HW4 – KMean Implementation

Yao-Chun Hsieh 462217691

In this assignment, I implemented K-Mean clustering using python without modern machine learning libraries.

Given a bunch of documents, first I do tokenization for each document. Second, I generate the TF matrix and the TF-iDF value for the terms in each document. The TF-iDF vector of each document is used for computing distance between documents during K-Mean clustering process. The detail of the K-Mean clustering implementation is explained with the following sub questions that affect the performance:

1. What is the value of K?

The value of K, which is 5, is a fixed number provided by professor Lee.

2. How to choose initial starting centers?

I randomly choose 5 documents as the initial documents. For each of them, I use its entire TF-iDF data as the cluster center.

3. How many times should we restart the clustering process?

In this assignment, I restart the clustering process 2 times of the amount of documents. For example, for directory with 10 documents, the restarting time is 10*2 = 20; for directory with 50 documents, the restarting time is 50*2 = 100. Moreover, each initial set is guaranteed to be unique.

The reason I set restarting time as 2 times of the amount of documents is due to the consideration of needed computation time. For a directory with 50 documents, it takes about 7 minutes to finish 100 convergences. The computation is time consuming, therefore I only use such small number for getting a local optimized result.

4. What is the definition of distance between two data?

The requirement asks student to represent each document using it's tf-idf vector, and calculate the distance between document with Euclidean Distance. The tf-idf formula is as below:

```
## For each term t in current document ##
# TF(t) = 0,                    if count(t) = 0,
# TF(t) = 1 + log10(count(t))   otherwise
# ------------------------------------------------------
# N = Total number of documnets
# DF(t) = the number of documents that t occurs
# iDF(t) = log10(N / DF(t))
# ------------------------------------------------------
# TF-iDF(t) = TF-weight * DF-weight = TF(t) * iDF(t)
```

5. **During clustering process, what is the criteria to terminate?**
   The clustering process is terminated when the cluster set remains stable. In other word, the combination of current cluster set is exactly the same as the previous one. For example, cluster set [[1,2], [3,4], [5]] equals to cluster set [[5], [4,3], [1,2]].

6. **What is the criteria for selecting best clusters?**
   RSS, is computed for each convergence, the formula is the sum of the distance of every document to the center of its belonging cluster. For all the acquired cluster sets, the one with minimum RSS is defined as the best cluster set.

The result of K-Mean clustering for 10 documents and 50 documents is in next page.

The result of K-Mean clustering for 10 documents and 50 documents is as below:

	10 Documents	50 Documents
Number of Terms	4455	17372
Restarting Amount	20	100
Total Execution Time	0:00:05.047524	0:06:58.922424
Avg Converge Time	0:00:00.02523	00:00:04.1892
Min RSS	61.58534	1381.0748720
Result Cluster Set ** Red Mark is the beginning of a cluster	[['turkey.html'], ['elizabeth_azcona_bocock.html', 'equator.html', 'hybrid_name.html', 'king.html', 'lorenzo_perosi.html', 'plagiosauridae.html'], ['ichthyosaur.html'], ['john_chrysostom.html'], ['turandot.html']]	[['alcoholism.html'], ['1828.html', '252.html', 'anal_glands.html', 'asterix.html', 'carrion.html', 'commissure.html', 'concordat_of_1801.html', 'dendrogram.html', 'eel.html', 'empanada.html', 'farnley.html', 'filter_(software).html', 'force.html', 'genome.html', 'hildegard.html', 'hooker_island.html', 'humerus.html', 'imprimatur.html', 'jason_miller.html', 'leeds_thomas_danby.html', 'lineage.html', 'list_of_church_of_england_dioceses.html', 'lyell_island.html', 'mikhail_lermontov.html', 'neuss.html', 'order.html', 'pin.html', 'pope_cornelius.html', 'scottish.html', 'skull.html', 'specific_name.html', 'the_rain_in_spain.html', 'variety_(universal_algebra).html'], ['ecology.html'], ['1969.html', 'argentina.html', 'bavaria.html', 'chevrolet.html', 'french_revolution.html', 'insect.html', 'july_10.html', 'june_21.html', 'lesotho.html', 'lichen.html', 'running.html', 'september_3.html', 'sociology.html', 'wisconsin.html'], ['recorder.html']]