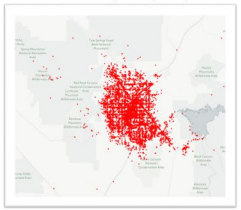
Restaurant Recommendation using Yelp Reviews

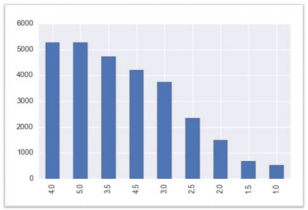
ver. Las Vegas

Dae Hyun Kim, Junsu Choi,
Sang Ha Park, Yong Hyun Kwon
https://junsu10291.github.io/CS1951ABlog/

Introduction

Yelp provides millions of reviews and ratings of restaurants written by millions of people across the world. While valuable, each user has a different standards of rating. One might give 5 only for life-changing experiences, while others might give out 5 for even the most mediocre restaurants. In short, ratings is a biased metric. In order to overcome such bias in ratings, we decided to build a restaurant recommendation system purely based on texts to match the user's need.





We used dataset provided Yelp, which included 4.1M reviews by 1M users for 144K businesses.
However, due to the large dataset, we decided to focus on
Las Vegas, Nevada, which included over 720000 review texts and 10000 businesses

To improve the quality of the recommendation, we further narrowed down the scope of the dataset. We chose users with over 50 reviews and restaurants with over 100 reviews. The remaining dataset had 49346 reviews, 1620 restaurants, and 558 users.

Methodology

To find the matching result for input, we decided to do k-means clustering on the review data. For the best, result, we looked for various ways to preprocess the data or change the parameters.

Preprocess : Stemming using SnowballStemmer (NLTK). Filtering stop words.
Tf-idf to transform reviews into vectors(Max_df: 0.5, min_df: 100) -> 1300 features
LSA (Latent Semantic Analysis) => 200 features, ~60% of variance explained

K-Means Clustering : Processed Tf-idf vectors => initialization using k-means++
Visualization : Distance metric => 1 – cosine similarity(LSA_TF-idf_matrix)
Reduce dimension of distance matrix to 2 by using T-SNE => plot clusters

Metric

To assess the quality of our restaurant and review clustering,

we used 3 different RMSE (root mean squared error).

$$User-based: \sqrt{\sum_{i=1}^{user}\sum_{c=1}^{rest_i}(r_{i,c}-\mu_{i,c})^2}$$

$$Cluster-based: \sqrt{\sum_{j=1}^{users_j}\sum_{c=1}^{rest_j}(r_{i,c}-\mu_{j,c})^2}$$

$$Restaurant-based: \sqrt{\sum_{l=1}^{k}\sum_{c=1}^{k}(u_{l,c}-\mu_{l,c})^2}$$

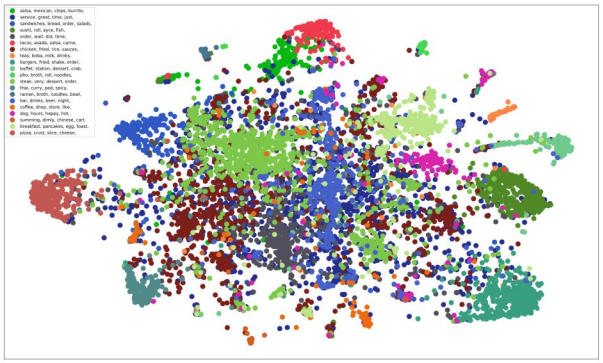
Here i = # of total users (558), j = # of user clusters

r = individual user's rating of each restaurant

u = rating of restaurants for each user

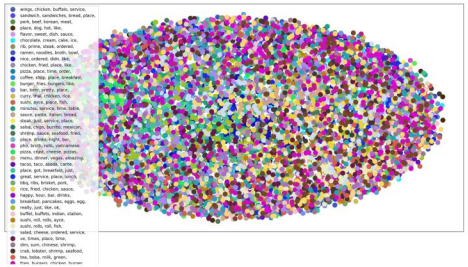
Clustering

Clustering the data only by text data :



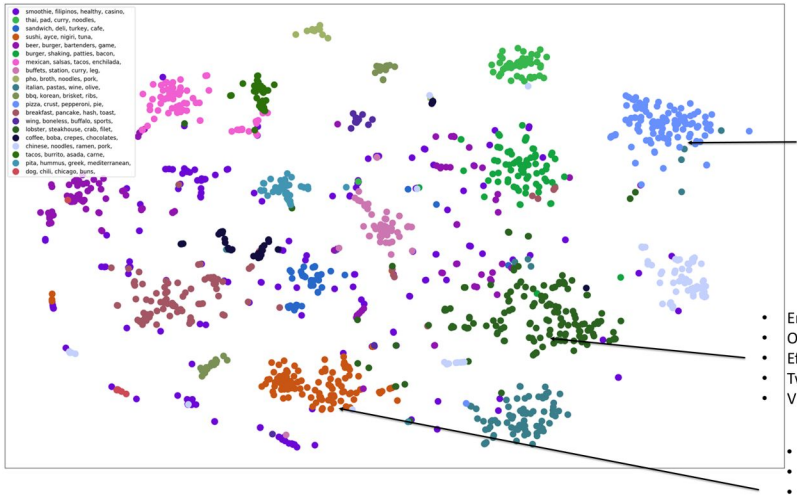
Challenges

Visualizing / Clustering High dimensional data
Bad example: Multi-dimensional Scaling (MDS) using Euclidean distance.



Results

K-means(k=20) clustering on text grouped by restaurants.



- California Pizza Kitchen
- Those Guys House
- Pizza Rev
- New York Pizza & Pasta
- Marco's Pizza

- Emeril's New Orleans Fish House
- Outback Steakhouse
- Effiel Tower Restauarant
- Twin Creeks
- Vic & Anthony Steakhouse

- Sushi Wow
- Sushi Twister
- Sushi Tachi
- Sushi Bar Sage
- Jumpice Rice & Rolls