Assignment 1 (due October 1)

In this assignment you will be given a news group dataset. Perform the following tasks and report the results accompanied by the codes and created files.

1- Load the news group dataset into R.
2- Clean, preprocess and convert the dataset to feature vectors where features are term frequencies (TF) and the category of each document is its class.
3- Perform a hierarchal clustering model over the loaded dataset using k=2 to 6. Analyze and choose the best k based on the clustering results using the entropy to measure the quality of a clustering solution with different k. For the definition, see Wikipedia page on entropy (information theory).
4- Extend the feature to IDF (IDF instead of TF.IDF). See tf-idf Wikipedia page for definition.
5- Analyze and choose the best k based on the new features as in step 3.
6- The default distance in "hclust" is *Euclidean* distance. *Cosine* similarity is often used for document data. To use cosine similarity as the proximity measure in clustering, you can use the proxy R package. Perform the hierarchical clustering based on the best features and best k using cosine similarity as a distance function. Provide the analysis of how distance function affects the results. Does optimal k and features change with different distance functions?