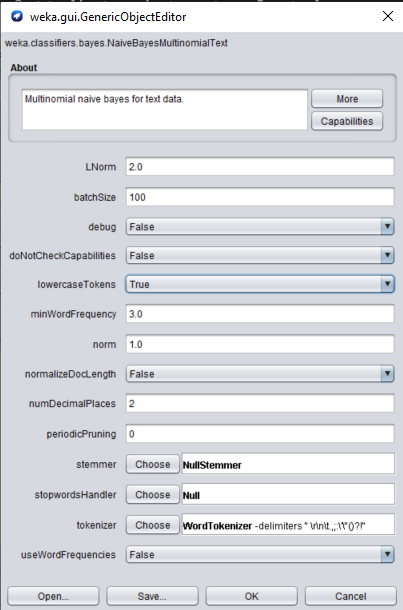
# Task 1: Split the Data set into Training and Test Sets

Test Instances:

|  |  |  |
| --- | --- | --- |
| Domain | Line Number | Index Number |
| Computer Science | 16 | 5 |
| Computer Science | 26 | 15 |
| Breast Cancer | 31 | 20 |
| Breast Cancer | 35 | 24 |
| Aircraft | 51 | 40 |
| Aircraft | 64 | 53 |

This table represents the original line numbers and their indices in the original dataset before extracting them out into a separate test dataset. 2 instances were selected randomly from each domain (comp sci, breast cancer, aircraft).

# Task 2: Naïve Bayes Classification using Weka

  
*Settings used for the classifier in Weka*

**Classifier Output:**

**=== Summary ===**

Correctly Classified Instances 5 83.3333 %

Incorrectly Classified Instances 1 16.6667 %

Kappa statistic 0.75

Mean absolute error 0.1267

Root mean squared error 0.3009

Relative absolute error 28.5159 %

Root relative squared error 63.8249 %

Total Number of Instances 6

**=== Detailed Accuracy By Class ===**

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Computer-Science

1.000 0.250 0.667 1.000 0.800 0.707 1.000 1.000 Breast-Cancer

0.500 0.000 1.000 0.500 0.667 0.632 1.000 1.000 Aircraft

*Weighted Avg. 0.833 0.083 0.889 0.833 0.822 0.780 1.000 1.000*

**=== Confusion Matrix ===**

a b c <-- classified as

2 0 0 | a = Computer-Science

0 2 0 | b = Breast-Cancer

0 1 1 | c = Aircraft

**Questions:**

What the is the dictionary size of the classifier model?

749

What is the total number of instances in the training set?

54

How many instances in each domain in the training set?

18

How many instances in the test set?

6

How many test instances were incorrectly classified?

1

What is the confusion matrix of the test results?

=== Confusion Matrix ===  
a b c <-- classified as  
2 0 0 | a = Computer-Science  
0 2 0 | b = Breast-Cancer  
0 1 1 | c = Aircraft

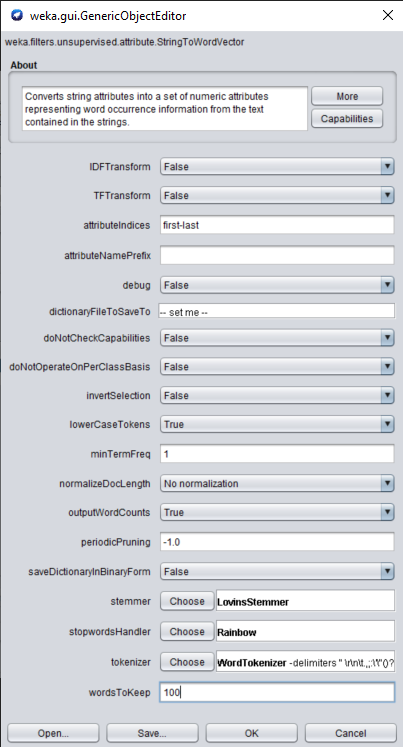
For each domain, list the recall, precision, and F1-Measure.

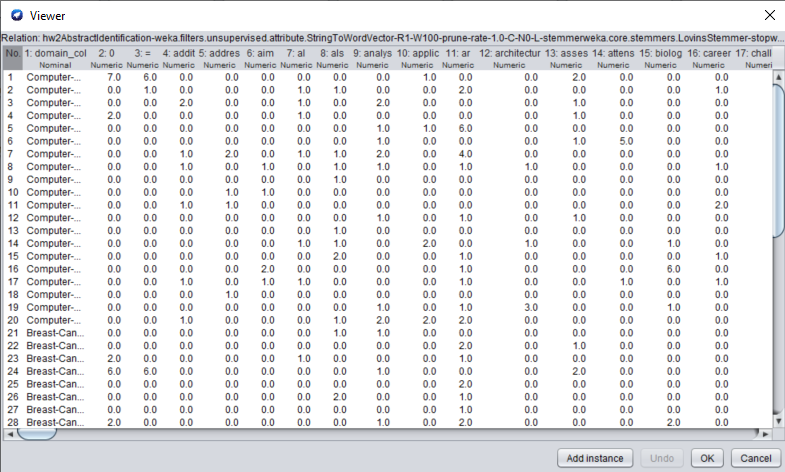
|  |  |  |  |
| --- | --- | --- | --- |
| Domain | Recall | Precision | F-Measure |
| Computer-Science | 1.000 | 1.000 | 1.000 |
| Breast-Cancer | 1.000 | 0.667 | 0.800 |
| Aircraft | 0.500 | 1.000 | 0.833 |

What is the average recall, precision, and F1-Measure of the Naïve Bayes Classifier on the datasets?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Recall | Precision | F-Measure |
| Weighted Avg. | 0.833 | 0.889 | 0.822 |

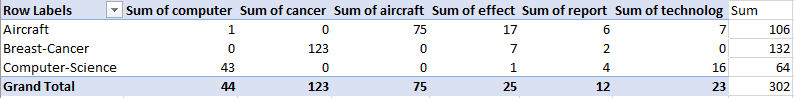
# Task 3: Naïve Bayes Classification by Manually Computing Conditional Probabilities

  
*Settings used to filter in Weka*

  
*Lots of the words ended up becoming abbreviated after running the filter. For example, “technology” became “technolog”.*

**Term Frequency of selected words:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Frequency of the Selected 6 Dictionary Words | | | | | |  |
| **Domain** | computer | cancer | aircraft | effect | report | “technolog” | **Grand Total** |
| **Computer Science** | 43 | 0 | 0 | 1 | 4 | 16 | 64 |
| **Breast Cancer** | 0 | 123 | 0 | 7 | 2 | 0 | 132 |
| **Aircraft** | 1 | 0 | 75 | 17 | 6 | 7 | 106 |

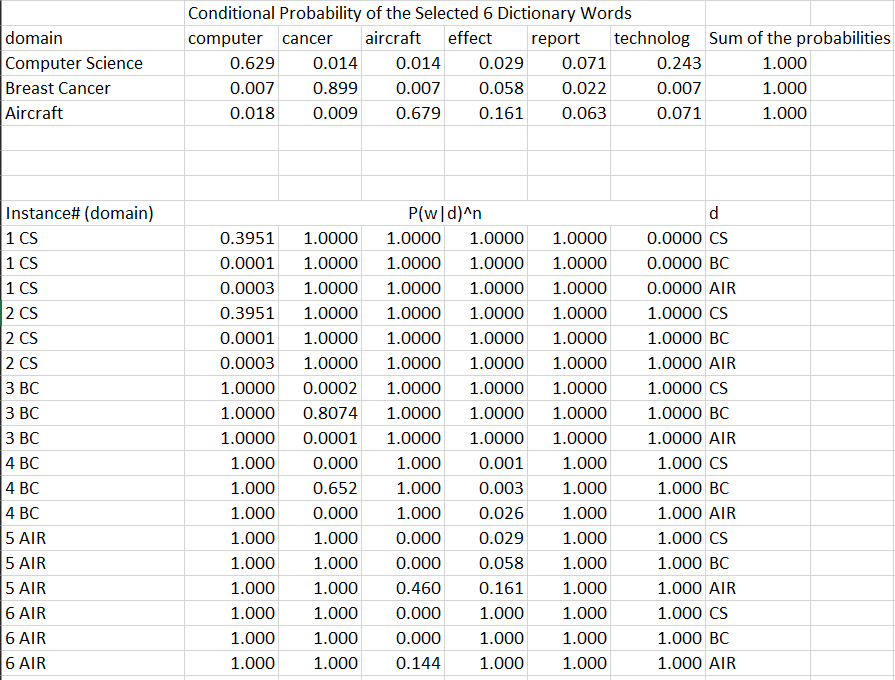
  
*Output from Excel’s Pivot Table*

**Conditional Probability Chart:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Conditional Probability of the Selected 6 Dictionary Words | | | | | |  |
| domain | computer | cancer | aircraft | effect | report | technolog | Sum of the probabilities |
| Aircraft | 0.018 | 0.009 | 0.679 | 0.161 | 0.063 | 0.071 | 1.000 |
| Breast Cancer | 0.007 | 0.899 | 0.007 | 0.058 | 0.022 | 0.007 | 1.000 |
| Computer Science | 0.629 | 0.014 | 0.014 | 0.029 | 0.071 | 0.243 | 1.000 |

**Predicting Testing Data Manually**

I started off by copying each row from the original data set and putting it in its own Excel sheet to make filtering easier. Using the index numbers from part 1, I was able to identify which rows in the term frequency output corresponded with the testing data. I created a new sheet in Excel and copied the header and the 6 instances over and then used the hide function in Excel to hide the irrelevant columns. I then took the conditional probabilities table, and the new term frequency table and created an auxiliary table that calculated P(w|d)^n for each word in each instance (see the auxiliary table below). I then used this auxiliary table to compute the product of each row to calculate the un-normalized conditional probability for each instance, given a different domain. Then taking the higher of the 3 values to predict what domain each instance belongs. To my surprise, running this test manually produced perfect results. Instance 1 came the closest to failing, as the values were very small. However, in the end, the value for Computer Science still ended up being the largest (in the table below it is listed as 0.0000 as I set the display settings to only allow 4 decimal points for cleaner screenshots.)

  
*Auxiliary table calculating P(w|d)^n in Excel*

**Final Table – Manual Classification of Abstracts using Conditional Probability:**

  
*Results table*

**Questions:**

Create a Confusion Matrix for your output:

a b c <-- classified as  
2 0 0 | a = Computer-Science  
0 2 0 | b = Breast-Cancer  
0 0 2 | c = Aircraft

Create a Performance Evaluation for your output:

|  |  |  |  |
| --- | --- | --- | --- |
| Domain | Precision | Recall | F1 Score |
| Computer Science | 1.000 | 1.000 | 1.000 |
| Breast Cancer | 1.000 | 1.000 | 1.000 |
| Aircraft | 1.000 | 1.000 | 1.000 |