DMML ASSIGNMENT

Group members

1)Jerin Biju (MDS202329)

2)Aparna C (MDS202310)

Our assignment consists of 5 functions

Jac_distance

Merge

Clustering

Find_centroid

File_convert

Jac_distance: this function has two list arguments . each list consists of vocabulary Ids of words which are present in that document .Jaccard distance is calculated as 1-(intersection of two documents / union of two documents) based on cardinality of the documents

This function returns a metric in range of [0,1] as a measure of distance between documents

Merge: This function is used to merge two documents in the sorted order. Here we are exploiting the fact that vocabulary Ids in each documents are given in sorted order so to merge in sorted order we can do in O(n+m).The final merged version of documents will be the union of the two argument document. That is there wont be repetition of words

Find_centroid: This function is having a list as argument . This list is actually a list of ids . each Ids corresponds to a document. So in fact the argument of this function is a list of documents . Now here we are assuming Union of ‘n’ documents as the centroid of ‘n’ documents. So we are using Merge function inside this function to find the centroid list and returns that list

Clustering: This function has 3 arguments data: which is a dictionary of documents, n:number of documents, k: number of clusters. In this function, we

initialises k random documents from data as centroids. We initialise a dictionary di of length k where each key value corresponds to a cluster which is associated with each centroids. Now we will travel through the entire data and assign each document to one of the clusters using jaccard distance between the document and centroid. Now we will recalculate the new centroid of each cluster. Again we will travel through the entire data and assign documents to clusters using the new centroids. This procedure is repeated until either the jaccard distance between two consecutive centroids become less than 1 or the number of iterations exceeds 100. After the iterations we will calculate the inertia, which is the sum of squared jaccard distances of each document from its centroid. Inertia is returned after executing this function.

File_convert: This function converts data given as csv file into a dictionary with key as document id and value as a list of word ids each document contain.

For each documents collection, we plotted number of clusters vs inertia. From the graphs, by observing an approximate elbow,

For the first document collection, enron, optimum k turned out to be 7.

For the second document collection, optimum k turned out to be 6.

For the third document collection, optimum k is 8