Prerak Mehta

MSDS 453 - Natural Language Processing

Assignment 3 - Cluster Analysis and Multidimensional Scaling

Abstract:

This research is an add on to the research conducted in the second assignment. The restaurant review dataset will be continued to be used in this assignment. Using the numerical matrices produced using the three vectorization methods in the second research assignment, K-Means and Hierarchical cluster analysis will be performed along with Latent Dirichlet Allocation (LDA) and Spectral biclustering on the full documents-by-terms matrix.

Introduction:

From the second research assignment it was concluded that the count vectorization method yielded the maximum performing random classifier model. And hence all the matrix that will be used throughout this research assignment will be the documents-by-terms matrix produced by the count vectorization approach. This research is being conducted to go deeper into analyzing our entire corpus and see if there are any common themes across any documents or can they fall into identifiable groups or clusters.

Literature Review:

As discussed in the previous assignment a lot of platforms out there have a dataset full of customer reviews and other attributes related to them for the restaurants they visit. A lot of analysis and machine learning algorithms have been built around the dataset to provide many functionalities to the platform users such as popular restaurants, popular items in a restaurant, restaurant recommendation, image analysis, etc. In this research assignment there will be various types of cluster analysis done on the numerical matrices and explore our findings.

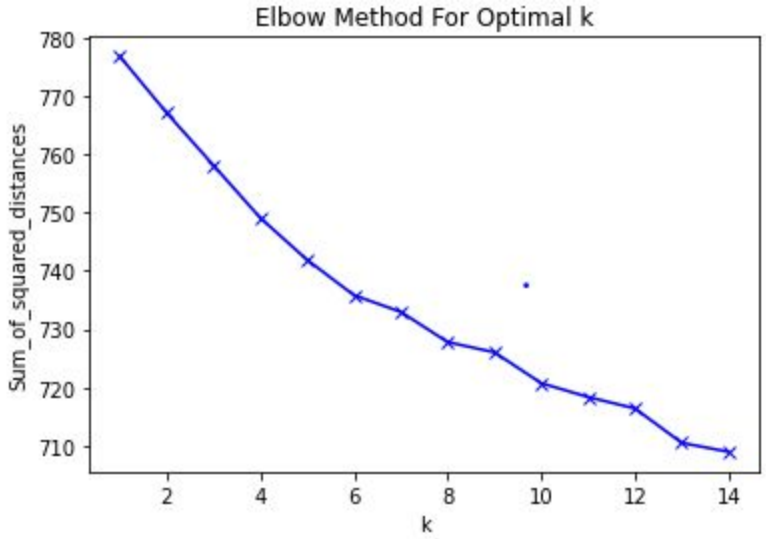
Methods:

Using the reduced matrix obtained from count vectorization in the previous assignment, cluster analysis (K-means) using multidimensional scaling and hierarchical scaling will be performed. The results from both these analyses will be shown in the python code. Elbow method was employed in order to determine the optimal number of clusters that should be used in the K-means cluster. Later LDA and Spectral biclustering will be employed on the full documents by terms matrix (reduced matrix as well).

Results:

The basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible. The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

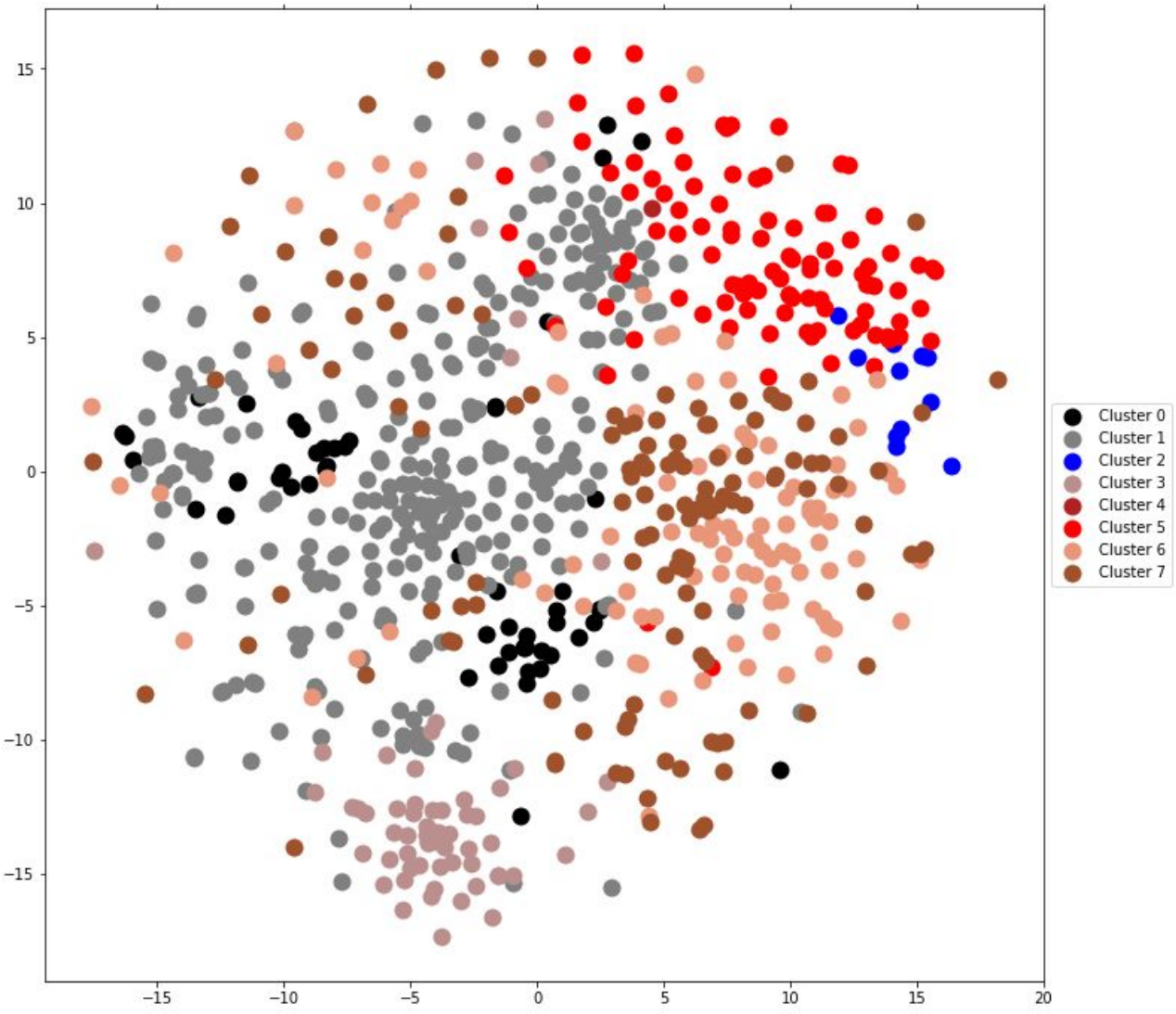
The following graph was obtained from the elbow method.



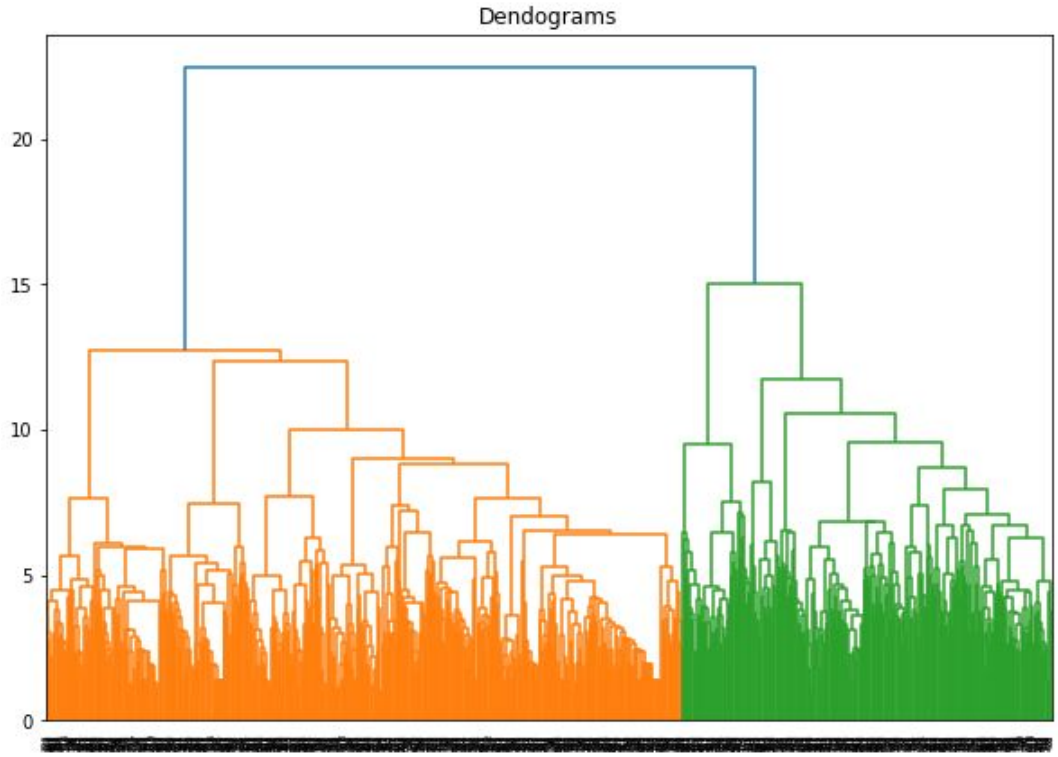
From this graph it can be determined that the elbow occurs when k=8. Later this k was used to fit a K-means cluster on the reduced matrix. A list of top terms from each cluster was then computer. A sample being:

Cluster 0: great, here, and, service, food, had, not, was, again, will.

Later multidimensional scaling was done using documents as objects. T-distributed stochastic neighbor embedding (t-SNE) was used for multidimensional scaling. The resulting plot was as below:



Another method used was hierarchical scaling. Using Dendograms, the number of clusters were determined. The resulting figure is as below:



From this graph we can see that the number of clusters to be used for hierarchical cluster is 5. The resulting array from fit_predict on the matrix is one of the outputs in the python file. Later LDA and Spectral biclustering were employed on the matrix as well. The resulting array outputs are shown in the python file as well.

Conclusion:

The goal of this research was to observe how useful the employed methods are to better understand the corpus. The K-means cluster provides the most wholesome idea of the corpus per the analysis in this research assignment. As mentioned in the previous assignment there are many ways the vectorization can be improved like expanding the corpus and adding extra layers of text preprocessing.