Cluster Analysis, ANN and Text Mining Project Report

CSC 177 - Section 01 and Section 02

Team: Import TeamName

Team Members: Lauren Prather, Vibhor Sagar, Rachel Mao, Santiago Bermudez, Mueed Khalid, Marco Toro, Jared Roque

Summary:

For this project, we had many different tasks to work on. Firstly, we had to perform clustering. For this part of the project, we would use the IMDB dataset and perform some basic preprocessing on the data before using it. We would sort the dataset by genre and remove unnecessary columns from it. We would then encode all of the columns with continuous values and then handle all of the missing values.After that, we would perform KMeans clustering with 11 clusters. We use 11 clusters in this case because there are 11 movie genres in the provided dataset. After this, we would try to find the optimum number of clusters for KMeans. By plotting the SSE vs the number of Clusters (*Elbow method), we were able to determine that the optimum number of clusters is 3. This is interesting because this is the number of unique title types there are in the data set. After this, we would perform hierarchical analysis and use different types of algorithms for clustering.We would use the single link, complete link and group average algorithms and then print out their dendrograms for the cluster analysis portion of our project. After this came text mining on the corpus that we were provided in the project. This part was fairly straightforward as all we had to do was create a count vector and a TF-IDF vector. For the count vector, we would use sk-text and vectorizer to print a sort of matrix where rows would represent the text and the columns would represent the words. After this, we would go through a similar process, but this time obtain the TF-IDF scores of all the words as opposed to their ‘count’ values. For the artificial neural network portion, we would first perform attribute value binarization by encoding certain columns to have values of either one or zero. We would then use artificial neural networks on the dataset and then other classification models as well for comparison.

Results:

ANN	0.86
ML Model	Accuracy on Test Set (Provide accuracy in %) 0.8505
Naïve Bayes	0.78
KNN	Also provide the K value for which you got the highest accuracy 0.8035 Our K value that got the highest accuracy was 100!
SVM	0.7
DT	1
Logit	0.7

Conclusion:

For the ANN and classification models portion of our project, that decision tree was the most accurate again, but this time followed by ANN. Our least accurate models this time were SVM and Logit. For this project, we learned how to use an SSE vs the number of clusters graph to find the optimum number of clusters, which was 3 and based on the number of unique title types. This result did surprise us a bit, as we were doing things based on the number of unique genres when working with KMeans and hierarchical analysis, which would have made the number of clusters 11. We were also able to visualize and further understand the process of hierarchical clustering by printing out the dendrograms for each algorithm that we used for clustering. We also learned that the purpose of TF-IDF is that it returns the value which is the product of the number of times that a word appears in a document and the inverse of the number of documents that contain that same word. It is used to help us determine the importance of some words and to help us find some of the most important words while ignoring the least important ones.