Restaurant Recommendation System.

Rutuja Shah[1]   *Shalvi Desai* [2]

shah.rut@northeastern.edu   desai.shal@northeastern.edu

Problem definition
- Picking a restaurant based on features specific to a user is time-consuming given the number of options available so we need to model an application that combats this.
- The purpose of the project is to generate restaurant recommendations based on following:
    1. Finding similar restaurants based on a user's preference using content-based filtering.
    2. Finding similar restaurants by finding users who have similar taste using Collaborative filtering.

Existing methods
- Content-based:
    1. Product recommendation for E-commerce business using K-means clustering for customer segmentation and PCA for reducing the dimensions of different product and customer features.
    2. Content based article recommendation system in which the articles with similar contextual meaning are grouped together using K-means and Agglomerative clustering.
- Collaborative filtering:
    1. Active learning algorithms have been stacked upon the baseline approach to get rid of the cold start problem.
    2. Bayesian decision trees can be used to maximize the probability of items that can be recommended by a particular user, hence adding a personalization component.

Proposed methods
- Content-based recommendation system: Clustering the restaurants based on the cuisines they offer and then returning recommendations based on the cluster a user x is assigned to. Achieved by:
    1. K-means clustering, an unsupervised machine learning algorithm, paired with PCA.
    2. Agglomerative clustering, a hierarchical bottom-up algorithm.
- Collaborative filtering: Projecting the input matrix to the latent space and rank the most similar users to predict the rating of the current user.
    1. SVD projects utility matrix into r latent components to predict the rating .
    2. Neural networks uses utility matrix as input features and identifies the nonlinear relationship

Data description and experimental setup
We have used business.json and review.json from the yelp dataset/.

- Choosing optimal no. of clusters for K-means using the elbow method and choosing the optimal no. of clusters for Agglomerative using the DB index and silhouette coefficient.
- For SVD we have used different number of components to determine the value which retains the majority variance.
- For neural networks we have tried different dimensions for extracting the embedding of the utility matrix.


Distortion Score Elbow for KMeans Clustering — elbow at k = 13, score = 95.729
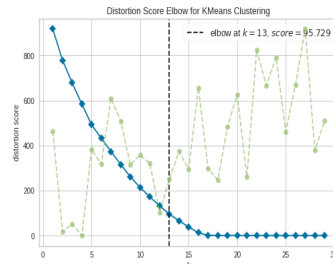
No. of clusters	DB Index	silhouette coefficient
6	0.508	0.584
7	0.477	0.631
8	0.438	0.678
9	0.396	0.725
10	0.355	0.766
11	0.314	0.803
12	0.269	0.839
13	0.219	0.875
14	0.179	0.903
15	0.133	0.932
16	0.081	0.961
17	0.039	0.974
18	0.057	0.878
19	0.090	0.865

Results and discussions
- Metrics David Boulder index and silhouette coefficient used to measure clustering quality. Both return similar results.
- Precision@k was used to assess the quality of predictions for SVD.
- RMSE for Neural Network : 1.5872

Method	Davis Boulder Index	silhouette coefficient
Kmeans	0.219	0.874
Agglomerative	0.219	0.875

Method	p@4	p@5	p@6
SVD	0.75	0.6	0.666

Ratings Histogram
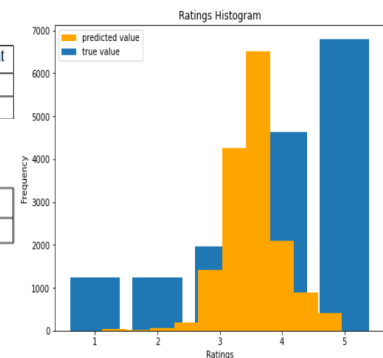
Take-Away points and future work
- Collaborative filtering requires no domain knowledge for extracting features for the model. Also users can explore new interests. This approach can be improved upon by adding certain extensions :
    1. Make the algorithm scalable by using a clustering approach beforehand.
    2. Make personalized user recommendation by adding bayesian network

- Content based recommendation systems using clustering techniques produced cohesive and well-separated clusters and as a result, returned good recommendations to a user instance. Can be enhanced by:
    1. Incorporating the use of collaborative filtering in the current model such that clusters are formed based on similarities between users in terms of their ratings and reviews provided.
    2. Incorporating the use of collaborative filtering in the current model which incorporates the use of sentiment analysis on reviews of similar users