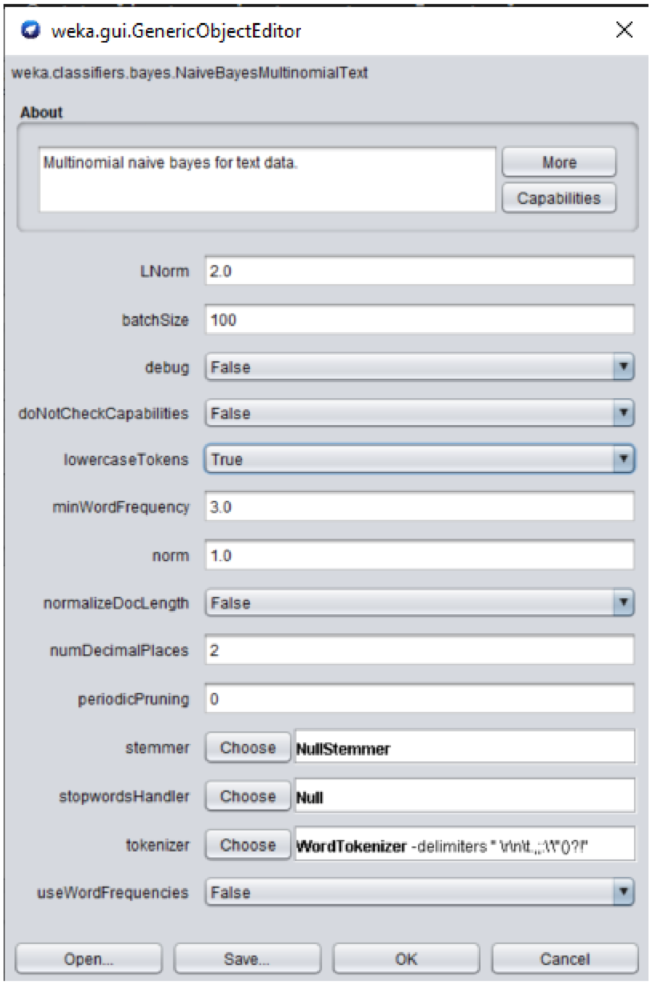
Task 1: Split the Data set into Training and Test Sets

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



weka.gui.GenericObjectEditor

weka.classifiers.bayes.NaiveBayesMultinomialText

About

Multinomial naive bayes for text data.

LNorm	2.0
batchSize	100
debug	False
doNotCheckCapabilities	False
lowercaseTokens	True
minWordFrequency	3.0
norm	1.0
normalizeDocLength	False
numDecimalPlaces	2
periodicPruning	0
stemmer	NullStemmer
stopwordsHandler	Null
tokenizer	WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"
useWordFrequencies	False

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
```

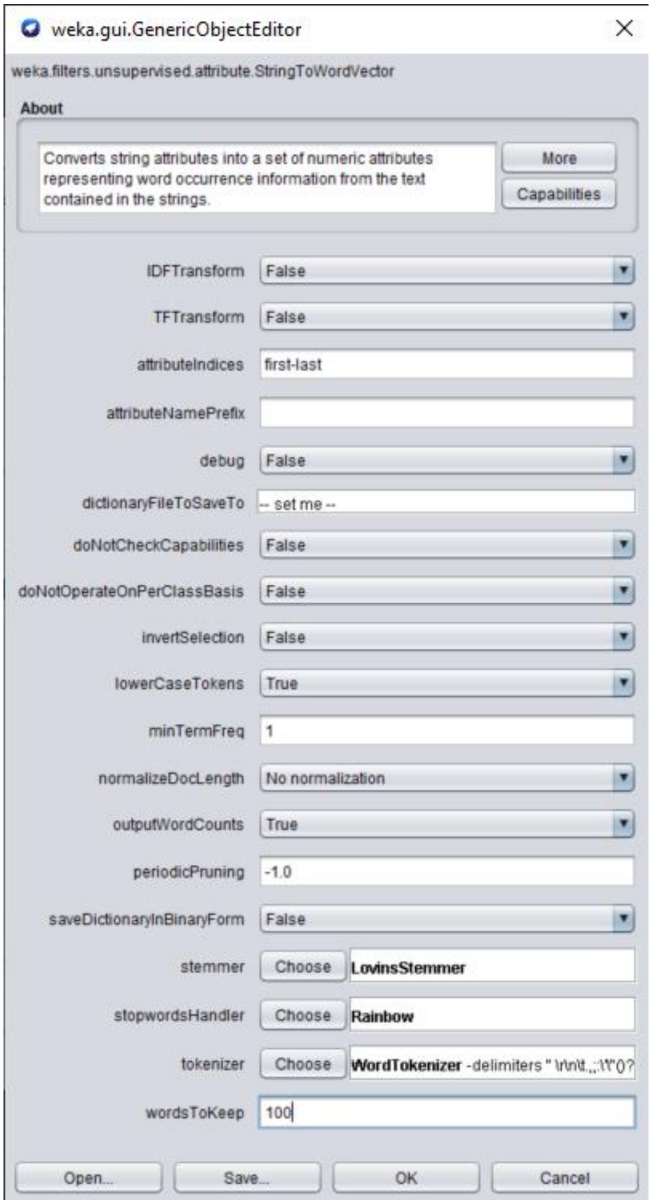
For each domain, list the recall, precision, and F1-Measure.

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



weka.gui.GenericObjectEditor

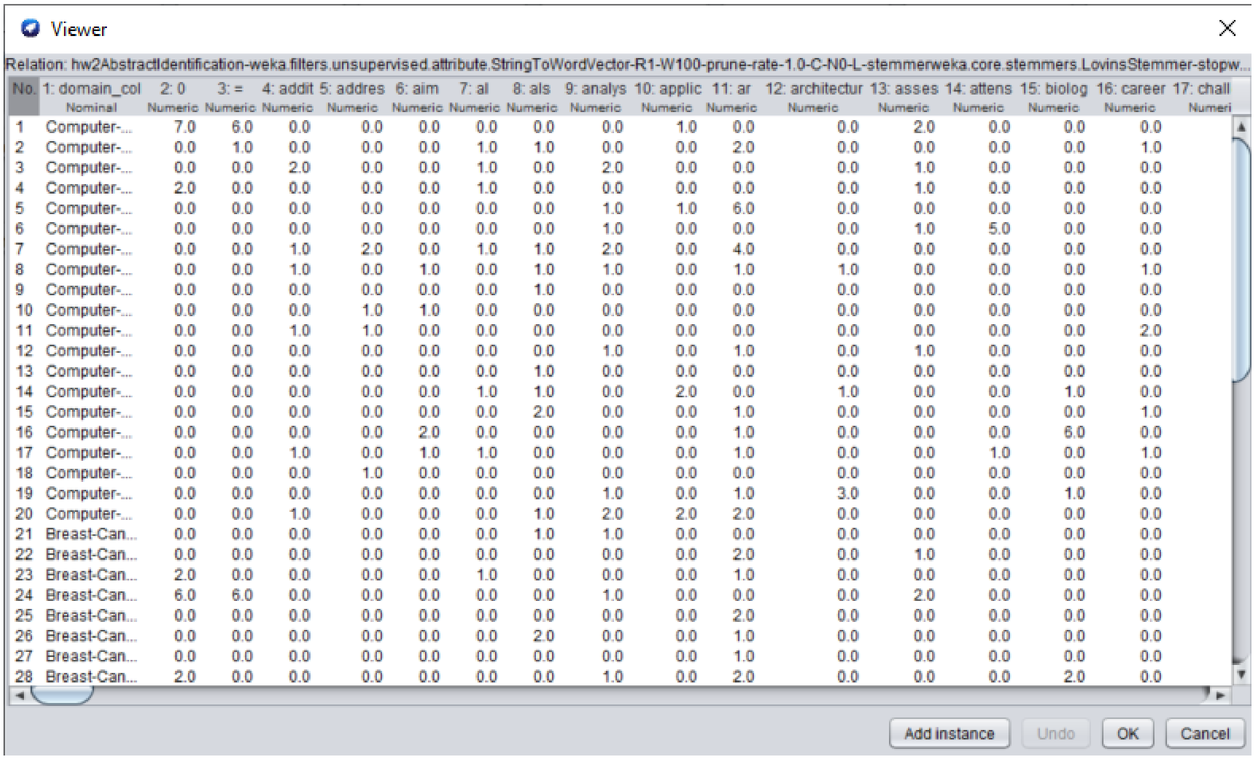
weka.filters.unsupervised.attribute.StringToWordVector

About

Converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings.

IDFTransform	False
TFTransform	False
attributeIndices	first-last
attributeNamePrefix	
debug	False
dictionaryFileToSaveTo	-- set me --
doNotCheckCapabilities	False
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	True
minTermFreq	1
normalizeDocLength	No normalization
outputWordCounts	True
periodicPruning	-1.0
saveDictionaryInBinaryForm	False
stemmer	LovinsStemmer
stopwordsHandler	Rainbow
tokenizer	WordTokenizer -delimiters " \r\n\t.,;:\'\"()?
wordsToKeep	100

Settings used to filter in Weka



Relation: hw2AbstractIdentification-weka.filters.unsupervised.attribute.StringToWordVector-R1-W100-prune-rate-1.0-C-N0-L-stemmerweka.core.stemmers.LovinsStemmer-stopw...

No.	1: domain_col	2: 0	3: =	4: addit	5: addres	6: aim	7: al	8: als	9: analys	10: applic	11: ar	12: architectur	13: asses	14: attens	15: biolog	16: career	17: chall
	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeri
1	Computer-...	7.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	
2	Computer-...	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	
3	Computer-...	0.0	0.0	2.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
4	Computer-...	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
5	Computer-...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	6.0	0.0	0.0	0.0	0.0	0.0	
6	Computer-...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	5.0	0.0	0.0	
7	Computer-...	0.0	0.0	1.0	2.0	0.0	1.0	1.0	2.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	
8	Computer-...	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	
9	Computer-...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	Computer-...	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
11	Computer-...	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	
12	Computer-...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	
13	Computer-...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14	Computer-...	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	2.0	0.0	1.0	0.0	0.0	1.0	0.0	
15	Computer-...	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	
16	Computer-...	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	6.0	0.0	
17	Computer-...	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	
18	Computer-...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
19	Computer-...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	3.0	0.0	0.0	1.0	0.0	
20	Computer-...	0.0	0.0	1.0	0.0	0.0	0.0	1.0	2.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	
21	Breast-Can...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
22	Breast-Can...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	0.0	
23	Breast-Can...	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
24	Breast-Can...	6.0	6.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	
25	Breast-Can...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	
26	Breast-Can...	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
27	Breast-Can...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
28	Breast-Can...	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	2.0	0.0	

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

Row Labels	Sum of computer	Sum of cancer	Sum of aircraft	Sum of effect	Sum of report	Sum of technolog	Sum
Aircraft	1	0	75	17	6	7	106
Breast-Cancer	0	123	0	7	2	0	132
Computer-Science	43	0	0	1	4	16	64
Grand Total	**44**	**123**	**75**	**25**	**12**	**23**	302

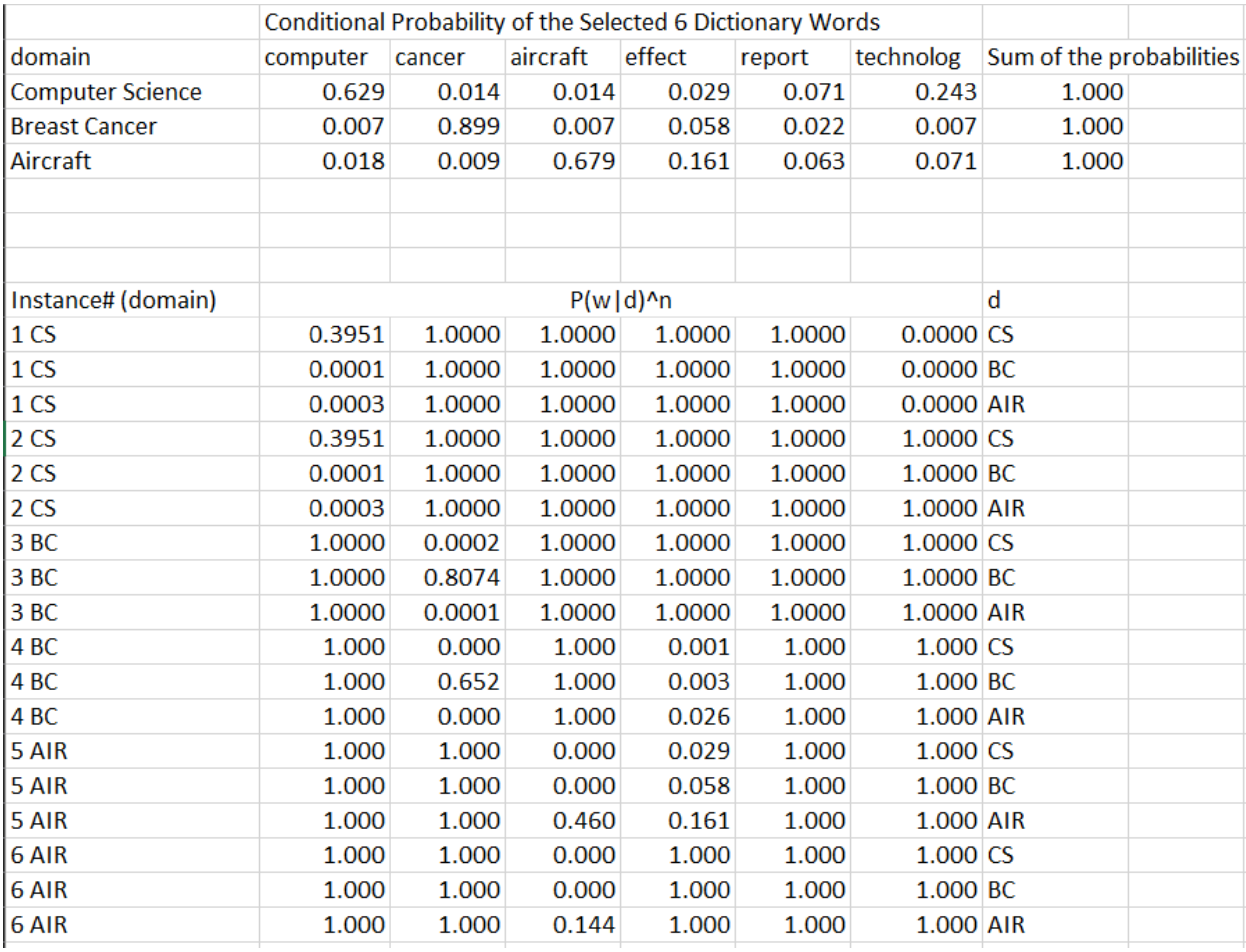
Output from Excel's Pivot Table

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

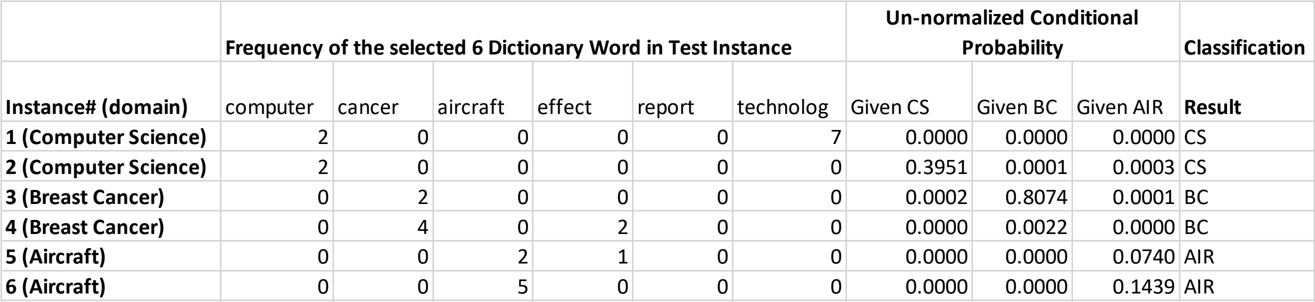
I started off by copying each row from the original data set and putting it in its own Excel sheet to make filtering easier. Using the index numbers from part 1, I was able to identify which rows in the term frequency output corresponded with the testing data. I created a new sheet in Excel and copied 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.)



	Conditional Probability of the Selected 6 Dictionary Words							
domain	computer	cancer	aircraft	effect	report	technolog	Sum of the probabilities	
Computer Science	0.629	0.014	0.014	0.029	0.071	0.243	1.000	
Breast Cancer	0.007	0.899	0.007	0.058	0.022	0.007	1.000	
Aircraft	0.018	0.009	0.679	0.161	0.063	0.071	1.000	
Instance# (domain)	P(w|d)^n						d	
1 CS	0.3951	1.0000	1.0000	1.0000	1.0000	0.0000	CS	
1 CS	0.0001	1.0000	1.0000	1.0000	1.0000	0.0000	BC	
1 CS	0.0003	1.0000	1.0000	1.0000	1.0000	0.0000	AIR	
2 CS	0.3951	1.0000	1.0000	1.0000	1.0000	1.0000	CS	
2 CS	0.0001	1.0000	1.0000	1.0000	1.0000	1.0000	BC	
2 CS	0.0003	1.0000	1.0000	1.0000	1.0000	1.0000	AIR	
3 BC	1.0000	0.0002	1.0000	1.0000	1.0000	1.0000	CS	
3 BC	1.0000	0.8074	1.0000	1.0000	1.0000	1.0000	BC	
3 BC	1.0000	0.0001	1.0000	1.0000	1.0000	1.0000	AIR	
4 BC	1.000	0.000	1.000	0.001	1.000	1.000	CS	
4 BC	1.000	0.652	1.000	0.003	1.000	1.000	BC	
4 BC	1.000	0.000	1.000	0.026	1.000	1.000	AIR	
5 AIR	1.000	1.000	0.000	0.029	1.000	1.000	CS	
5 AIR	1.000	1.000	0.000	0.058	1.000	1.000	BC	
5 AIR	1.000	1.000	0.460	0.161	1.000	1.000	AIR	
6 AIR	1.000	1.000	0.000	1.000	1.000	1.000	CS	
6 AIR	1.000	1.000	0.000	1.000	1.000	1.000	BC	
6 AIR	1.000	1.000	0.144	1.000	1.000	1.000	AIR	

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

Final Table – Manual Classification of Abstracts using Conditional Probability:



	Frequency of the selected 6 Dictionary Word in Test Instance						**Un-normalized Conditional Probability**			**Classification**
Instance# (domain)	computer	cancer	aircraft	effect	report	technolog	Given CS	Given BC	Given AIR	**Result**
1 (Computer Science)	2	0	0	0	0	7	0.0000	0.0000	0.0000	CS
2 (Computer Science)	2	0	0	0	0	0	0.3951	0.0001	0.0003	CS
3 (Breast Cancer)	0	2	0	0	0	0	0.0002	0.8074	0.0001	BC
4 (Breast Cancer)	0	4	0	2	0	0	0.0000	0.0022	0.0000	BC
5 (Aircraft)	0	0	2	1	0	0	0.0000	0.0000	0.0740	AIR
6 (Aircraft)	0	0	5	0	0	0	0.0000	0.0000	0.1439	AIR

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