CMSC 476/676 Information Retrieval
Programming Assignment 5 - Due NLT 11:59pm Monday May 6, 2019
extra credit if submitted on or before Wednesday May 1
Do this assignment individually

In this assignment we perform some analysis of the 504-document HTML test corpus, and introduce some concepts related to document clustering.

A similarity matrix is defined as a matrix in which entry *i,j* is the similarity of documents i and j, computed using the cosine score or some other metric. A document is perfectly similar to itself, so the entries on the main diagonal are all 1. Similarity is also symmetric, i.e. *sim(i,j) = sim(j,i)*, so the similarity matrix is upper triangular in form. In this assignment, you'll need to construct a similarity matrix. You should code this section of the homework on your own.

Certain clustering algorithms use a similarity matrix to do their work. In hierarchical agglomerative clustering, two objects are merged if their centroids are closer to each other than are the centroids of any other pair of objects. An object could be a single document, or an existing cluster. Note that a document not yet in any cluster is in fact a trivial singleton cluster, with itself as the centroid. You may assume that the intersection between any two clusters is the empty set.

Your program is to execute agglomerative clustering, using the group average link method, or single link method. As each step of the algorithm proceeds, indicate which objects are being merged (clustered) together, where again an object can be a single document, or an existing cluster. The clustering will cease when no two clusters (or documents) have similarity greater than 0.4. Clustering is discussed in the textbook, and more information is available online. You have the option of using a proceedure call from a package or, if you code this yourself, you will recieve extra credit.

Remember that we are using similarity rather than distance. A good resource can be found in Chapter 17 of Manning's book on Information Retrieval.

Since the output may get lengthy, we'll accept the first 100 lines of output, where the first line shows which two documents are most similar, and are therefore the first to be merged into a cluster.

In your report, you'll need to answer the following questions:
Which pair of HTML documents is the most similar?
Which pair of documents is the most dissimilar?
Which document is the closest to the corpus centroid?

You'll also need to describe your method for giving names to clusters, and explain how you implemented the similarity matrix and any other major data structures you used. Please include listings of your code. As usual, make a zip file and email it to our TA.