Project Proposal

Chuning Yuan

**Project description:**

Based on the Airbnb data and the Facility data of San Francisco, both of the data set contains the geographic information of Airbnb and other Facility in SF, including latitude and longitude, and reviews detail, prices, jurisdiction of the facility. I think it will be interesting to explore the data set of Airbnb by its reviews using the text mining analysis, and use the geographic information to seek the connection between the distribution of Airbnb and the facility in a city and how the price of the Airbnb different base on the facility of the neighborhood. I can group the Airbnb by Latitude and Longitude. If I can cluster Airbnb by geographic information and review information, and check if there are a lot difference between these cluster criteria.

**The main questions to explore from the data:**

* How positive or negative are the reviews of Airbnb?
* How to cluster Airbnb according to geographic information?
* How to cluster Airbnb according to reviews information?
* How to cluster Airbnb according to facility information of the same city?
* Do the Airbnb and facility clusters in a similar geographic information?

**The presumed analysis procedure:**

* Clean data and extract restaurant observations from the original dataset.
* Modify, reconstruct and quantify restaurant feature information.
* Apply text analysis on restaurants and calculate similarity coefficient between each restaurant.
* Apply cluster analysis based on review, feature and geographic information.

**The potential problem:**

* There are many NA of the data set, and there are few different languages are used of the reviews data that I need to deal with.
* I need to combine the review Airbnb dataset with the Airbnb listing data set by the room id because the review data set does not contain the geographic information of each room, but the listing data set will have more detail information base on each Airbnb room.