Experiment 7
CLUSTERING- KMEANS AND HAC

***Abstract*—The documents were clustered using K-Means and Hierarchical Agglomerative clustering utilising the Reuter's dataset as the corpus, which consisted of texts belonging to various classes. To test the power of clustering algorithms, the results were compared to the real clusters (classes) that were present as a ground truth, as well as the corpus.**

I. EXPERIMENTATION DONE

- Computation of K-means clustering on a subset of Reuters-21578 dataset.
- Computation of Single-link, Complete-link, GAAC and centroid of the documents in the Reuters-21578 dataset.

II. APPROACH

A. Data preprocessing and modelling

- Downloaded the reuter's dataset from the nltk library [1]
- The document ids of all the documents that belong to a small number of classes were taken from Reuter's dataset, which has 90 classes of documents.
- The papers used to implement Kmeans belonged to ten different classes, but the documents used to execute HAC belonged to three separate classes only.
- All papers that belonged to more than one class in the same set of classes were discarded.
- Used the ground truth data regarding the types of documents contained in the dataset and assigned them a label ranging from 0-9 for documents that need to be clustered using Kmeans and 0-2 for documents that require HAC.
- Created a dictionary of document ids and terms using the text content of the document ids selected in the previous phases.
- Built a TF-IDF matrix of the corpus of documents picked using sklearn's TfidfVectorizer.

B. Kmeans

- Using the Tf-Idf matrix, we performed Kmeans clustering using both our own implementation and the sklearn library's implementation.
- The linear assignment utility of the sklearn library was used to align the labels obtained by Kmeans clustering and the labels given before.
- To compare ground truth classes and clustering outcomes, we used various metrics such as Purity, Nmi, RI, and F1 scores.

C. HAC

- The dendograms using different linkage methods, such as single-link, complete-link, average, centroid, and ward, were plotted using the scipy package.
- Using the Tf-Idf matrix, the sklearn library was used to conduct Hierarchical Agglomerative clustering.
- The linear assignment tool of the sklearn library was used to align the labels generated by the hac for different methods with the labels given before.
- For each of the four linkage methods used for clustering, the adjusted rand index was calculated.

III. RESULT

The overall accuracy of the Kmeans implementation was around 51%. Table 1 compares the purity normalised mutual information and adjusted rand index for Kmeans clustering to the ground truth classes.

Purity	**Normalized Mutual Information**	**Adjusted Rand Index**
0.8456	0.5414	0.3570

Table 1. Purity, NMI and RI for Kmeans

In almost all of the classes, Kmeans clustering produced false positives and false negatives. Detailed description for each class label:-

Class	**True Positive**	**True Negative**	**False Positive**	**False Negative**
0	2239	5335	547	80
1	0	7395	806	0
2	380	7726	17	78
3	944	4270	1	2986
4	6	7143	945	107
5	225	7774	137	65
6	349	7313	389	150
7	0	7551	484	166
8	70	7625	151	355
9	0	7689	511	1

Table 2. True Positives, True Negatives, False Positives and False Negatives for kmeans clustering

Only the ward linkage approach produced well-defined clusters when the dendograms were plotted for all of the different linking methods. Clusters produced by the single-link, complete-link, centroid, and average approaches were quite unsatisfactory. The clusters generated by the centroid technique were the worst, and there was little variation between them. The computed Adjusted Rand Index for the

various linking methods reveals similar results, namely, very poor clustering in the case of single, complete, and average linkage methods, and some improved results in the case of ward linkage approaches. Table 3 contains the rand indexes for various approaches.

Linkage	**Adjusted Rand Index**
Single-Link	-6.0790e-05
Complete-Link	0.078515
Average	0.078515
Ward	0.384413

IV. CONCLUSION

On a small subset of texts from the Reuters dataset, we were able to apply Kmeans and Hierarchical Agglomerative Clustering and compare the results to the ground truth. We also analysed the various HAC linkage methods, and found that the centroid linkage approach produced the lowest results, with essentially no clusters forming, while the ward method produced the best results.

V. GITHUB LINK

Click here to visit the Lab7 GitHub repository.

REFERENCES

[1] URL: https://www.cs.bgu.ac.il/ elhadad/nlp16/ReutersDataset.html