Group Members: Dae Hyun Kim(dk58), Yong Hyun Kwon(yk3), Sang Ha Park(sp86), Junsu Choi (jchoi21)

Vision:

Our vision is to create a recommendation system based on the review text in the Yelp Dataset. We plan to run clustering algorithms / LDA on review text in order to cluster users together based on their reviews and to cluster restaurants based on these user clusters. In the process, we'll need to analyze the review text extensively, including extracting latent variables (important topics) via LDA or other methods such as TF-IDF. We will attempt to build a recommendation algorithm based on this process.

Data:

- **Dataset (size, parsing/cleaning, etc)**
  - **Size:** ~1.8G for Yelp dataset (+ for additional data - Uber, foursquare, tripadvisor) + CHD Experts dataset for Nevada
  - **Content:** 4 million reviews, 900,000 tips, 1 million visit attributes, aggregated check-ins, etc
  - **Parsing / Cleansing**: JSON format

Progress:

- **Blog Post I:**

In this blog post, we focused on data exploration. Here, we decided to focus on Nevada within the dataset, which is why we explored how many restaurants were in each region in Nevada (Las Vegas, Boulder City, etc). We also made several visualizations: number of ratings for each rating (1 - 5), heatmaps, distribution by zip code, and more.

- **Blog Post II:**

In this phase of the project, we decided to come back to the larger dataset (not just Nevada) - we are using an EC2 Instance on AWS to run our data analysis. Next, we decided to run a Linear SVM on review text to predict ratings based on the text with a resulting precision value of ~0.8 (which is significantly better than random guessing on three classes -- positive, neutral, negative). It was also interesting to print out the top 10 features with the greatest weight for each of the three classes. Lastly, we also ran the same kind of analysis for restaurant open / closed - can we predict whether a restaurant is permanently closed / currently open based on review text?

Plan:

In order to cluster restaurants and users based on review text, we need to figure out a way to choose features. While we will experiment with a variety of different methods, for now we plan to place our focus on extracting latent variables (topics) via LDA and compare the results with simple top TF-IDF words within the review text. After obtaining several sets of features, we will attempt to find user and restaurant clusters.

We think it is important to place the cluster results into perspective - thus, we must come up with a evaluation criteria and run many analysis via visualization to see if we have discovered meaningful topics, user clusters, and restaurant clusters.

Lastly, our team must consider the workings of the recommendation system. An interesting idea might be to have a recommendation based on free-form text. Other ideas include preference-based recommendations (location, price range, etc) and selection of specific restaurants (I like restuarant X - recommend me some similar ones).