Machine Learning Assignment 2

Ankush Dey (MDS202108), Anjali Pugalia (MDS202107)

PROCESS

- We started by exporting all the necessary libraries. Then we read and processed the text files.
- We then used numpy and a for loop to create a sparse matrix where the wordIDs were the rows and the docIDs the columns. We did this manually.
- Then using the inbuilt python function pairwise distance we calculated the jaccard index and after subtracting it from one we got the jaccard similarity matrix to get a square matrix with dimension the number of unique docIds.
- From the library sklearn we utilised the inbuilt kmeans function. We ran a for loop and found out the inertia for a range of k values. On plotting the inertia for all these different kmeans we found the best fit (Elbow Method).
- We reduced the dimension of the Jaccard Matrix using PCA and then plotted the subsequent points which helped in visualising the clusters.

REMARKS

- NIPS DATASET : This dataset had approximately 1500 documents and 12419 unique words in the vocabulary. The dataset was not very huge and ran smoothly. The time taken to run the entire algorithm was 50 secs. According to the graph of the different values of inertia k=4 would give the best fit.
- KOS DATASET: This was a relatively small dataset with 3430 documents and 6906 unique words. The process to generate the optimal value of the number of clusters was the same as that of the NIPS dataset. The best value for k was 2 according to the elbow method. The time complexity for this program was 59 secs.
- ENRON DATASET: This dataset was vast and huge. It had approximately 39861 documents and 28102 unique words. It was impossible to use the earlier method on the dataset without optimising it. So we reduced the dataset by taking stratified sampling. So here we have multiple stratas o based on the frequency of the word. On this

reduced dataset we used the same procedure to get the jaccard similarity matrix and then the kmeans clustering. This dataset was huge and our laptops stopped working a few times in between, so it takes much higher time to solve the dataset. The process to generate the optimal value of the number of clusters was the same as that of the NIPS dataset. The best value for k was 3 according to the elbow method.

	Nips	Kos	Enron
Time taken	50 seconds	81 seconds	3401 seconds
Memory Used	341 MiB	880 MiB	4446 MiB