R·I·T

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

Lab 5
Document Clustering

Overview

This lab consists of two major tasks:

- Preprocess a document collection to construct a VSM representation of documents
- Use a K-means clustering algorithm to cluster the documents

Resources

- You should have read Chapters 16 and 17 of Introduction to Information Retrieval.
- Go over the Text Clustering lecture content.

Task 1: Preprocess documents to construct VSM representations

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

1. Complete the following method and class in Clustering.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 K-means algorithm to cluster a set of documents.

1. Complete the cluster method in Clustering.java

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

	}
```

2. Output the cluster assignments for each document. This can be done in the cluster method. The expected cluster assignments are noted in Clustering.java in the main method.