The review user gives come with two parts, star and text. The star ranges from 1-5. Here I want to build a predictor based on the review text and predict if the user like the restaurant or not. In this case, I assign the 5 star rating to like, and other as dislike. You might wonder “Hey 4 star is a good rating”, Yes I agree, however in many people’s text review of the four star rating, instead of saying how much he/she like the restaurant, they are talking about the reason why they did not give a 5 star. In this case, classify the 4 star into dislike makes more sense.

I Used NPL techniques, such as stemming, lemmatization and TF-IDF, to extract features from review text data. Then built language understanding models to classify positive and negative reviews using Logistic Regression and Random Forests. The model achieves 84% accuracy.

From a restaurant point of view, if a customer makes suggestion if might be a hint to improve somewhere. And if many customers make the same suggestion, the hint to improve is stronger. Here I have built a similar review search engine, when you see one review and want to know if anyone else is saying the same thing, the similar review search will be a great tool.

Here I extracted the NLP representation of the text data and run cosine similarity between the reviews. This can provide a recommendation for the business to improve.

I used unsupervised learning (Kmean) to cluster user into different groups. I noticed when K is small, the two cluster typical are like or dislike, as K increase we start seeing different type of food forming its own cluster.