R·I·T

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

Lab 4 (3+2 points)
Text Classification using a Naïve Bayes (Multinomial) 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
- (Bonus) Using the classifier to classify a set of test documents and report the classification accuracy

Resources

- You should have read Chapter 13 of Introduction to Information Retrieval.
- Carefully read the lecture examples of week 10 and understand the technical details.
- Go over the lecture notes of week 10.
- The data folder 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.

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 16, 2020.

Task 1: Building the Naïve Bayes Classifier (2 points)

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

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 trainDataFoler the training document folder
          */
         public NBClassifier(String trainDataFolder)
         {

         }




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

         }
```

Task 2: Classifying individual testing documents (1 points)

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 (Bonus): Classifying all the testing documents in the test folder and report the classification accuracy (2 points)

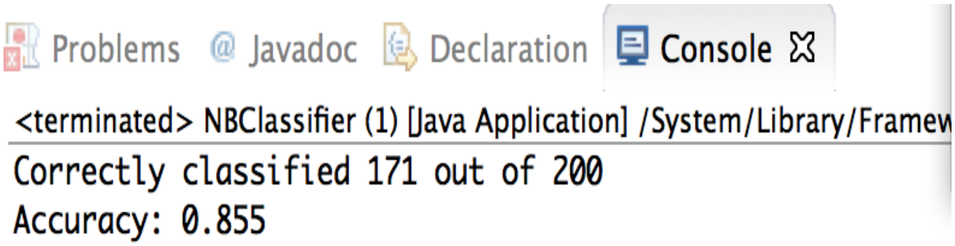
In this task, you need to implement the following method that uses the Naïve Bayes classifier to assign the class labels to all the testing documents in the test folder, compare the assigned label with the true class label, and report the overall classification accuracy.

```
/**
 *  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 the screenshot of the result:

Submit your programs to a lab drop box in MyCourses before meeting with Instructor/TA.



```
<terminated> NBClassifier (1) [Java Application] /System/Library/Framew
Correctly classified 171 out of 200
Accuracy: 0.855
```