08/11 Project Outline

# Zocdoc Potential ratings for reviews

* I am planning to predict user ratings given user reviews. I will be predicting user ratings across 3 dimensions, Overall Rating, Bedside Manner, Wait Time
* The data is scraped from zocdoc.com. I am fetching all Primary care doctors at Manhattan. There are about 20,000 records.  
  After extraction and cleanup, I will be generating a csv file with following format:  
  Doctor Id, Doctor Name, Doctor Specialization, Zip Code, Patient Name, Review Date, Overall Rating, Bedside Rating, Wait Time, Review Text
* My idea here is to use user reviews and ratings, to train a machine learning model. This model will be able to predict ratings of new reviews.
* Since the problem here is a classification problem that uses text, I would like to use text features and classification models like Naïve Bayes, Logistic regression, Random Forests, KNN etc. to see which model fits the best here.
* Since this is a classification problem, it will be measured by % accurate predictions.
* Challenge: Identifying relevant text features, ngrams etc. Also, most users have rated high and there are missing reviews. So it would be a challenge for the model to predict low ratings.
* Applications:
  + To quantify user reviews. We can potentially scrape blogs of people regarding some doctor and say what would they rate this doctor as
  + To predict missing ratings
  + To group similar users/ doctors
  + A reverse of this application is also useful. Where users give ratings but they don’t provide review. We can predict what are the potential buzz words in the review of this user, given his ratings.