CSE 5800 Mining/Learning and the Internet—HW3
Due Oct 21, Wed, 6:30pm
Submit Server: course= ml-internet , project=hw3

1. Implement these clustering algorithms, each outputs K clusters:
   (a) K-means
   (b) Bisecting K-means with largest cluster to split
   (c) Bisecting K-means with least overall similarity to split
   (d) Aggolermerative Hierarchical Clustering with Intra-Cluster Similarity technique (IST)
   (e) Aggolermerative Hierarchical Clustering with Centroid Similarity technique (CST)
   (f) Aggolermerative Hierarchical Clustering with UPGMA
   (g) Aggolermerative Hierarchical Clustering with UPGMA to seed K-means

2. Each document is represented by a TF-IDF unit vector, each component is: $tf_i \times idf_i$, where:
   - tf_i is the frequency of term i in the document divided by the total number of terms in the document and
   - $idf_i = \log(D/df_i)$, where df_i is the number of documents that contain term i and D is the total number of documents
   - to get a unit vector, divide each component by the magnitude of the vector

3. Allow these parameters:
   (a) number of (final) clusters (K in the paper)
   (b) number of iterations ($ITER$ in the paper) for Bisecting K-means

4. Measure performance of final clusters using:
   (a) Entropy
   (b) F-measure
   (c) Overall Similarity
   (d) Silhouette Coefficient

5. Three data sets:
   (a) toy data set on the course web site
   (b) news data set on the course web site
   (c) your own data set

6. A report (in pdf) that discusses the following:
   (a) Sensitivity analysis of parameters: for the second data set,
      i. for bisecting k-means, vary $ITER$ from 2 to 10 with increment of 2 (with K=6)
      ii. for each algorithm, vary K (keep $ITER$ constant for bisecting k-means based on the previous experiment) [assuming the desired number of clusters is not known]
      iii. calculate each performance measurement,
      iv. plot performance vs. value of a parameter,
      v. discuss the value for each parameter that seems to achieve the highest performance and possible reasons.
      vi. discuss if any of the performance measurements can help determine the value of K (which is usually not known in advance)
   (b) Compare the clustering algorithms: for the second data set,
      i. use the plot(s) for performance vs. number of clusters for different algorithms
      ii. discuss the relative performance of different algorithms

7. Implementation:
   (a) preferably use one of these programming languages: C, C++, Java, Python, or LISP.
   (b) input files: a file for the topic names; each topic has a file, which has multiple documents, each document starts with `--DocID--`
   (c) three modules:
      i. Preprocess: input the documents, output TF-IDF vectors
      ii. Cluster: input the TF-IDF vectors; for each cluster, output DocID's in the cluster and the top 3 words in the centroid
      iii. Evaluate: input DocID's, their class labels and cluster membership; output performance

8. Submission:
   (a) source code
   (b) your data set
   (c) report in pdf
   (d) README.txt (how to compile and run your program/experiments on code.fit.edu or hopper.cs.fit.edu)