Programming Assignment 4 - Clustering Report

Submitted By: Ankita Mehta

November 27 , 2017

1 Clustering

1.1 Implementing Agglomerative Clustering

After creating compressed, uncompressed indices and Retrieval models (in the last assignments) and next task is to cluster the similar documents present in the shakespeare dataset. This has been implemented in the following steps:

1. Firstly indexes and vocabulary is created (part of previous assignment) and then for each document, a document vector is made as a numpy array of length equals vocabulary length. Document vector has normalised tf*idf weights for all the terms present in that document.
2. After the document vectors are created, agglomerative clustering is performed. Beginning with the 1st document present in the index, documents are clustered in the order of their occurrence in the index.
3. Agglomerative is an online clustering, so as the documents are coming into the stream, either they have been added to the existing cluster or made a new cluster based on the threshold. This has been experimented with a range of thresholds from 0.05 to 0.095 with step size 0.05. Four linking choices (min, max, average , mean) have been considered while finding out the distance been document and the cluster.
4. All the linking choices, VSM implementation has been written in one code and indexing, clustering has been written in different code.

1.2 Design Tradeoffs:

1. Cluster numbers are starting from 1 whereas document numbers are starting from 0.
2. Sparse vectors have been stored in the numpy arrays for faster implementation.
3. 1 - Cosine similarity has been used as a distance measure. Since it was already implemented in the previous assignment, so this will test the working of VSM model as well clustering.

1.3 Various Questions and how they were approached:

Various difficulties that were faced while doing the project and I figured out these problems after having discussion with the professor and peers

1. The very fist question that came when I started working on this assignment was Can we use sklearn agglomeartive clustering ?
2. What all things to be tested and run ? Results should only be for mean or for all choices ?
3. What should be the step size for the thresholds ?
4. Is it fine to use any method of distance computation , either 1 - cosine similarity or euclidean distance ?

2 What happens as the threshold value increases ??

Solution:

Following are the clustering results of 4 linking choices with different threshold values:

Min Linkage :

Threshold Value	Number of Clusters	Size of Clusters
0.05	748	All clusters with size 1
0.1	748	All clusters with size 1
0.15	748	All clusters with size 1
0.2	748	All clusters with size 1
0.25	748	All clusters with size 1
0.3	748	All clusters with size 1
0.35	748	All clusters with size 1
0.4	748	All clusters with size 1
0.45	748	All clusters with size 1
0.5	748	All clusters with size 1
0.55	748	All clusters with size 1
0.6	747	1,2
0.65	746	1,2
0.7	739	1,2,3
0.75	686	1,2,3,4,5
0.8	512	1,2,3,4,5,6,7,8,9,10,12
0.85	233	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,26,32,35,46
0.9	31	1,2,3,4,5,10,26,34,35,598
0.95	1	748

Max Linkage :

Threshold Value	Number of Clusters	Size of Clusters
0.05	748	All clusters with size 1
0.1	748	All clusters with size 1
0.15	748	All clusters with size 1
0.2	748	All clusters with size 1
0.25	748	All clusters with size 1
0.3	748	All clusters with size 1
0.35	748	All clusters with size 1
0.4	748	All clusters with size 1
0.45	748	All clusters with size 1
0.5	748	All clusters with size 1
0.55	748	All clusters with size 1
0.6	747	1,2
0.65	746	1,2
0.7	741	1,2
0.75	698	1,2,3
0.8	581	1,2,3,4,5,6
0.85	400	1,2,3,4,5,6,8
0.9	237	1,2,3,4,5,6,7,8,9
0.95	117	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16

Average Linkage :

Threshold Value	Number of Clusters	Size of Clusters
0.05	748	All clusters with size 1
0.1	748	All clusters with size 1
0.15	748	All clusters with size 1
0.2	748	All clusters with size 1
0.25	748	All clusters with size 1
0.3	748	All clusters with size 1
0.35	748	All clusters with size 1
0.4	748	All clusters with size 1
0.45	748	All clusters with size 1
0.5	748	All clusters with size 1
0.55	748	All clusters with size 1
0.6	747	1,2
0.65	746	1,2
0.7	741	1,2
0.75	694	1,2,3,4
0.8	563	1,2,3,4,5,6,8
0.85	348	1,2,3,4,5,6,8,9
0.9	153	1-15,18,19
0.95	43	1-11,13-21,23,25, 26, 28, 40, 82, 85

Mean Linkage :

Threshold Value	Number of Clusters	Size of Clusters
0.05	748	All clusters with size 1
0.1	748	All clusters with size 1
0.15	748	All clusters with size 1
0.2	748	All clusters with size 1
0.25	748	All clusters with size 1
0.3	748	All clusters with size 1
0.35	748	All clusters with size 1
0.4	748	All clusters with size 1
0.45	748	All clusters with size 1
0.5	748	All clusters with size 1
0.55	748	All clusters with size 1
0.6	747	1,2
0.65	746	1,2
0.7	737	1,3,4
0.75	657	1,2,3,4,5,6,7,8,9,10
0.8	295	1,2,3,4,5,6,7,8,9,10,12,13,18,293
0.85	105	1,2,3,4,5,6,7,8,9,13,14,16,19,22,483
0.9	22	1,2,3,4,9,14,22,670
0.95	2	745,3

As we can see from the results above: Independent of the linking choice, as the threshold value increases, number of clusters decreases. This happened because with the increase in the cap, more documents can lie in a single cluster. Hence number of clusters decreases.

3 Explain difference between the four linking strategies.

Solution:
All four linkage methods (min , max, average and mean) computes the cost between two clusters by comparing each instance of one cluster to every instance of other cluster in different ways.

Min Linkage : It relies only on the minimum distance between 2 clusters or between a cluster and a document . It doesn't consider how far apart the remainder of instances in a cluster can be. So in this case, clusters can be very long.

Max Linkage : It relies only on the maximum distance between 2 clusters or between a cluster and a document . Clusters tend to be more compact and less spread out as in min linkage.

This can be seen in the above results as well that the maximum cluster size with threshold 0.9 is 598 for min linkage whereas it is 9 for max linkage.

Average Linkage : It is like a compromise between min and max linkage. As min takes the nearest neighbour , max takes the farthest neighbour whereas average linkage averages of all the pairwise distances between 2 clusters or between a cluster and document. Here the size of clusters formed is largely dependent on the structure of clusters because here we are taking an average of all the distances.

It can also be seen in the above results that the size of clusters obtained from average linking are not as big as min linkage and not as short as max linkage.

Mean Linkage : It is closely related to the average linkage. It firstly computes centroid of the cluster which is the average of all the instances present in the cluster and then computes distance between two centroids (if we are taking two clusters). Otherwise it computes the distance between document and the centroid.

It can easily be seen from the above results that for the same threshold, mean linkage gave either same or lesser number of clusters than the number of clusters made by max, min or average linkages. Also mean linkage make very big clusters as compared to other linkage choices for the same threshold.