**Analysis on Yelp Reviews Data**

Huitong Kou, Zihang Wang, Peibin Rui

**Introduction**

There are tens of thousands of reviews for different types of businesses. Our analysis mainly focuses on restaurants which serve steak in several US cities such as Madison and Pittsburgh. Among these restaurants, our specific goals are:

* Investigate the relationship between ratings and words in reviews from two different aspects: foods, non-food items.
* Provide useful advice not only for improving the ratings of existed steak restaurants on Yelp, but also for opening a new steak restaurant based on our analysis.

**Data Cleaning**

Our dataset contains a subset of million reviews from restaurants in Madison (U.S.), Cleveland (U.S.), Pittsburgh (U.S.) and Urbana-Champaign (U.S.) released by Yelp. The restaurants with at least 3 reviews older than 14 days are included and only reviews that were recommended at the time of the data collection are included

Limited by the computing power, we chose to focus on restaurants whose category contained the word “steak”. Then we used some tools in nltk to split, clean and recreate the reviews.

The yelp.ipynb in code folder is used to clean the data and create the word frequency table and

the large word embedding table with original views.

After filtering and combination, the steak\_cleaned.csv has 33629 rows × 6085 columns.