R·I·T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **School of Information**

Lab 4
Text Classification using a Naïve Bayes Classifier

Overview

This lab consists of three major tasks:

- Building a Naïve Bayes classifier using a set of training documents
- Using the classifier to classify individual test documents and report the class label
- Using the classifier to classify a set of test documents and report the classification accuracy

Resources

- Read Chapter 13 of Introduction to Information Retrieval.
- Carefully read the Text Classification lecture examples and understand the technical details.
- The Lab4_data folder (stored in Lab4_data.zip on myCourses) consists of two subfolders: train and test.
    - The train folder consists of two subfolders: pos and neg, each of which consists of 900 positive and negative movie reviews, respectively.
    - The test folder consists of two subfolders: pos and neg, each of which consists of 100 positive and negative movie reviews, respectively.

Task 1 (40 points): Building the Naïve Bayes Classifier

In this task, you need to construct the Naïve Bayes classifier:

1. Complete the following two methods in NBClassifier.java, using which you can construct a Naïve Bayes classifier.

```
 * Build a Naive Bayes classifier using a training document set
 * @param trainDataFolder the training document folder
 */
public NBClassifier(String trainDataFolder)
{

}

 * Load the training documents
 * @param trainDataFolder
 */
public void preprocess(String trainDataFolder)
{

}
```

Task 2 (20 points): Classifying individual testing documents

In this task, you need to implement the following method that uses the Naïve Bayes classifier to assign the class label to a given testing document.

1. Complete the classify method in NBClassifier.java

```
/**
 * Classify a test doc
 * @param doc test doc
 * @return class label
 */
public int classify(String doc) {

}
```

Task 3 (40 points): Classify all the test documents in the test folder and report the classification accuracy

In this task, you need to implement the following method that uses the Naïve Bayes classifier to assign the class labels to all the testing documents in the test folder, compare the assigned label with the true class label, and report the overall classification accuracy.

```
/**
 *  Classify a set of testing documents and report the accuracy
 * @param testDataFolder fold that contains the testing documents
 * @return classification accuracy
 */
public double classifyAll(String testDataFolder)
{

}
```

Here is a screenshot of the expected classification result: