R·I·T

Rochester Institute of Technology
Golisano College of Computing and Information Sciences
Department of Information Sciences and Technology

Lab 5 (2 points + 3 bonus points)
Document Clustering

Overview

This lab consists of three major tasks:

- Preprocess a document collection to construct a vector representation of documents
- Use a Kmeans clustering algorithm to cluster the documents
- (Bonus) Use a hierarchical clustering algorithm (choose one from single linkage or complete linkage) to cluster the documents

Resources

- You should have read Chapters 16 and 17 of Introduction to Information Retrieval.
- Go over the lecture notes of week 12.

Note: Make JavaDoc comments in your Java programs including Course #, Lab #, Your name, and main functional description of each method with @param & @return if applicable at the minimum.

Ref. http://www.oracle.com/technetwork/articles/java/index-137868.html

Submit your programs to a lab drop box in MyCourses by April 23, 2020.

Task 1: Preprocess documents to construct vector representations

In this task, you need to construct the vector representations for the documents to be clustered

1. Complete the following method and class in Cluster.java. Instead of using a tf-idf weighting mechanism, we only use the tf information here to simplify the task.
2.

```
/**
   * Load the documents to build the vector representations
   * @param docs
   */
        public void preprocess(String[] docs){
                //TO BE COMPLETED
        }


/**
```

```
 *
 * Document class for the vector representation of a document
 */
class Doc{
        //TO BE COMPLETED
}
```

Task 2: Cluster documents

In this task, you need to implement the following method that uses the Kmeans algorithm to cluster a set of documents.

1. Complete the cluster method in Cluster.java

```
  /**
   * Cluster the documents
   * For kmeans clustering, use the first and the ninth documents as the initial
centroids
   */
        public void cluster(){
                //TO BE COMPLETED

        }
```

[Bonus] Task 3: Cluster documents using a hierarchical clustering algorithm (choose either single linkage or complete linkage algorithms) (3 points)

In this task, you need to implement a hierarchical algorithm for document clustering. Hint: refer to the pseudo code SIMPLEHAC in section 17.1 of Introduction to Information Retrieval. Display the merging process as the final output.