the linguistic relationships between restaurants, discover new groupings of restaurants, and present these relationships in an easy-to-read graph. On one hand, these generated groupings should have a strong relationship to the categories currently in use, so that we know that these groupings correspond in some way to the intuitive notion of a "type" of restaurant; on the other hand, these groupings should also show us new restaurant categories that teach us something new about how people pick restaurants based on similarities in food options, other patrons, and restaurant self-image.