CMSC 476/676 Information Retrieval
Homework 5 REPORT
JA52979
Sai Teja Challa

Steps to run the program:

In the IDE terminal, You can run the program described below:
python3 cluster.py.

Here, 'cluster.py' is the main python program. You must have all libraries installed and your current working directory containing python files in order to run the method described above.

Detailed summary of the code:

In this phase, we need to find the similarity matrix of the entire corpus and perform agglomerative clustering. Here, we are not taking any input directory paths from the user. So, I used the output files from phase-2 i.e,. term weights of each document and included the path of those files in the code to make document vectors.

So, to make document vectors for each document data, I created a function *'load_document_vectors'*, which iterates over all the output files of phase-2. For each file, it initializes a dictionary where the keys are terms and the values are their corresponding TF-IDF weights. The function parses out the TF-IDF value and stores it in the dictionary. The function returns a dictionary where each key is a document ID and the value is another dictionary (the document vector). Thus, using these document vectors cosine similarity is calculated and a similarity matrix is built.

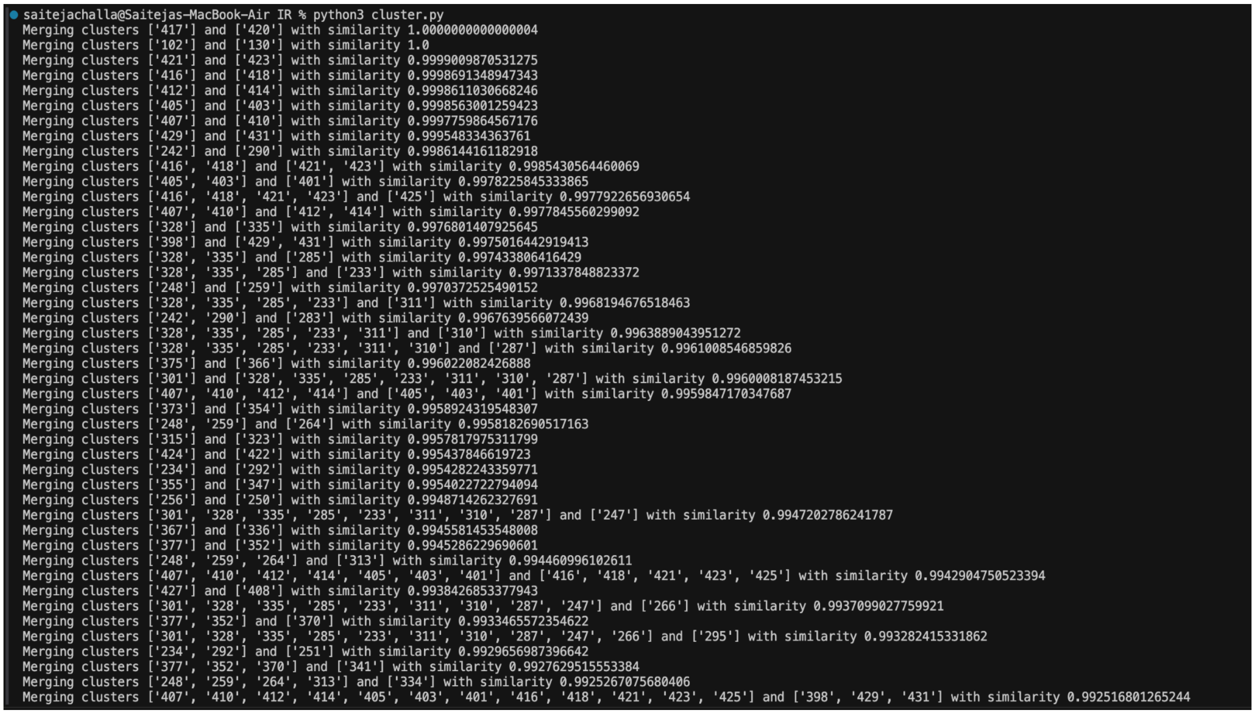
Here, a similarity matrix is built using the *'build_similarity_matrix'* function. This function initializes a square matrix of size equal to the number of documents, with zeros, and later fills the matrix by computing the cosine similarity between each pair of document vectors using the *'cosine_similarity function'*. Then it returns the similarity matrix along with a list of document IDs corresponding to the indices in the matrix. Also, the function *'cosine_similarity function'* takes two document vectors as input, and identifies the intersection of terms between the two documents. Then calculates the dot product of the vectors for the intersecting terms and the norm (Euclidean length) of each vector. Thus, the dot product is divided by the product of the norms, resulting in 1 indicates an identical direction, while 0 indicates orthogonality (no similarity).

Now, using the similarity matrix, I performed the hierarchical clustering using the *‘hierarchical_clustering’* function. Here, in the function each document is initially treated as a singleton cluster. This is represented by a dictionary where each key is an index and the value is a list containing the document ID corresponding to that index. A set named *‘active_indices’* is maintained to keep track of clusters that are still active (i.e., have not been merged into another cluster). The function repeatedly finds the two closest clusters and merges them until only one cluster remains or until no pair of clusters has a similarity above a predefined threshold (e.g., 0.4). It iterates over all active clusters to find the pair with the highest similarity, by checking the current similarity matrix where each entry (i, j) represents the similarity between clusters i and j.

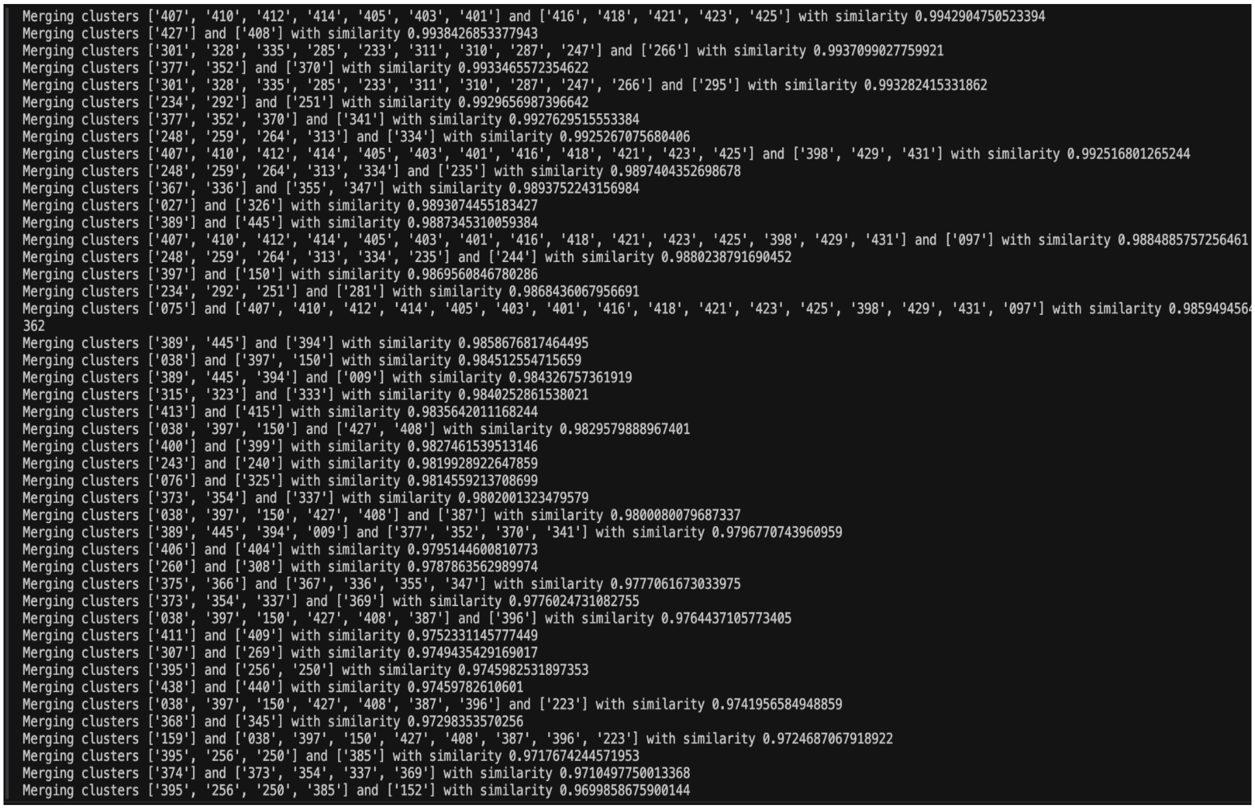
Once the closest pair is found, the clusters are merged. This involves combining the document lists of the two clusters. The merged cluster retains the index of one of the original clusters, and the other index is removed from *‘active_indices’*. The document list for the merged cluster is updated to include documents from both original clusters.

After merging, the similarity matrix needs to be updated. The similarities involving the merged cluster are recalculated as the maximum similarity between the merged cluster and all other active clusters. This choice of updating rule depends on the linkage criterion used; in this case, it resembles the complete linkage method where the similarity between two clusters is the maximum of the similarities between any member of one cluster to any member of the other. The row and column in the similarity matrix corresponding to the removed cluster are marked inactive (e.g., set to -1). The process stops when there is only one cluster left or when the highest similarity between any two clusters falls below the threshold, indicating that the remaining clusters are sufficiently dissimilar that further merging would not be meaningful.

Hence, the clustering process is done, the function prints messages indicating which clusters are being merged and their similarity.

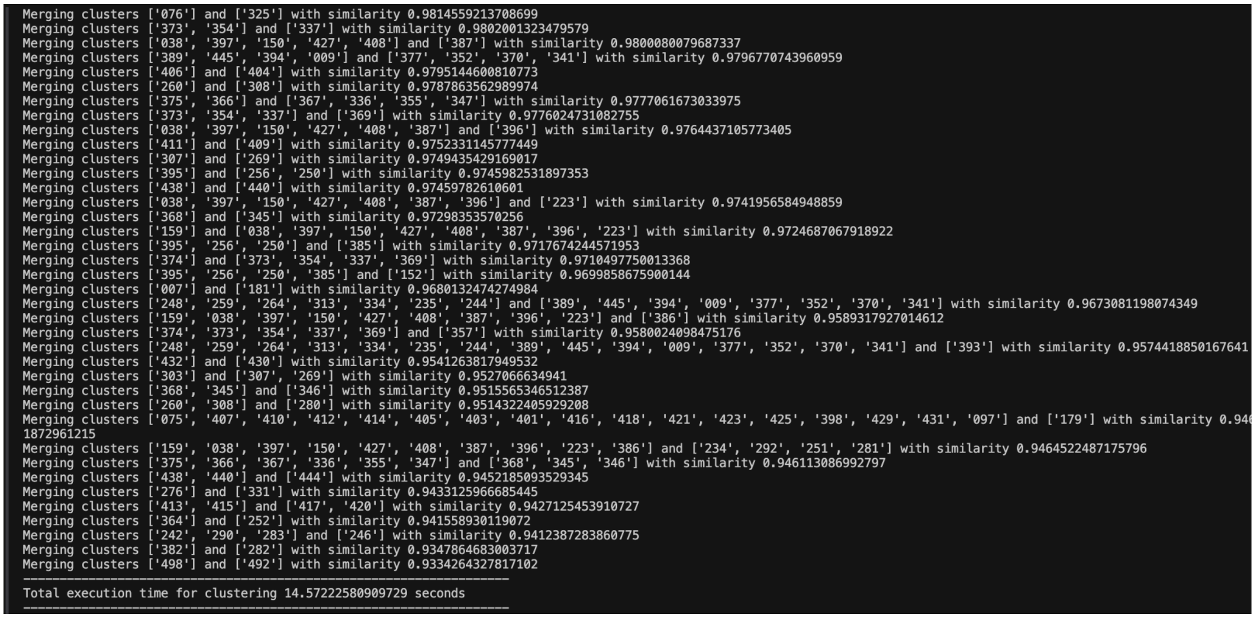


```
● saitejachalla@Saitejas-MacBook-Air IR % python3 cluster.py
  Merging clusters ['417'] and ['420'] with similarity 1.0000000000000004
  Merging clusters ['102'] and ['130'] with similarity 1.0
  Merging clusters ['421'] and ['423'] with similarity 0.9999009870531275
  Merging clusters ['416'] and ['418'] with similarity 0.9998691348947343
  Merging clusters ['412'] and ['414'] with similarity 0.9998611030668246
  Merging clusters ['405'] and ['403'] with similarity 0.9998563001259423
  Merging clusters ['407'] and ['410'] with similarity 0.9997759864567176
  Merging clusters ['429'] and ['431'] with similarity 0.999548334363761
  Merging clusters ['242'] and ['290'] with similarity 0.9986144161182918
  Merging clusters ['416', '418'] and ['421', '423'] with similarity 0.9985430564460069
  Merging clusters ['405', '403'] and ['401'] with similarity 0.9978225845333865
  Merging clusters ['416', '418', '421', '423'] and ['425'] with similarity 0.9977922656930654
  Merging clusters ['407', '410'] and ['412', '414'] with similarity 0.9977845560299092
  Merging clusters ['328'] and ['335'] with similarity 0.9976801407925645
  Merging clusters ['398'] and ['429', '431'] with similarity 0.9975016442919413
  Merging clusters ['328', '335'] and ['285'] with similarity 0.997433806416429
  Merging clusters ['328', '335', '285'] and ['233'] with similarity 0.9971337848823372
  Merging clusters ['248'] and ['259'] with similarity 0.9970372525490152
  Merging clusters ['328', '335', '285', '233'] and ['311'] with similarity 0.9968194676518463
  Merging clusters ['242', '290'] and ['283'] with similarity 0.9967639566072439
  Merging clusters ['328', '335', '285', '233', '311'] and ['310'] with similarity 0.9963889043951272
  Merging clusters ['328', '335', '285', '233', '311', '310'] and ['287'] with similarity 0.9961008546859826
  Merging clusters ['375'] and ['366'] with similarity 0.996022082426888
  Merging clusters ['301'] and ['328', '335', '285', '233', '311', '310', '287'] with similarity 0.9960008187453215
  Merging clusters ['407', '410', '412', '414'] and ['405', '403', '401'] with similarity 0.9959847170347687
  Merging clusters ['373'] and ['354'] with similarity 0.9958924319548307
  Merging clusters ['248', '259'] and ['264'] with similarity 0.9958182690517163
  Merging clusters ['315'] and ['323'] with similarity 0.9957817975311799
  Merging clusters ['424'] and ['422'] with similarity 0.995437846619723
  Merging clusters ['234'] and ['292'] with similarity 0.9954282243359771
  Merging clusters ['355'] and ['347'] with similarity 0.9954022722794094
  Merging clusters ['256'] and ['250'] with similarity 0.9948714262327691
  Merging clusters ['301', '328', '335', '285', '233', '311', '310', '287'] and ['247'] with similarity 0.9947202786241787
  Merging clusters ['367'] and ['336'] with similarity 0.9945581453548008
  Merging clusters ['377'] and ['352'] with similarity 0.9945286229690601
  Merging clusters ['248', '259', '264'] and ['313'] with similarity 0.994460996102611
  Merging clusters ['407', '410', '412', '414', '405', '403', '401'] and ['416', '418', '421', '423', '425'] with similarity 0.9942904750523394
  Merging clusters ['427'] and ['408'] with similarity 0.9938426853377943
  Merging clusters ['301', '328', '335', '285', '233', '311', '310', '287', '247'] and ['266'] with similarity 0.9937099027759921
  Merging clusters ['377', '352'] and ['370'] with similarity 0.9933465572354622
  Merging clusters ['301', '328', '335', '285', '233', '311', '310', '287', '247', '266'] and ['295'] with similarity 0.993282415331862
  Merging clusters ['234', '292'] and ['251'] with similarity 0.9929656987396642
  Merging clusters ['377', '352', '370'] and ['341'] with similarity 0.9927629515553384
  Merging clusters ['248', '259', '264', '313'] and ['334'] with similarity 0.9925267075680406
  Merging clusters ['407', '410', '412', '414', '405', '403', '401', '416', '418', '421', '423', '425'] and ['398', '429', '431'] with similarity 0.992516801265244
```



```
  Merging clusters ['407', '410', '412', '414', '405', '403', '401'] and ['416', '418', '421', '423', '425'] with similarity 0.9942904750523394
  Merging clusters ['427'] and ['408'] with similarity 0.9938426853377943
  Merging clusters ['301', '328', '335', '285', '233', '311', '310', '287', '247'] and ['266'] with similarity 0.9937099027759921
  Merging clusters ['377', '352'] and ['370'] with similarity 0.9933465572354622
  Merging clusters ['301', '328', '335', '285', '233', '311', '310', '287', '247', '266'] and ['295'] with similarity 0.993282415331862
  Merging clusters ['234', '292'] and ['251'] with similarity 0.9929656987396642
  Merging clusters ['377', '352', '370'] and ['341'] with similarity 0.9927629515553384
  Merging clusters ['248', '259', '264', '313'] and ['334'] with similarity 0.9925267075680406
  Merging clusters ['407', '410', '412', '414', '405', '403', '401', '416', '418', '421', '423', '425'] and ['398', '429', '431'] with similarity 0.992516801265244
```

```
  Merging clusters ['248', '259', '264', '313', '334'] and ['235'] with similarity 0.9897404352698678
  Merging clusters ['367', '336'] and ['355', '347'] with similarity 0.9893752243156984
  Merging clusters ['027'] and ['326'] with similarity 0.9893074455183427
  Merging clusters ['389'] and ['445'] with similarity 0.9887345310059384
  Merging clusters ['407', '410', '412', '414', '405', '403', '401', '416', '418', '421', '423', '425', '398', '429', '431'] and ['097'] with similarity 0.9884885757256461
  Merging clusters ['248', '259', '264', '313', '334', '235'] and ['244'] with similarity 0.9880238791690452
  Merging clusters ['397'] and ['150'] with similarity 0.9869560846780286
  Merging clusters ['234', '292', '251'] and ['281'] with similarity 0.9868436067956691
  Merging clusters ['075'] and ['407', '410', '412', '414', '405', '403', '401', '416', '418', '421', '423', '425', '398', '429', '431', '097'] with similarity 0.9859494564
  362
  Merging clusters ['389', '445'] and ['394'] with similarity 0.9858676817464495
  Merging clusters ['038'] and ['397', '150'] with similarity 0.984512554715659
  Merging clusters ['389', '445', '394'] and ['009'] with similarity 0.984326757361919
  Merging clusters ['315', '323'] and ['333'] with similarity 0.9840252861538021
  Merging clusters ['413'] and ['415'] with similarity 0.9835642011168244
  Merging clusters ['038', '397', '150'] and ['427', '408'] with similarity 0.9829579888967401
  Merging clusters ['400'] and ['399'] with similarity 0.9827461539513146
  Merging clusters ['243'] and ['240'] with similarity 0.9819928922647859
  Merging clusters ['076'] and ['325'] with similarity 0.9814559213708699
  Merging clusters ['373', '354'] and ['337'] with similarity 0.9802001323479579
  Merging clusters ['038', '397', '150', '427', '408'] and ['387'] with similarity 0.9800080079687337
  Merging clusters ['389', '445', '394', '009'] and ['377', '352', '370', '341'] with similarity 0.9796770743960959
  Merging clusters ['406'] and ['404'] with similarity 0.9795144600810773
  Merging clusters ['260'] and ['308'] with similarity 0.9787863562989974
  Merging clusters ['375', '366'] and ['367', '336', '355', '347'] with similarity 0.9777061673033975
  Merging clusters ['373', '354', '337'] and ['369'] with similarity 0.9776024731082755
  Merging clusters ['038', '397', '150', '427', '408', '387'] and ['396'] with similarity 0.9764437105773405
  Merging clusters ['411'] and ['409'] with similarity 0.9752331145777449
  Merging clusters ['307'] and ['269'] with similarity 0.9749435429169017
  Merging clusters ['395'] and ['256', '250'] with similarity 0.9745982531897353
  Merging clusters ['438'] and ['440'] with similarity 0.97459782610601
  Merging clusters ['038', '397', '150', '427', '408', '387', '396'] and ['223'] with similarity 0.9741956584948859
  Merging clusters ['368'] and ['345'] with similarity 0.97298353570256
  Merging clusters ['159'] and ['038', '397', '150', '427', '408', '387', '396', '223'] with similarity 0.9724687067918922
  Merging clusters ['395', '256', '250'] and ['385'] with similarity 0.9717674244571953
  Merging clusters ['374'] and ['373', '354', '337', '369'] with similarity 0.9710497750013368
  Merging clusters ['395', '256', '250', '385'] and ['152'] with similarity 0.9699858675900144
```

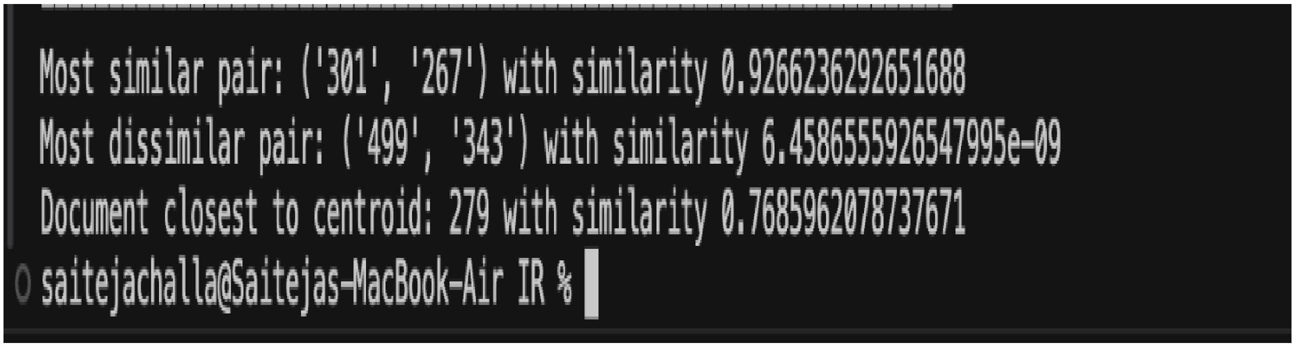


```
Merging clusters ['076'] and ['325'] with similarity 0.9814559213708699
Merging clusters ['373', '354'] and ['337'] with similarity 0.9802001323479579
Merging clusters ['038', '397', '150', '427', '408'] and ['387'] with similarity 0.9800080079687337
Merging clusters ['389', '445', '394', '009'] and ['377', '352', '370', '341'] with similarity 0.9796770743960959
Merging clusters ['406'] and ['404'] with similarity 0.9795144600810773
Merging clusters ['260'] and ['308'] with similarity 0.9787863562989974
Merging clusters ['375', '366'] and ['367', '336', '355', '347'] with similarity 0.9777061673033975
Merging clusters ['373', '354', '337'] and ['369'] with similarity 0.9776024731082755
Merging clusters ['038', '397', '150', '427', '408', '387'] and ['396'] with similarity 0.9764437105773405
Merging clusters ['411'] and ['409'] with similarity 0.9752331145777449
Merging clusters ['307'] and ['269'] with similarity 0.9749435429169017
Merging clusters ['395'] and ['256', '250'] with similarity 0.9745982531897353
Merging clusters ['438'] and ['440'] with similarity 0.97459782610601
Merging clusters ['038', '397', '150', '427', '408', '387', '396'] and ['223'] with similarity 0.9741956584948859
Merging clusters ['368'] and ['345'] with similarity 0.97298353570256
Merging clusters ['159'] and ['038', '397', '150', '427', '408', '387', '396', '223'] with similarity 0.9724687067918922
Merging clusters ['395', '256', '250'] and ['385'] with similarity 0.9717674244571953
Merging clusters ['374'] and ['373', '354', '337', '369'] with similarity 0.9710497750013368
Merging clusters ['395', '256', '250', '385'] and ['152'] with similarity 0.9699858675900144
Merging clusters ['007'] and ['181'] with similarity 0.9680132474274984
Merging clusters ['248', '259', '264', '313', '334', '235', '244'] and ['389', '445', '394', '009', '377', '352', '370', '341'] with similarity 0.9673081198074349
Merging clusters ['159', '038', '397', '150', '427', '408', '387', '396', '223'] and ['386'] with similarity 0.9589317927014612
Merging clusters ['374', '373', '354', '337', '369'] and ['357'] with similarity 0.9580024098475176
Merging clusters ['248', '259', '264', '313', '334', '235', '244', '389', '445', '394', '009', '377', '352', '370', '341'] and ['393'] with similarity 0.9574418850167641
Merging clusters ['432'] and ['430'] with similarity 0.9541263817949532
Merging clusters ['303'] and ['307', '269'] with similarity 0.9527066634941
Merging clusters ['368', '345'] and ['346'] with similarity 0.9515565346512387
Merging clusters ['260', '308'] and ['280'] with similarity 0.9514322405929208
Merging clusters ['075', '407', '410', '412', '414', '405', '403', '401', '416', '418', '421', '423', '425', '398', '429', '431', '097'] and ['179'] with similarity 0.946
1872961215
Merging clusters ['159', '038', '397', '150', '427', '408', '387', '396', '223', '386'] and ['234', '292', '251', '281'] with similarity 0.9464522487175796
Merging clusters ['375', '366', '367', '336', '355', '347'] and ['368', '345', '346'] with similarity 0.946113086992797
Merging clusters ['438', '440'] and ['444'] with similarity 0.9452185093529345
Merging clusters ['276'] and ['331'] with similarity 0.9433125966685445
Merging clusters ['413', '415'] and ['417', '420'] with similarity 0.9427125453910727
Merging clusters ['364'] and ['252'] with similarity 0.941558930119072
Merging clusters ['242', '290', '283'] and ['246'] with similarity 0.9412387283860775
Merging clusters ['382'] and ['282'] with similarity 0.9347864683003717
Merging clusters ['498'] and ['492'] with similarity 0.9334264327817102
-------------------------------------------------------------------
Total execution time for clustering 14.57222580909729 seconds
-------------------------------------------------------------------
```

Above images show the first 100 lines of output.

Finding most similar and most dissimilar pairs:

Using the *'find_similar_and_dissimilar_pairs'* function, the most similar and dissimilar pairs are found; It initializes variables to track the maximum and minimum similarities and their corresponding document pairs, iterates over the upper triangle of the similarity matrix (excluding the diagonal) to identify the pairs with the highest and lowest similarity values and returns the document pairs and their similarities for both the most similar and most dissimilar pairs. The document closest to the corpus centroid is found using the *'find_closest_to_centroid'* function. It initializes variables to track the highest similarity and the corresponding document, iterates over all document vectors, computing the cosine similarity to the centroid for each. Hence, the document with the highest similarity to the centroid is considered the closest, using cosine similarity as the metric.



```
Most similar pair: ('301', '267') with similarity 0.9266236292651688
Most dissimilar pair: ('499', '343') with similarity 6.458655926547995e-09
Document closest to centroid: 279 with similarity 0.7685962078737671
○ saitejachalla@Saitejas-MacBook-Air IR %
```

So, the most similar and most dissimilar pair of html documents, and the document closest to centroid is depicted above.

Major data structures used:

Here, each document is represented as a vector where each dimension corresponds to a term's TF-IDF weight. This is stored in a dictionary where each key is a document ID, and the value is another dictionary mapping terms to their TF-IDF weights. Clusters are implemented using a dictionary where the key is an index (initially corresponding to document IDs) and the value is a list of document IDs in the cluster. This data structure supports efficient merging of clusters and updating of the similarity matrix. A set *'active_indices'* used to keep track of active cluster indices in the similarity matrix. This helps in iterating only over currently active clusters during the merging process.

Complexity:

The computation and updating of the similarity matrix throughout the merging process are the main factors influencing the complexity of the hierarchical agglomerative clustering algorithm. Since cosine similarity must be used to compare each pair of documents, creating the initial similarity matrix for n documents needs O(n^2) operations. Depending on the linkage criterion, each merging operation, which is usually performed n−1 times, includes updating the matrix, which can also be difficult. Because of the repeated matrix updates, this technique has an O(n^3) overall time complexity.