Sanjivani Rural Education Society's
Sanjivani College of Engineering, Kopargaon-423 603
(An Autonomous Institute, Affiliated to Savitribai Phule Pune University, Pune)
NACC 'A' Grade Accredited, ISO 9001:2015 Certified

Department of Computer Engineering
(NBA Accredited)

Subject- Data Mining and Warehousing Lab (CO319)
Assignment No.5

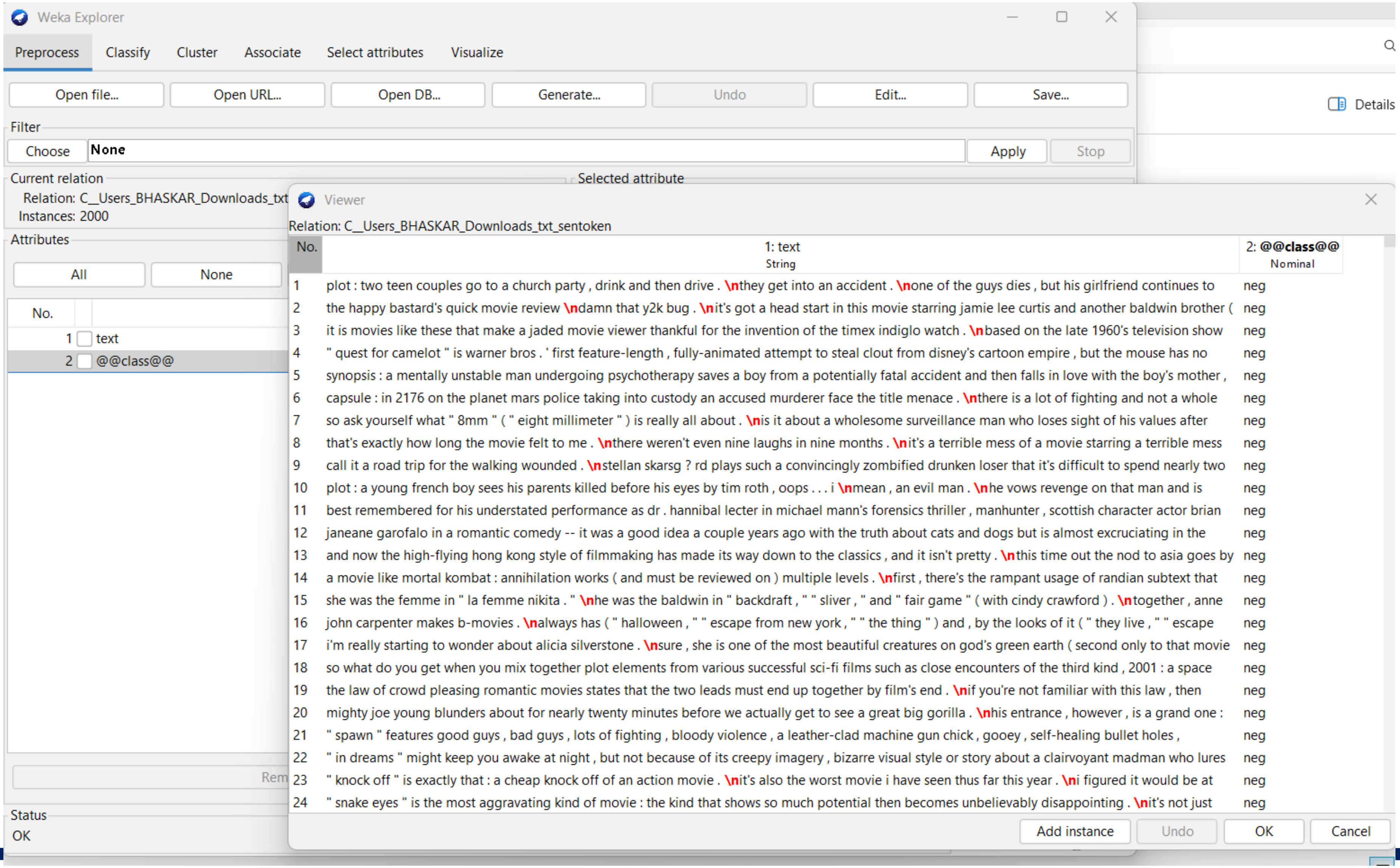
Prof. S.A.Shivarkar
Assistant Professor
Contact No.8275032712
Email- shivarkarsandipcomp@sanjivani.org.in

Problem Statement

- Consider a suitable text dataset. Remove stop words, apply stemming and feature selection techniques to represent documents as vectors. Classify documents and evaluate precision, recall (For Ex. Movie Review Dataset)

View the data

Select Class→ Click on Edit→ View the data



None

Relation: C__Users_BHASKAR_Downloads_txt
Instances: 2000

No.	
1	text
2	@@class@@

Status
OK

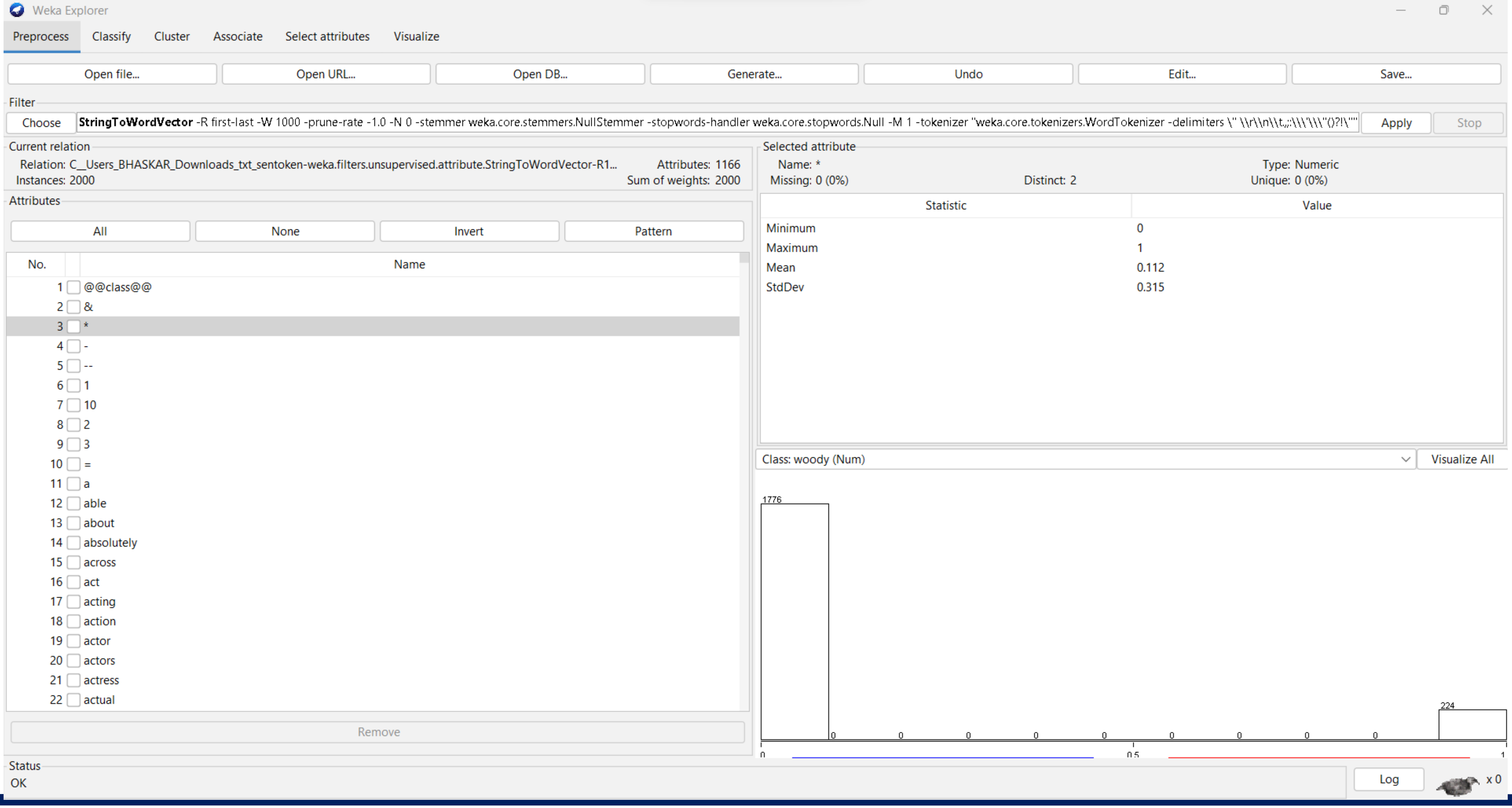
Relation: C__Users_BHASKAR_Downloads_txt_sentoken

No.	1: text String	2: @@class@@ Nominal
1	plot : two teen couples go to a church party , drink and then drive . \nthey get into an accident . \none of the guys dies , but his girlfriend continues to	neg
2	the happy bastard's quick movie review \ndamn that y2k bug . \nit's got a head start in this movie starring jamie lee curtis and another baldwin brother (	neg
3	it is movies like these that make a jaded movie viewer thankful for the invention of the timex indiglo watch . \nbased on the late 1960's television show	neg
4	" quest for camelot " is warner bros . ' first feature-length , fully-animated attempt to steal clout from disney's cartoon empire , but the mouse has no	neg
5	synopsis : a mentally unstable man undergoing psychotherapy saves a boy from a potentially fatal accident and then falls in love with the boy's mother ,	neg
6	capsule : in 2176 on the planet mars police taking into custody an accused murderer face the title menace . \nthere is a lot of fighting and not a whole	neg
7	so ask yourself what " 8mm " (" eight millimeter ") is really all about . \nis it about a wholesome surveillance man who loses sight of his values after	neg
8	that's exactly how long the movie felt to me . \nthere weren't even nine laughs in nine months . \nit's a terrible mess of a movie starring a terrible mess	neg
9	call it a road trip for the walking wounded . \nstellan skarsg ? rd plays such a convincingly zombified drunken loser that it's difficult to spend nearly two	neg
10	plot : a young french boy sees his parents killed before his eyes by tim roth , oops . . . i \nmean , an evil man . \nhe vows revenge on that man and is	neg
11	best remembered for his understated performance as dr . hannibal lecter in michael mann's forensics thriller , manhunter , scottish character actor brian	neg
12	janeane garofalo in a romantic comedy -- it was a good idea a couple years ago with the truth about cats and dogs but is almost excruciating in the	neg
13	and now the high-flying hong kong style of filmmaking has made its way down to the classics , and it isn't pretty . \nthis time out the nod to asia goes by	neg
14	a movie like mortal kombat : annihilation works (and must be reviewed on) multiple levels . \nfirst , there's the rampant usage of randian subtext that	neg
15	she was the femme in " la femme nikita . " \nhe was the baldwin in " backdraft , " " sliver , " and " fair game " (with cindy crawford) . \ntogether , anne	neg
16	john carpenter makes b-movies . \nalways has (" halloween , " " escape from new york , " " the thing ") and , by the looks of it (" they live , " " escape	neg
17	i'm really starting to wonder about alicia silverstone . \nsure , she is one of the most beautiful creatures on god's green earth (second only to that movie	neg
18	so what do you get when you mix together plot elements from various successful sci-fi films such as close encounters of the third kind , 2001 : a space	neg
19	the law of crowd pleasing romantic movies states that the two leads must end up together by film's end . \nif you're not familiar with this law , then	neg
20	mighty joe young blunders about for nearly twenty minutes before we actually get to see a great big gorilla . \nhis entrance , however , is a grand one :	neg
21	" spawn " features good guys , bad guys , lots of fighting , bloody violence , a leather-clad machine gun chick , gooey , self-healing bullet holes ,	neg
22	" in dreams " might keep you awake at night , but not because of its creepy imagery , bizarre visual style or story about a clairvoyant madman who lures	neg
23	" knock off " is exactly that : a cheap knock off of an action movie . \nit's also the worst movie i have seen thus far this year . \ni figured it would be at	neg
24	" snake eyes " is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing . \nit's not just	neg

Converting string to Numeric attribute

- Select Class→ Choose Filter→ Filter→ Unsupervised→ Attribute→ StrinToWordVector→Apply
- This will convert each word in string field to numeric attribute.
- Name of the attribute is word itself.
- The value of each attribute is 0 (absent) or 1 (present) in the current document.

Converting string to Numeric attribute



Filter: StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M 1 -tokenizer "weka.core.tokenizers.WordTokenizer -delimiters \" \\r\\n\\t.,;:\\\'\\\"()?!\""

Current relation

Relation: C__Users_BHASKAR_Downloads_txt_sentoken-weka.filters.unsupervised.attribute.StringToWordVector-R1...

Attributes: 1166

Instances: 2000

Sum of weights: 2000

No.	Name
1	@@class@@
2	&
3	*
4	-
5	--
6	1
7	10
8	2
9	3
10	=
11	a
12	able
13	about
14	absolutely
15	across
16	act
17	acting
18	action
19	actor
20	actors
21	actress
22	actual

Selected attribute

Name: *

Type: Numeric

Missing: 0 (0%)

Distinct: 2

Unique: 0 (0%)

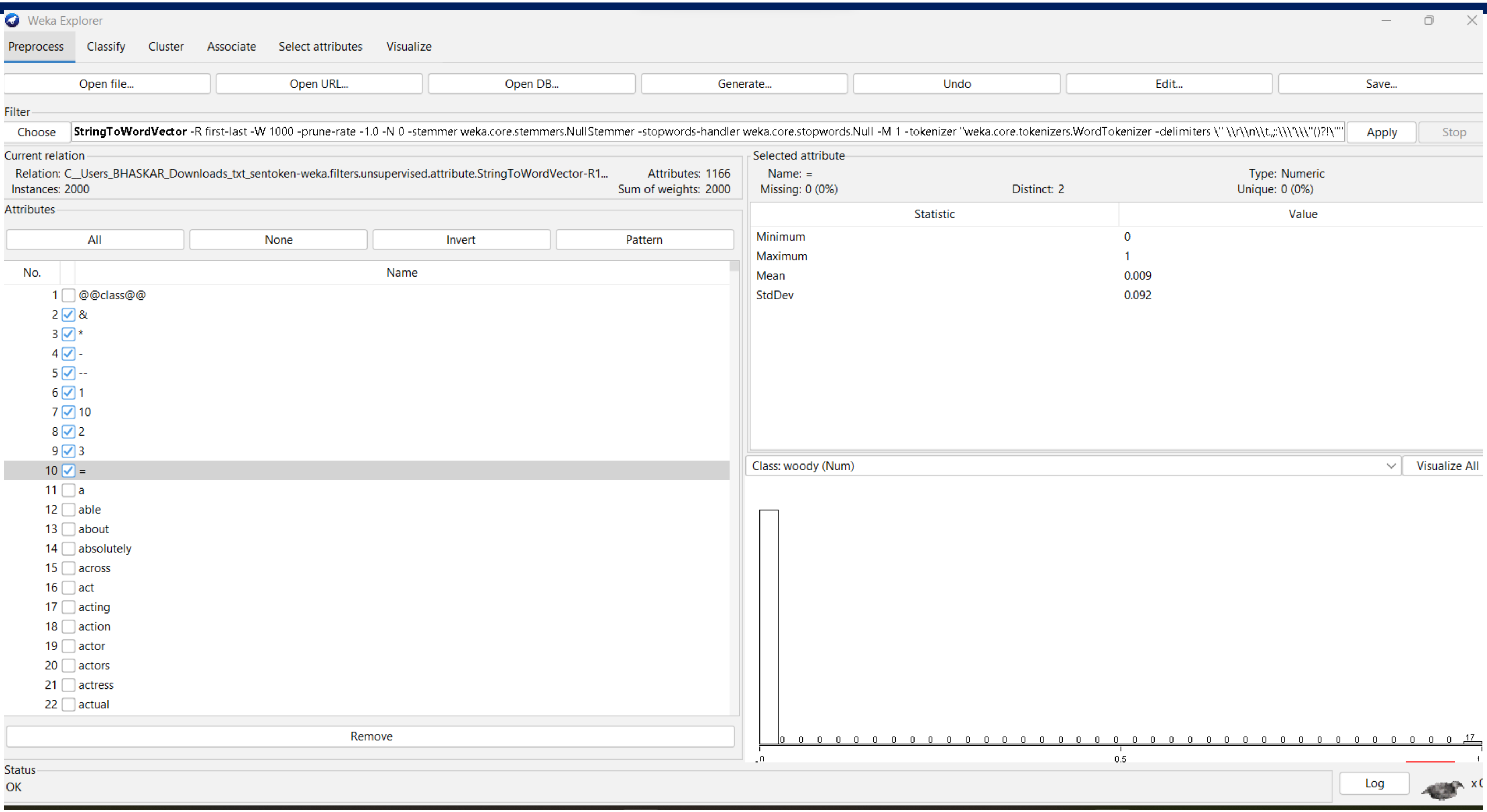
Statistic	Value
Minimum	0
Maximum	1
Mean	0.112
StdDev	0.315

Class: woody (Num)

Status

OK

Remove the selected attribute



Filter

StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M 1 -tokenizer "weka.core.tokenizers.WordTokenizer -delimiters \" \\r\\n\\t.,;:\\\'\\\"()?!\""

Current relation

Relation: C__Users_BHASKAR_Downloads_txt_sentoken-weka.filters.unsupervised.attribute.StringToWordVector-R1... Attributes: 1166
Instances: 2000 Sum of weights: 2000

Attributes

No.		Name
1	□	@@class@@
2	☑	&
3	☑	*
4	☑	-
5	☑	--
6	☑	1
7	☑	10
8	☑	2
9	☑	3
10	☑	=
11	□	a
12	□	able
13	□	about
14	□	absolutely
15	□	across
16	□	act
17	□	acting
18	□	action
19	□	actor
20	□	actors
21	□	actress
22	□	actual

Selected attribute

Name: = Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.009
StdDev	0.092

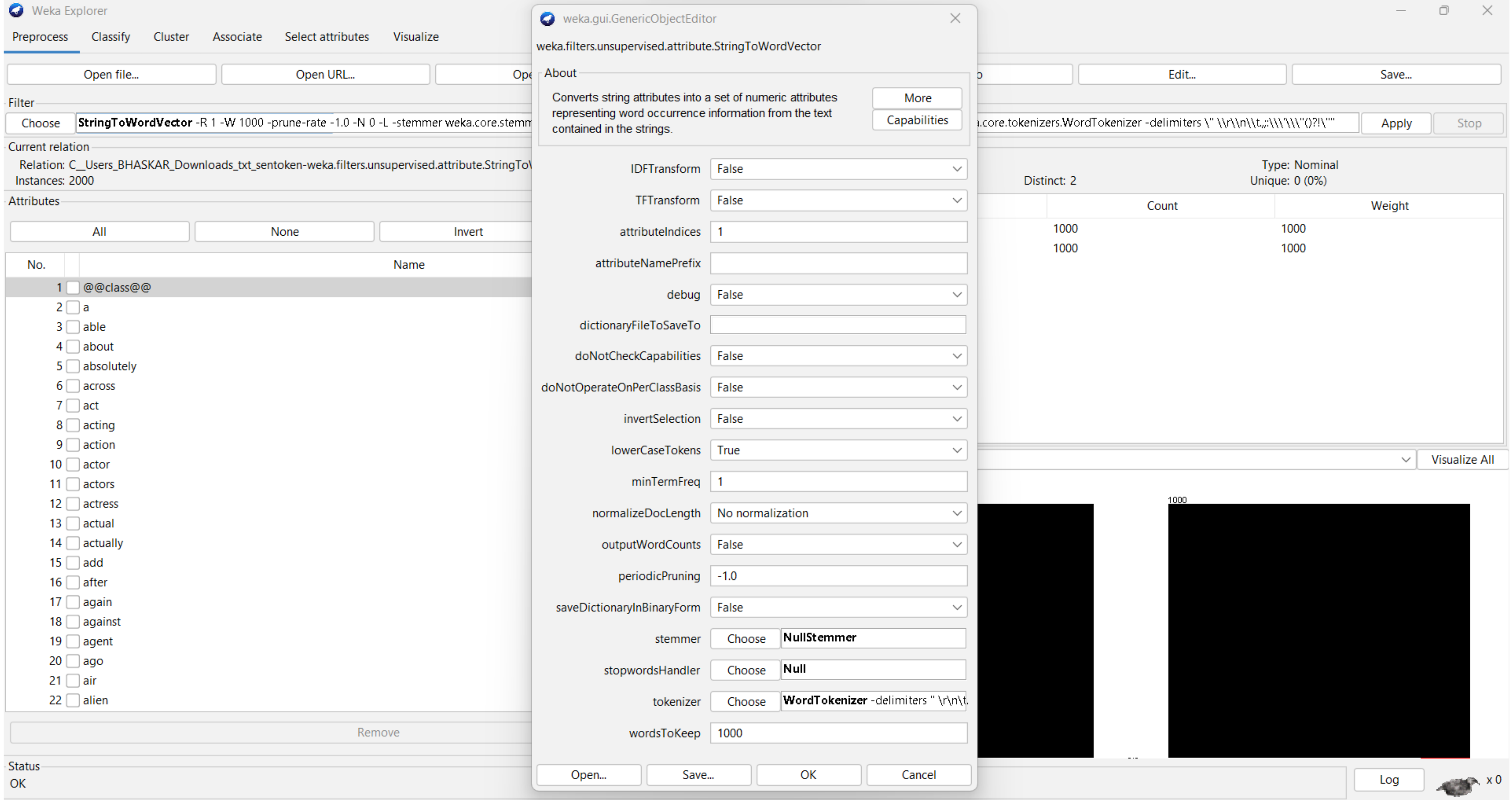
Class: woody (Num)

Status
OK

Converting string to Numeric attribute

- Click on StringToWordVector in filter.
- Set the parameters attributeIndices=1 and lowerCaseTokens=True, rest keep default.
- Click on Ok

Converting string to Numeric attribute



Filter: StringToWordVector -R 1 -W 1000 -prune-rate -1.0 -N 0 -L -stemmer weka.core.stemm … .core.tokenizers.WordTokenizer -delimiters \" \\r\\n\\t.,;:\\\'\\\"()?!\""

Current relation

Relation: C__Users_BHASKAR_Downloads_txt_sentoken-weka.filters.unsupervised.attribute.StringTo\

Instances: 2000

Distinct: 2

Type: Nominal

Unique: 0 (0%)

Count	Weight
1000	1000
1000	1000

Attributes

No.	Name
1	@@class@@
2	a
3	able
4	about
5	absolutely
6	across
7	act
8	acting
9	action
10	actor
11	actors
12	actress
13	actual
14	actually
15	add
16	after
17	again
18	against
19	agent
20	ago
21	air
22	alien

Status

OK

weka.filters.unsupervised.attribute.StringToWordVector

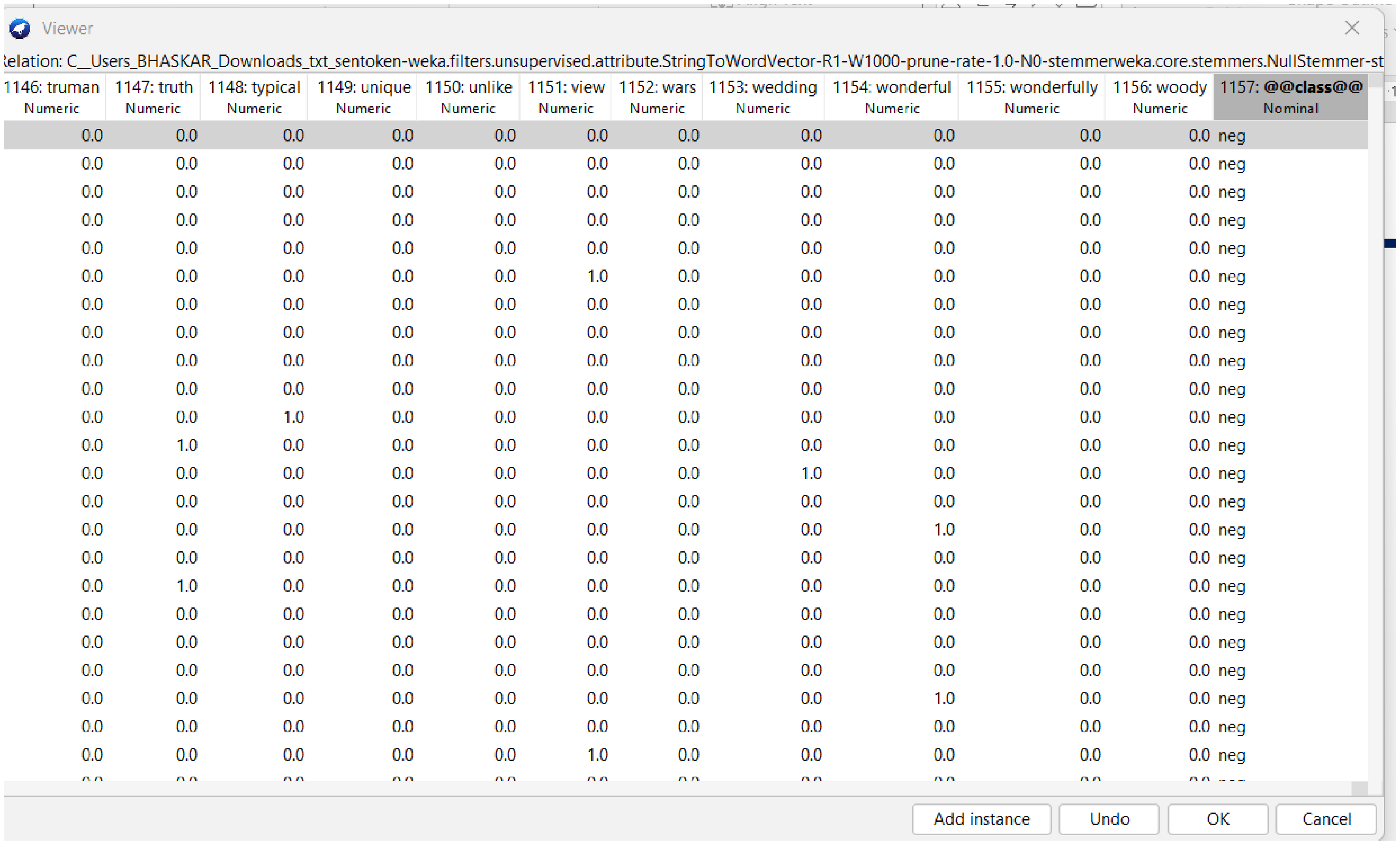
About

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	1
attributeNamePrefix	
debug	False
dictionaryFileToSaveTo	
doNotCheckCapabilities	False
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	True
minTermFreq	1
normalizeDocLength	No normalization
outputWordCounts	False
periodicPruning	-1.0
saveDictionaryInBinaryForm	False
stemmer	Choose NullStemmer
stopwordsHandler	Choose Null
tokenizer	Choose WordTokenizer -delimiters " \r\n\t.
wordsToKeep	1000

Move class attribute to the end of WordVector

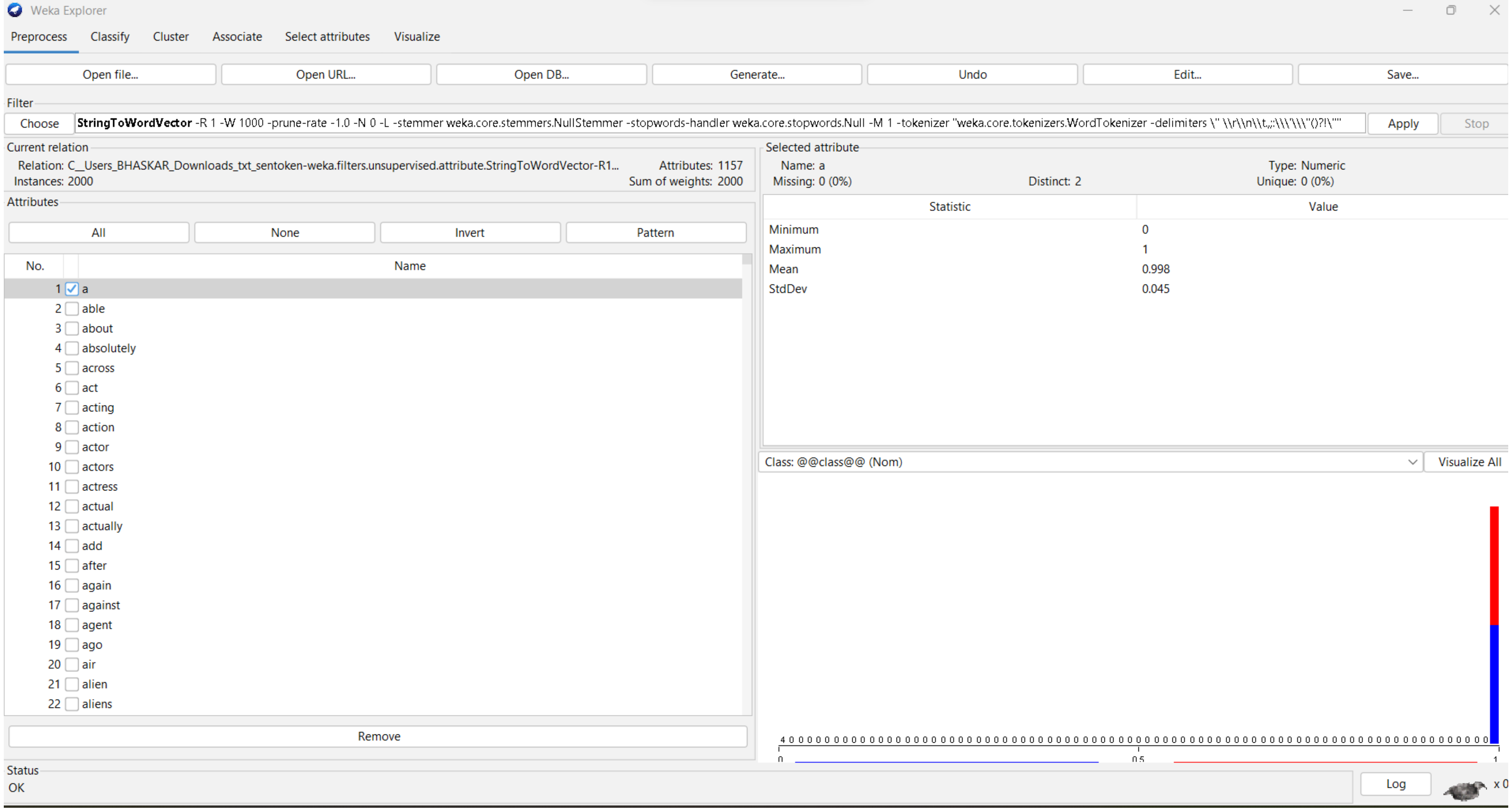
Click on Edit→ Right click on @@class@@ attribute→ Select attribute as class



elation: C__Users_BHASKAR_Downloads_txt_sentoken-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-st

1146: truman Numeric	1147: truth Numeric	1148: typical Numeric	1149: unique Numeric	1150: unlike Numeric	1151: view Numeric	1152: wars Numeric	1153: wedding Numeric	1154: wonderful Numeric	1155: wonderfully Numeric	1156: woody Numeric	1157: @@class@@ Nominal
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	neg
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	neg

Remove junk words



Filter

StringToWordVector -R 1 -W 1000 -prune-rate -1.0 -N 0 -L -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M 1 -tokenizer "weka.core.tokenizers.WordTokenizer -delimiters \" \\r\\n\\t.,;:\\\'\\\"()?!\""

Current relation

Relation: C__Users_BHASKAR_Downloads_txt_sentoken-weka.filters.unsupervised.attribute.StringToWordVector-R1...
Attributes: 1157
Instances: 2000
Sum of weights: 2000

Attributes

No.		Name
1	☑	a
2	□	able
3	□	about
4	□	absolutely
5	□	across
6	□	act
7	□	acting
8	□	action
9	□	actor
10	□	actors
11	□	actress
12	□	actual
13	□	actually
14	□	add
15	□	after
16	□	again
17	□	against
18	□	agent
19	□	ago
20	□	air
21	□	alien
22	□	aliens

Selected attribute

Name: a
Type: Numeric
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.998
StdDev	0.045

Class: @@class@@ (Nom)

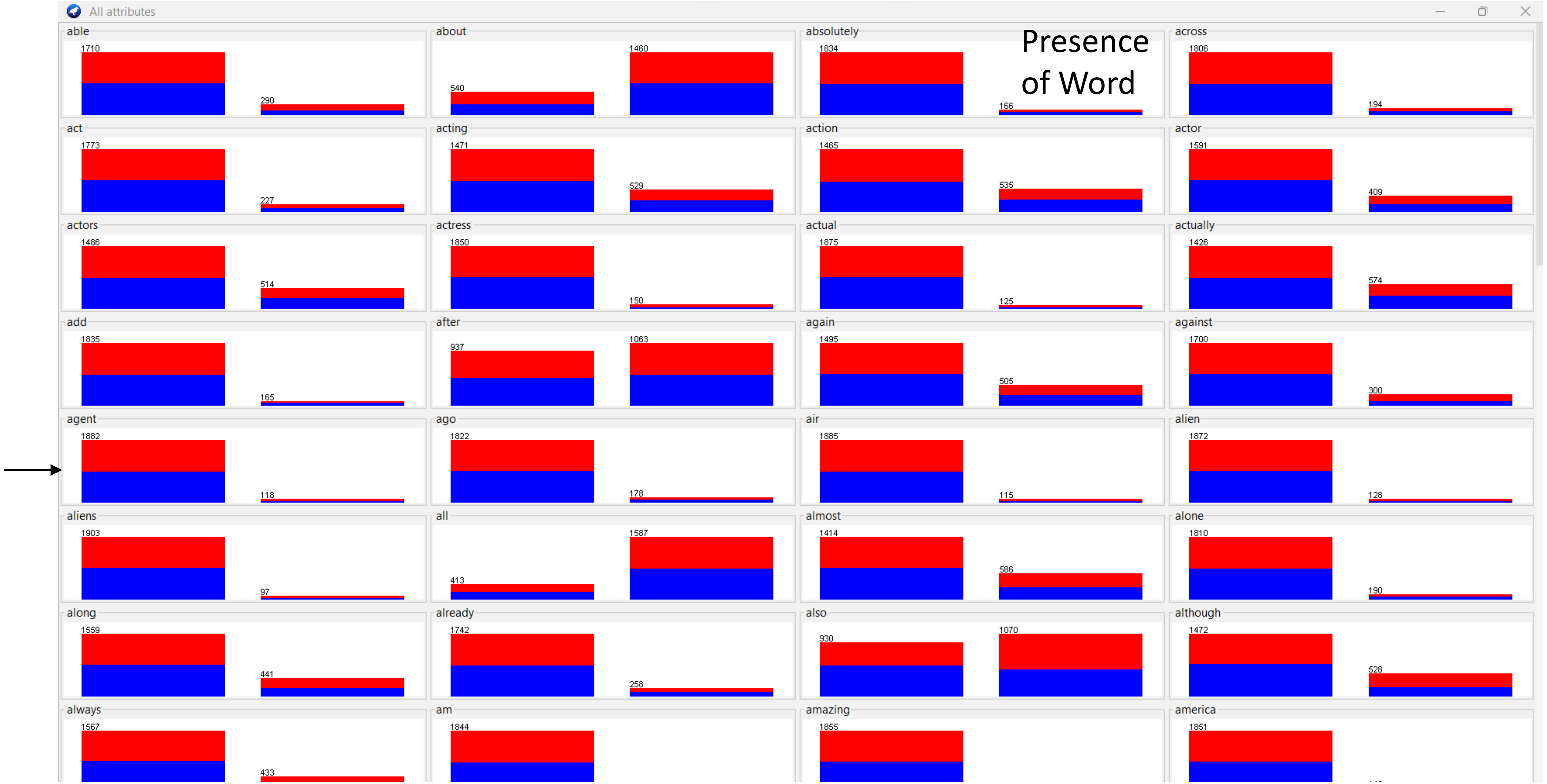
Status

OK

Converting Numeric to Nominal attribute

- Choose Filter→ Filter→ Unsupervised→ Attribute→ Numeric→Apply
- This will convert each word in string field to numeric attribute.
- Name of the attribute is word itself.
- The value of each attribute is 0 (absent) or 1 (present) in the current document.

Visualize each attribute as Histogram

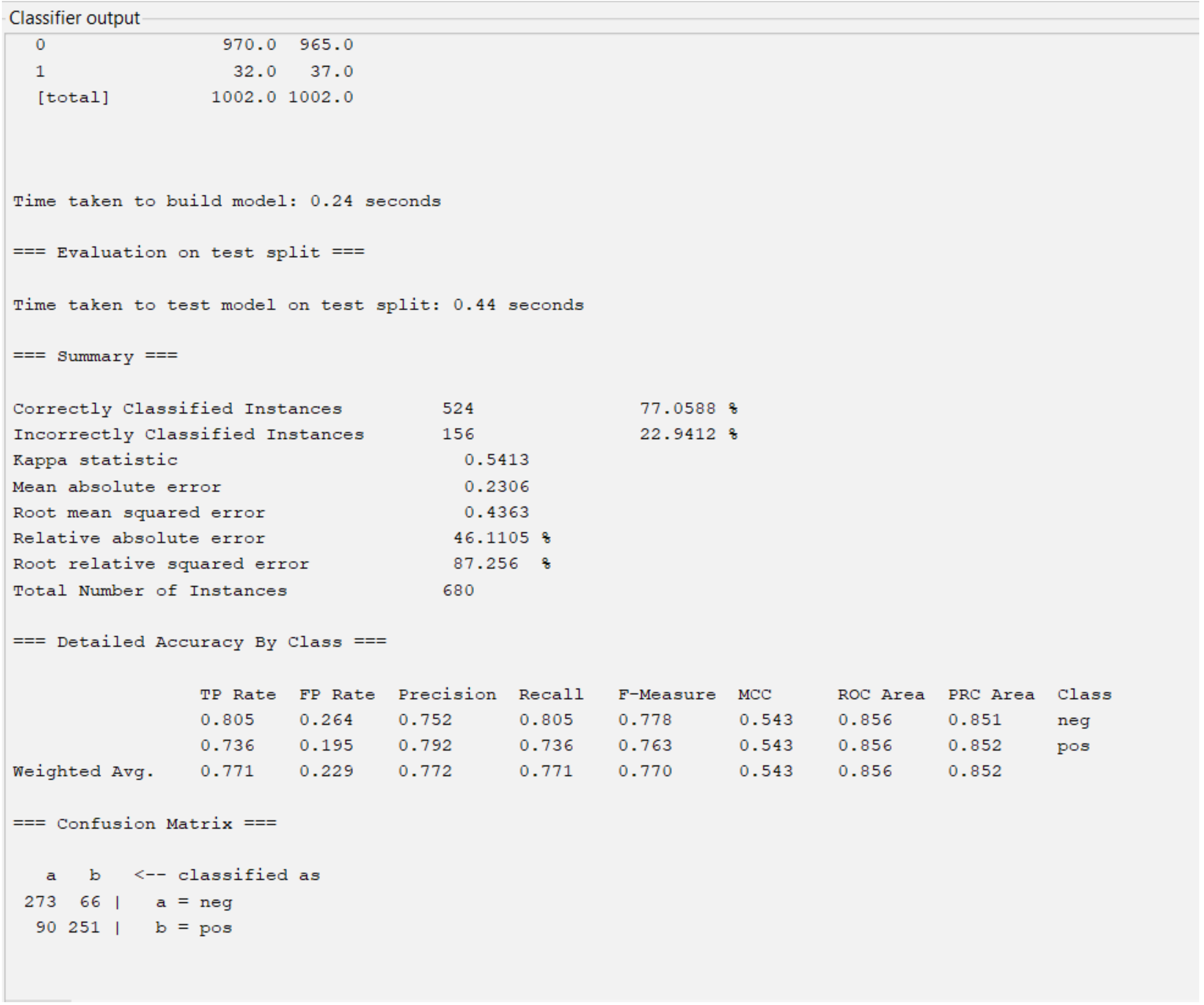


Presence of Word

Presence of Word →

Classification using Naive Bayes Classifier

- Classify→Choose Classifier→Bayes → NaiveBayes→ Select Default option Percentage Split 66% → Start



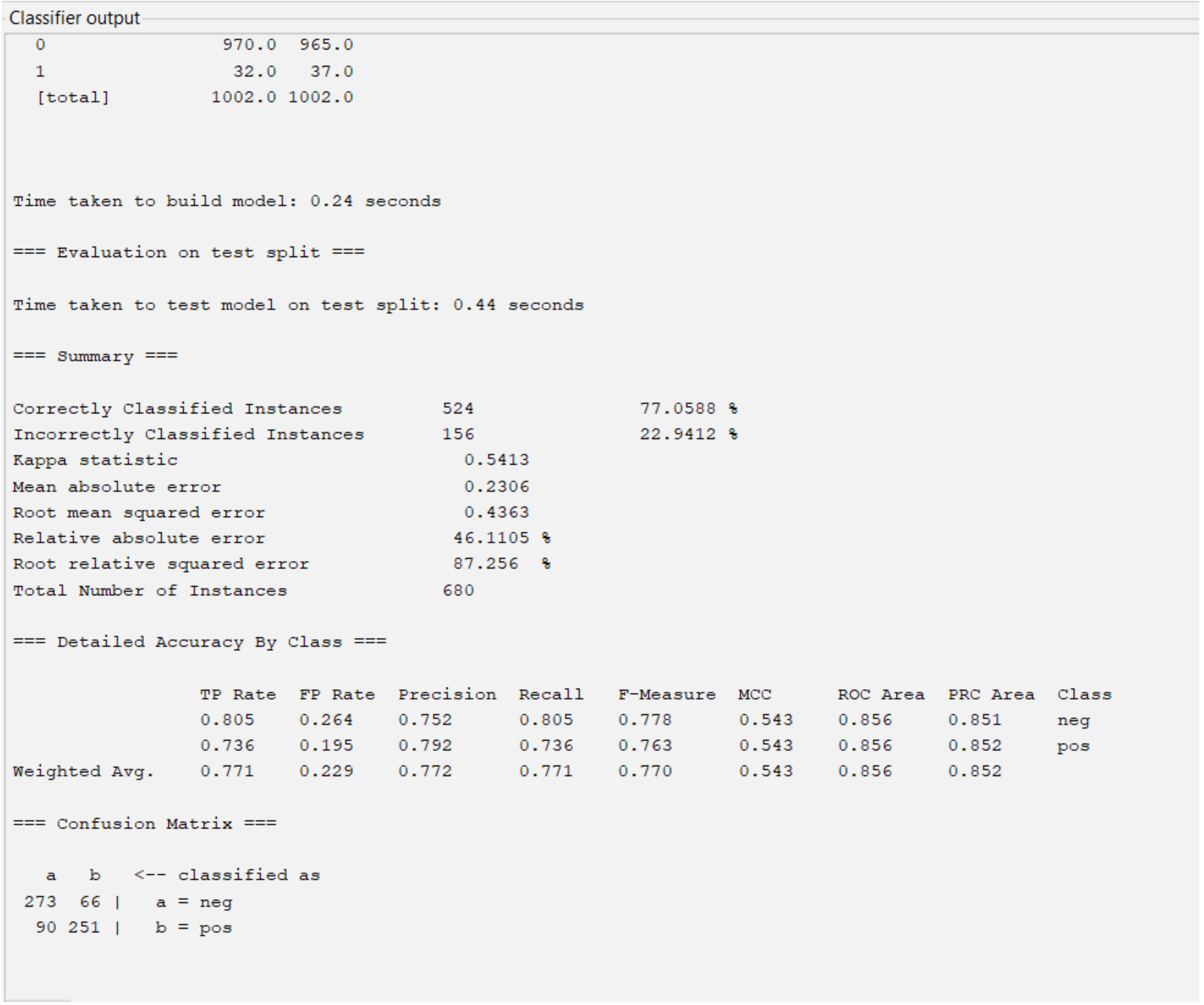
Classifier output

```
  0                 970.0  965.0
  1                  32.0   37.0
  [total]          1002.0 1002.0



Time taken to build model: 0.24 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.44 seconds

=== Summary ===

Correctly Classified Instances         524               77.0588 %
Incorrectly Classified Instances       156               22.9412 %
Kappa statistic                          0.5413
Mean absolute error                      0.2306
Root mean squared error                  0.4363
Relative absolute error                 46.1105 %
Root relative squared error             87.256  %
Total Number of Instances              680

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.805    0.264    0.752      0.805    0.778      0.543    0.856     0.851     neg
                 0.736    0.195    0.792      0.736    0.763      0.543    0.856     0.852     pos
Weighted Avg.    0.771    0.229    0.772      0.771    0.770      0.543    0.856     0.852

=== Confusion Matrix ===

   a   b   <-- classified as
 273  66 |   a = neg
  90 251 |   b = pos
```

Classification using Naive Bayes Classifier with all attributes

- Classify→Choose Classifier→Bayes → NaiveBayes→ Select Default option Percentage Split 66% → Start

Accuracy 77.0588 % for all attributes



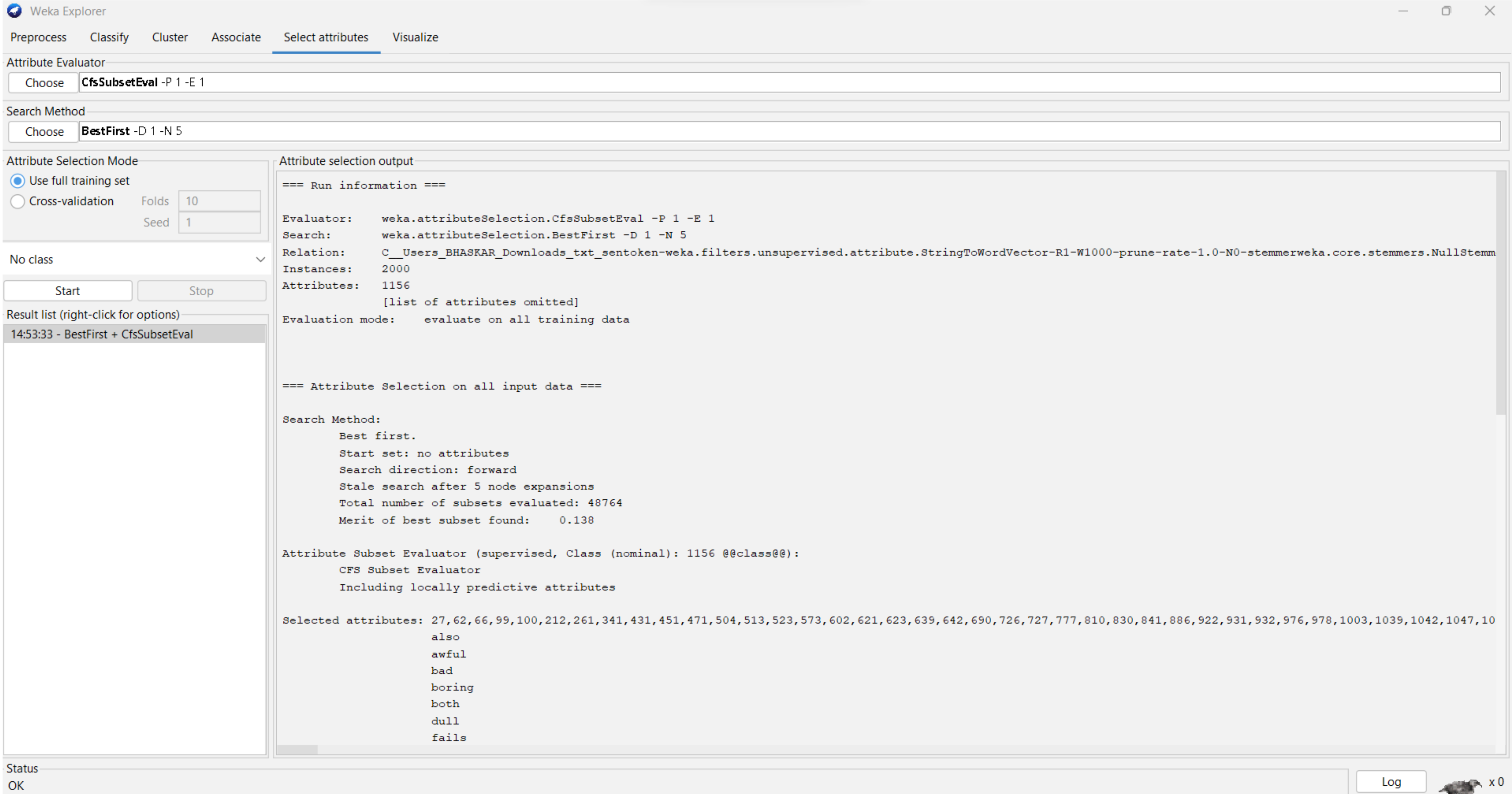
```
  0                  970.0   965.0
  1                   32.0    37.0
  [total]           1002.0  1002.0



Time taken to build model: 0.24 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.44 seconds

=== Summary ===

Correctly Classified Instances         524               77.0588 %
Incorrectly Classified Instances       156               22.9412 %
Kappa statistic                          0.5413
Mean absolute error                      0.2306
Root mean squared error                  0.4363
Relative absolute error                 46.1105 %
Root relative squared error             87.256  %
Total Number of Instances              680

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.805    0.264    0.752      0.805    0.778      0.543    0.856     0.851     neg
                 0.736    0.195    0.792      0.736    0.763      0.543    0.856     0.852     pos
Weighted Avg.    0.771    0.229    0.772      0.771    0.770      0.543    0.856     0.852

=== Confusion Matrix ===

   a   b   <-- classified as
 273  66 |   a = neg
  90 251 |   b = pos
```

Attribute Selection

- Select Attributes→Attribute Evauator=default, Search Method=default→Attribute Selection Mode=Use full trining set→ Start
- After getting output for attribute selection save that dataset with new name
- Right click on result list→ Save reduced data

Initially we have 1156 attributes, but after attribute selection only 52

Attribute Selection



Attribute Evaluator: CfsSubsetEval -P 1 -E 1

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode

- ● Use full training set
- ○ Cross-validation Folds 10 Seed 1

No class

Result list (right-click for options)

14:53:33 - BestFirst + CfsSubsetEval

Attribute selection output

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     C__Users_BHASKAR_Downloads_txt_sentoken-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemm
Instances:    2000
Attributes:   1156
              [list of attributes omitted]
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 48764
        Merit of best subset found:    0.138

Attribute Subset Evaluator (supervised, Class (nominal): 1156 @@class@@):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 27,62,66,99,100,212,261,341,431,451,471,504,513,523,573,602,621,623,639,642,690,726,727,777,810,830,841,886,922,931,932,976,978,1003,1039,1042,1047,10
                     also
                     awful
                     bad
                     boring
                     both
                     dull
                     fails
```

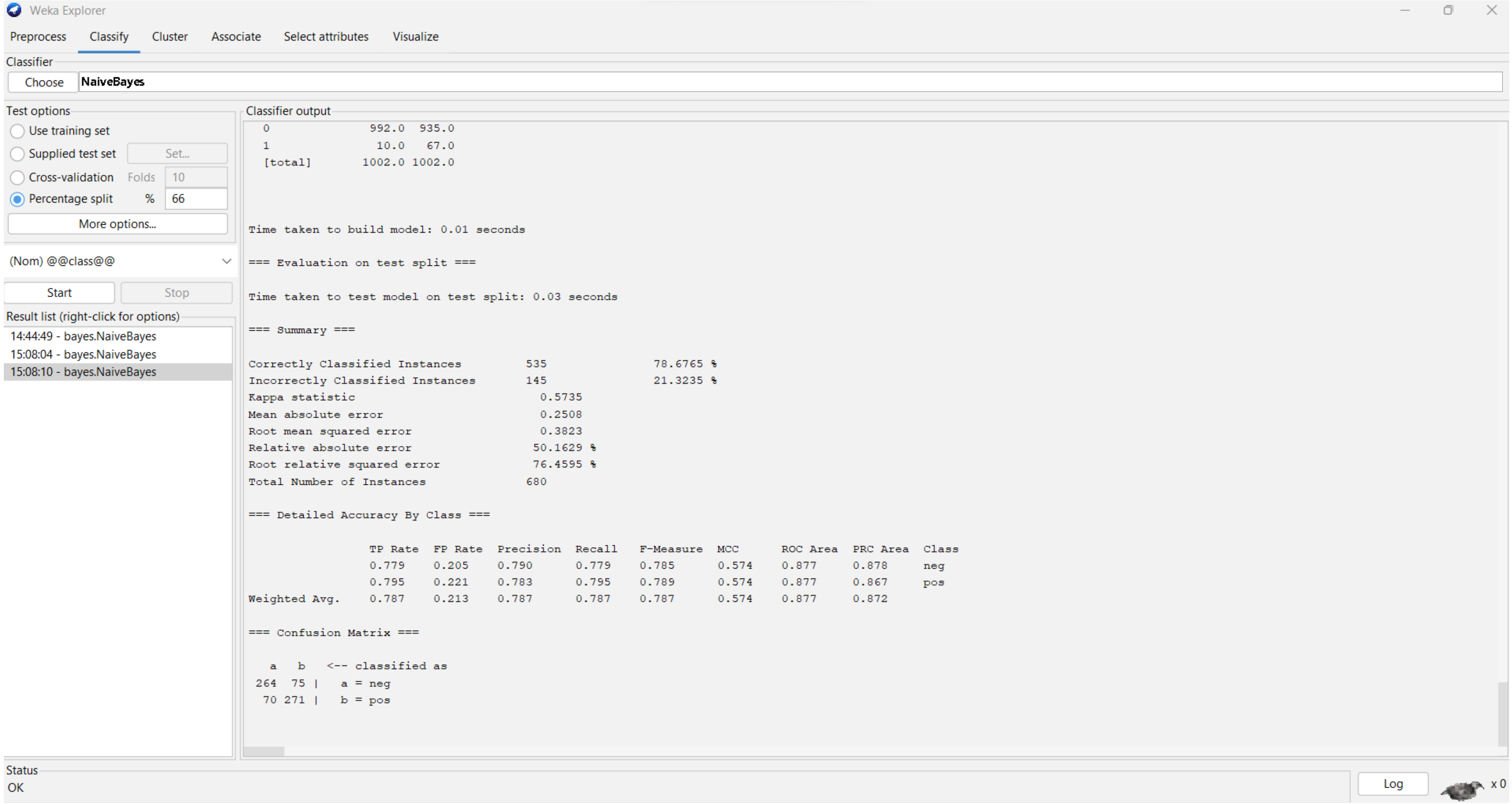
Status

OK

Classification using Naive Bayes Classifier with selected attributes

Initially
Accuracy 77.0588 %
for all attributes

After attribute Selection
Accuracy 78.6765 %



Classifier: NaiveBayes

Test options
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) @@class@@

Result list (right-click for options)
14:44:49 - bayes.NaiveBayes
15:08:04 - bayes.NaiveBayes
15:08:10 - bayes.NaiveBayes

Classifier output

```
 0               992.0  935.0
 1                10.0   67.0
 [total]        1002.0 1002.0



Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances         535               78.6765 %
Incorrectly Classified Instances       145               21.3235 %
Kappa statistic                          0.5735
Mean absolute error                      0.2508
Root mean squared error                  0.3823
Relative absolute error                 50.1629 %
Root relative squared error             76.4595 %
Total Number of Instances              680

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.779    0.205    0.790      0.779    0.785      0.574    0.877     0.878     neg
                 0.795    0.221    0.783      0.795    0.789      0.574    0.877     0.867     pos
Weighted Avg.    0.787    0.213    0.787      0.787    0.787      0.574    0.877     0.872

=== Confusion Matrix ===

   a   b   <-- classified as
 264  75 |   a = neg
  70 271 |   b = pos
```

Status
OK

Reference

- Han, Jiawei Kamber, Micheline Pei and Jian, “Data Mining: Concepts and Techniques”,Elsevier Publishers, ISBN:9780123814791, 9780123814807.