Text Classification and Clustering with WEKA

A guided example by

Sergio Jiménez

UNIVERSIDAD NACIONAL DE COLOMBIA

The Task

Building a model for movies revisions in English for classifying it into positive or negative.

Sentiment Polarity Dataset Version 2.0

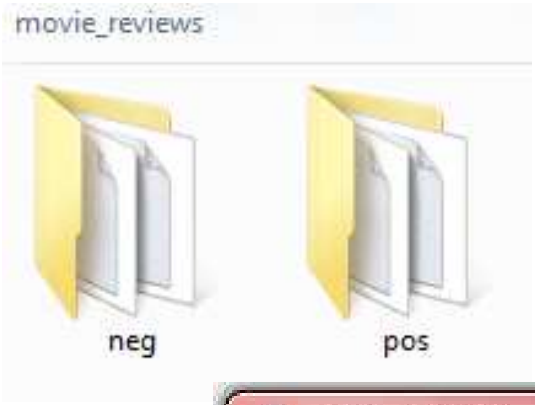
1000 positive movie review and 1000 negative review texts from:

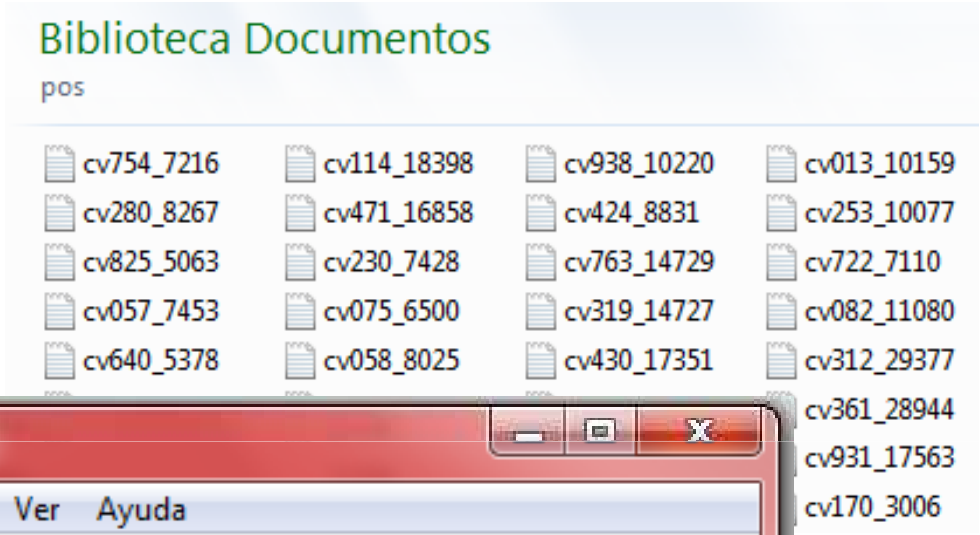
Thumbs up? Sentiment Classification using Machine Learning Techniques. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Proceedings of EMNLP, pp. 79--86, 2002.

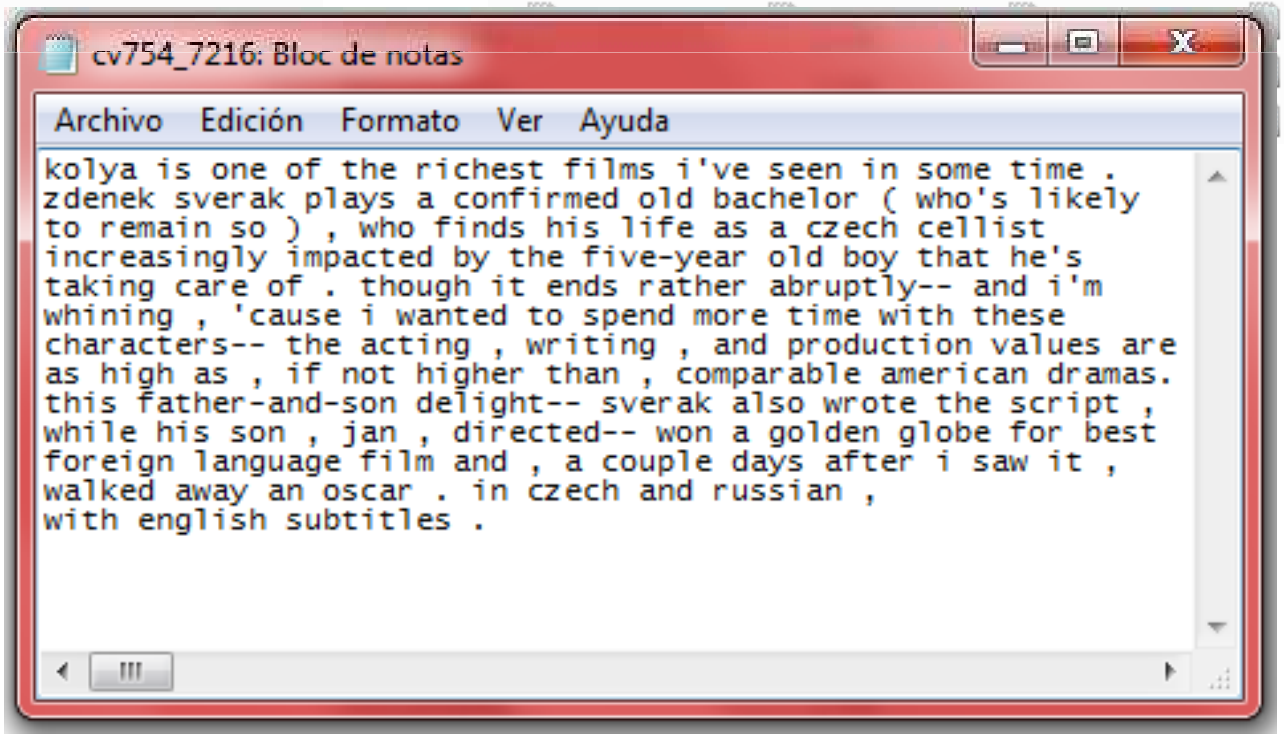
"Our data **source** was the **Internet Movie Database** (IMDb) archive of the rec.arts.movies.reviews newsgroup.3 We selected only reviews where the **author rating** was **expressed** either with stars or some **numerical value** (other conventions varied too widely to allow for automatic processing). Ratings were automatically extracted and converted into one of three categories: positive, negative, or neutral. For the work described in this paper, we concentrated **only** on discriminating between **positive** and **negative** sentiment."

http://www.cs.cornell.edu/people/pabo/movie-review-data/

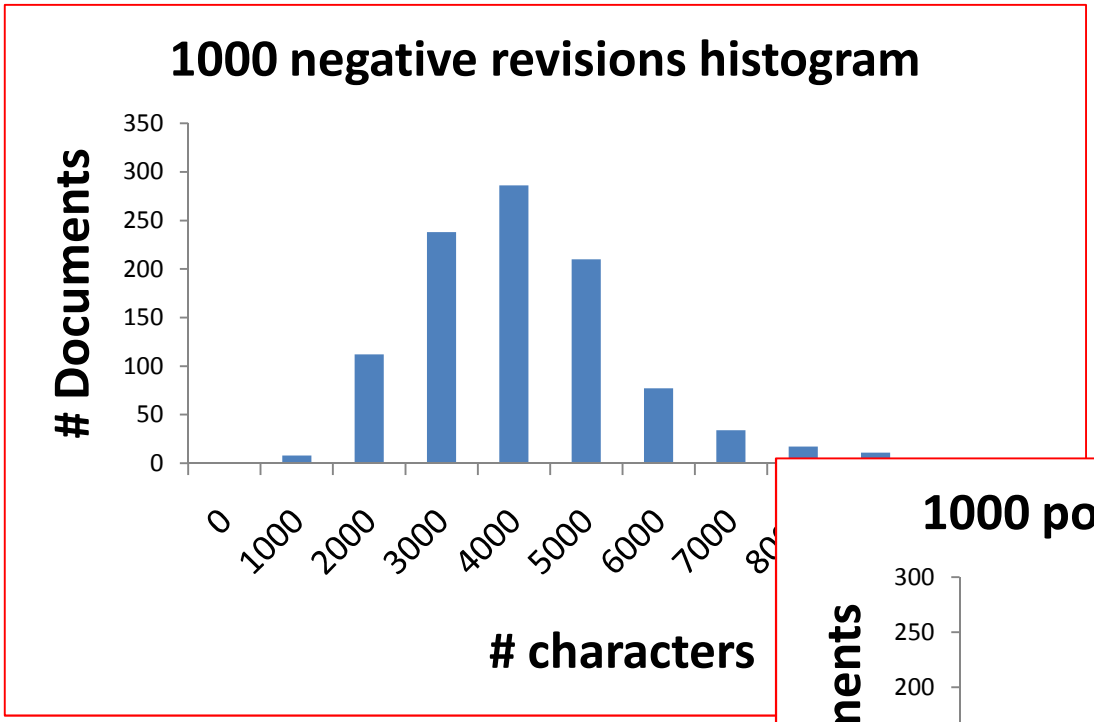
The Data (1/2)

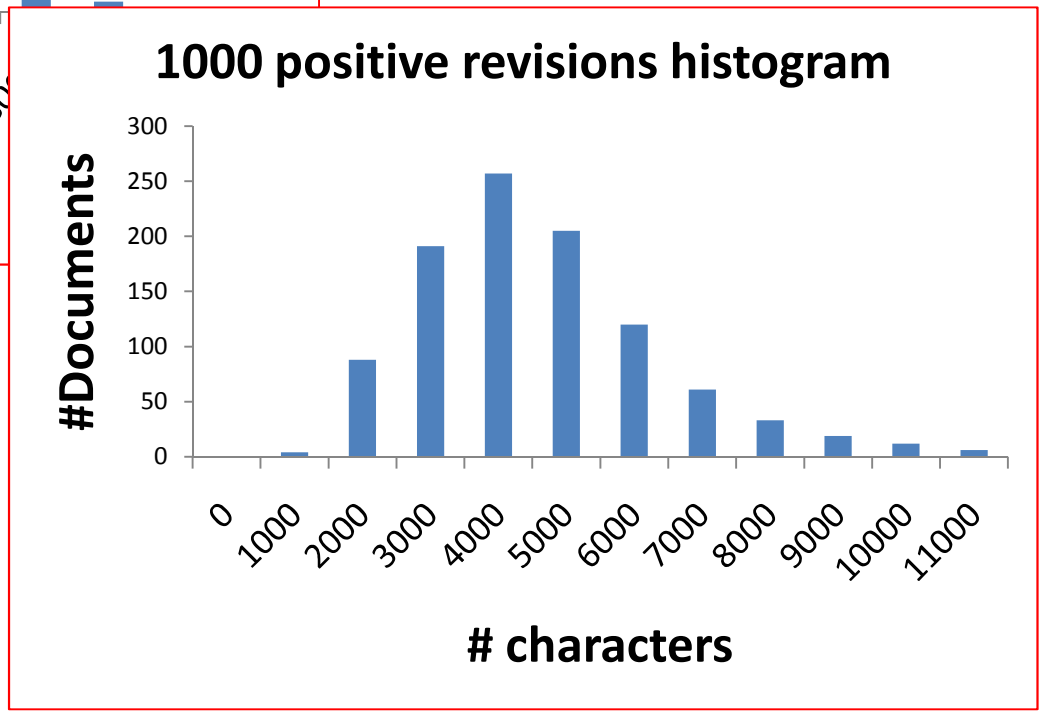

movie_reviews
neg
pos


Biblioteca Documentos
pos
cv754_7216
cv280_8267
cv825_5063
cv057_7453
cv640_5378
cv114_18398
cv471_16858
cv230_7428
cv075_6500
cv058_8025
cv938_10220
cv424_8831
cv763_14729
cv319_14727
cv430_17351
cv013_10159
cv253_10077
cv722_7110
cv082_11080
cv312_29377
cv361_28944
cv931_17563
cv170_3006


cv754_7216: Bloc de notas
Archivo Edición Formato Ver Ayuda
kolya is one of the richest films i've seen in some time .
zdenek sverak plays a confirmed old bachelor (who's likely
to remain so) , who finds his life as a czech cellist
increasingly impacted by the five-year old boy that he's
taking care of . though it ends rather abruptly-- and i'm
whining , 'cause i wanted to spend more time with these
characters-- the acting , writing , and production values are
as high as , if not higher than , comparable american dramas.
this father-and-son delight-- sverak also wrote the script ,
while his son , jan , directed-- won a golden globe for best
foreign language film and , a couple days after i saw it ,
walked away an oscar . in czech and russian ,
with english subtitles .

The Data (2/2)

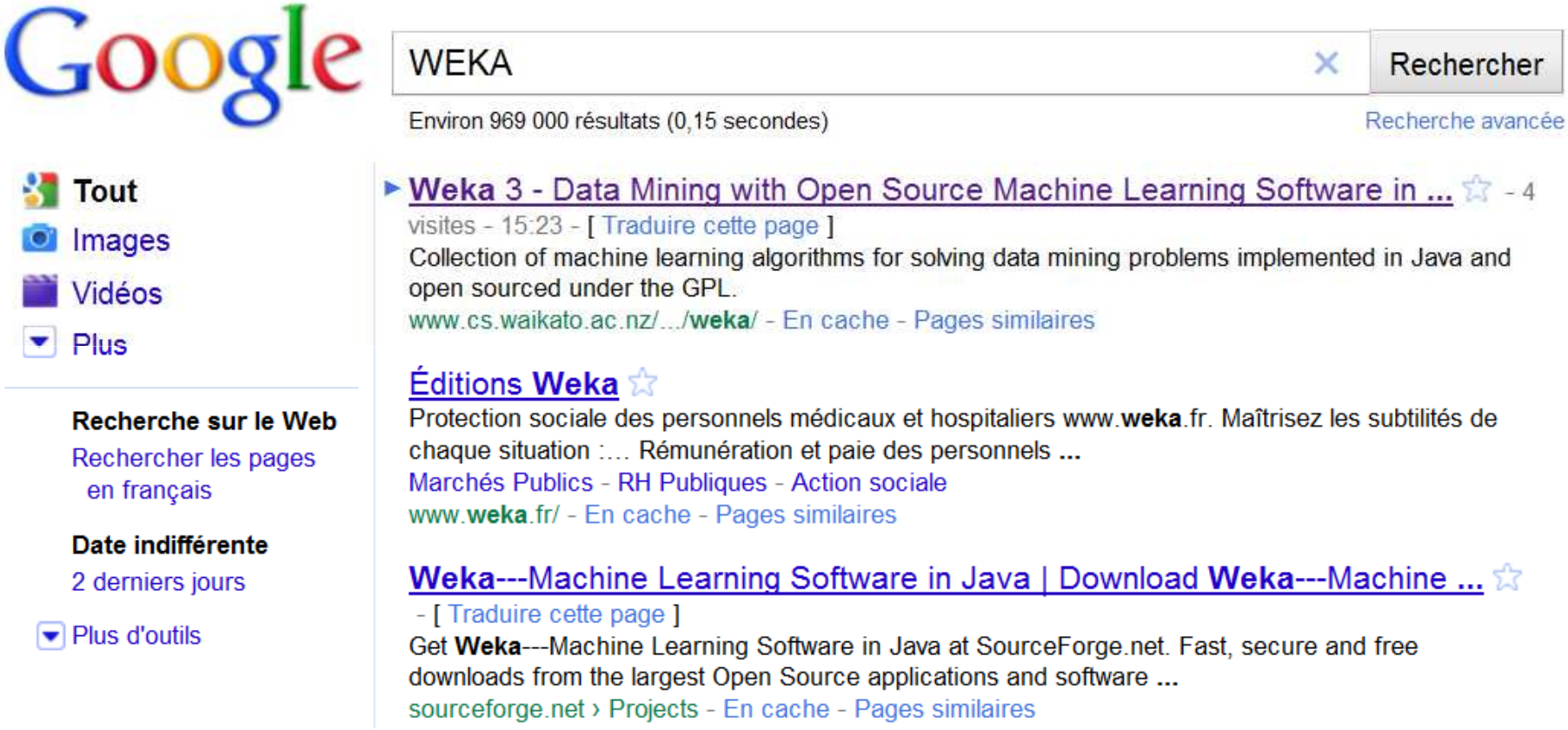

1000 negative revisions histogram
Documents
350
300
250
200
150
100
50
0
0
1000
2000
3000
4000
5000
6000
7000
characters


1000 positive revisions histogram
#Documents
300
250
200
150
100
50
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
characters

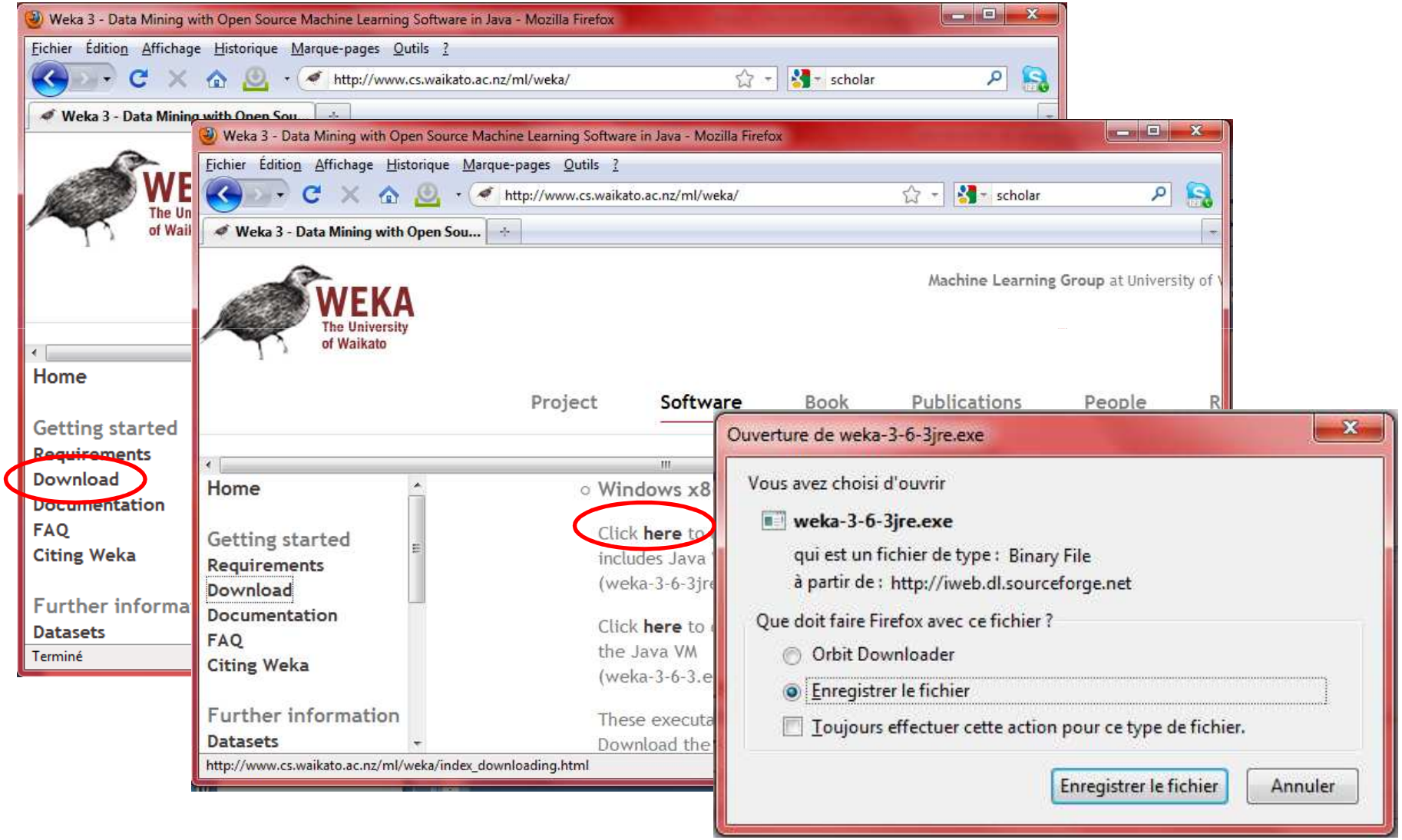
What WEKA is?

- “Weka is a collection of machine learning algorithms for data mining tasks”.
- “Weka contains tools for:
  - data pre-processing,
  - classification,
  - regression,
  - clustering,
  - association rules,
  - and visualization”

Where to start?

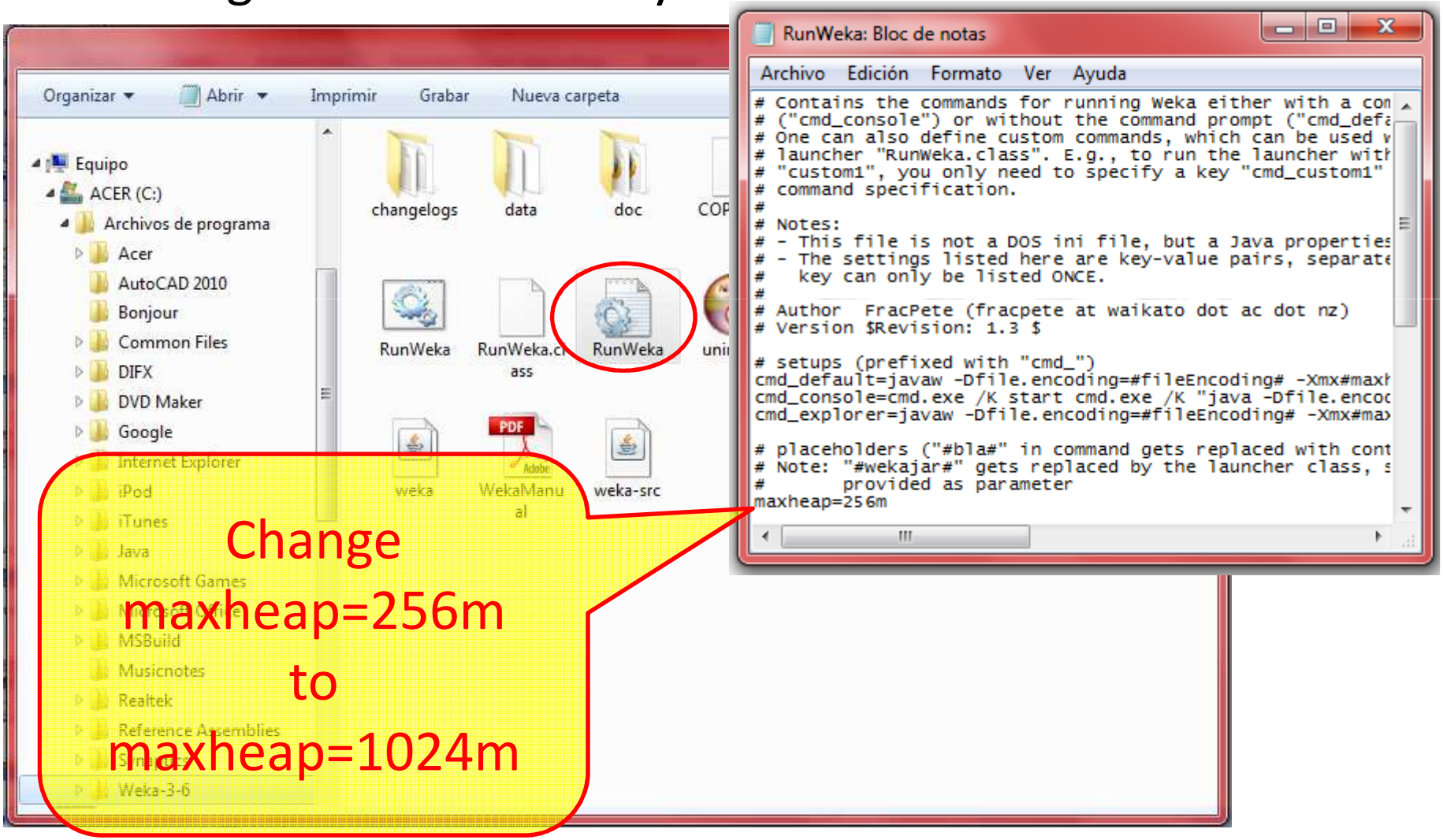


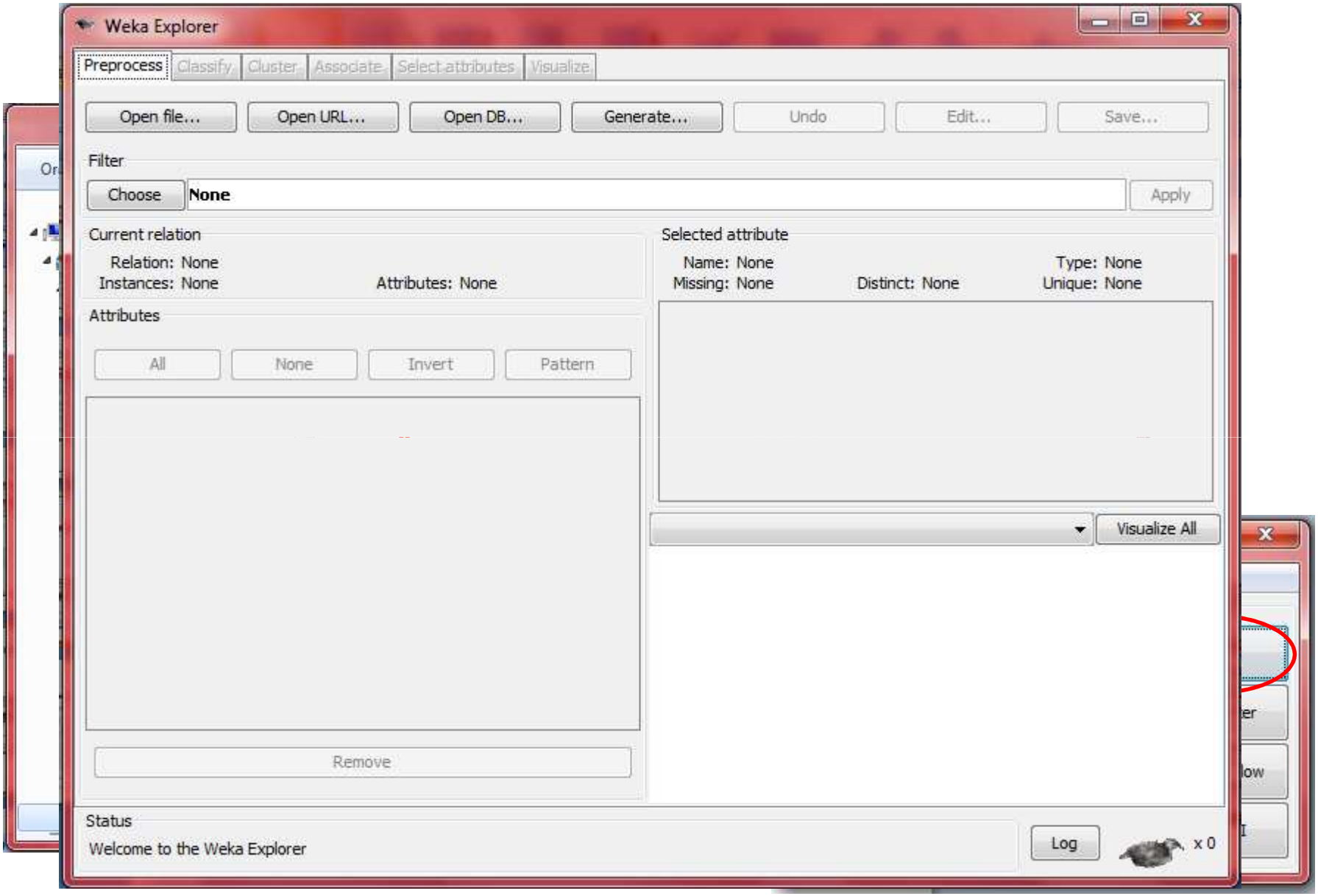
Getting WEKA



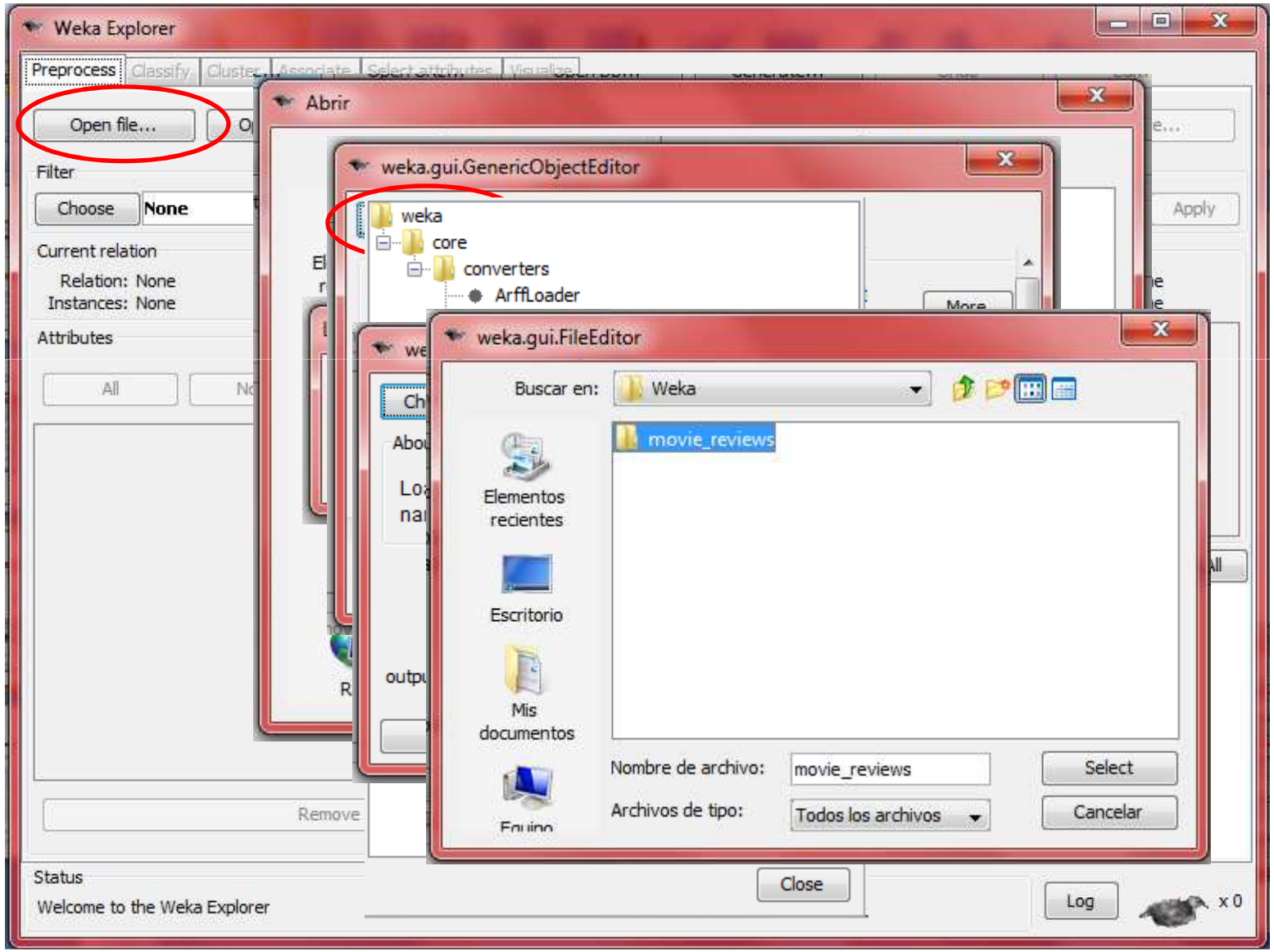
Before Running WEKA

Increasing available memory for Java in RunWeka.ini

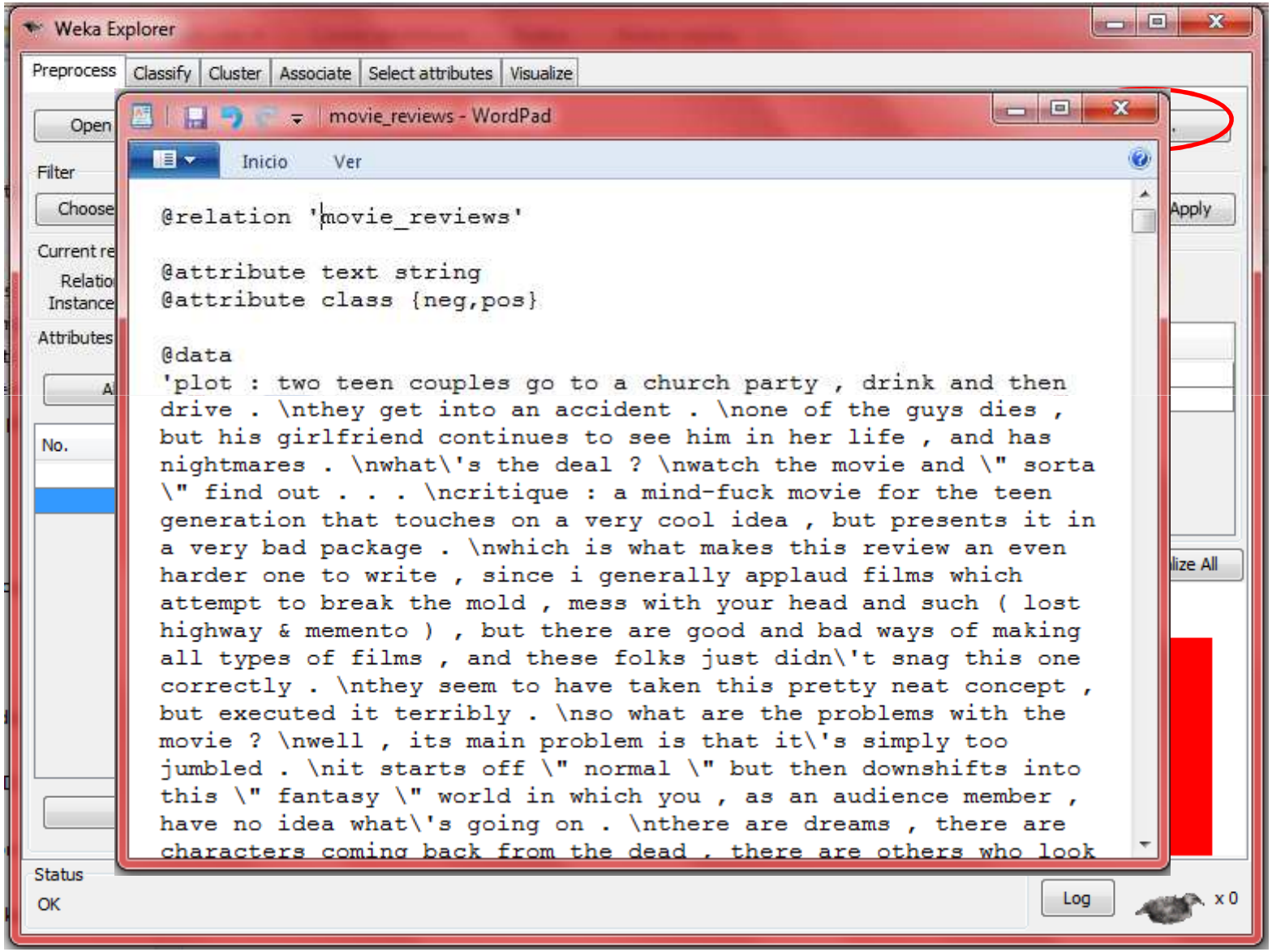



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Missing: None
Distinct: None
Type: None
Unique: None
Attributes
All
None
Invert
Pattern
Visualize All
Remove
Status
Welcome to the Weka Explorer
Log
x 0

Creating a .arff dataset



Saving the .arff dataset



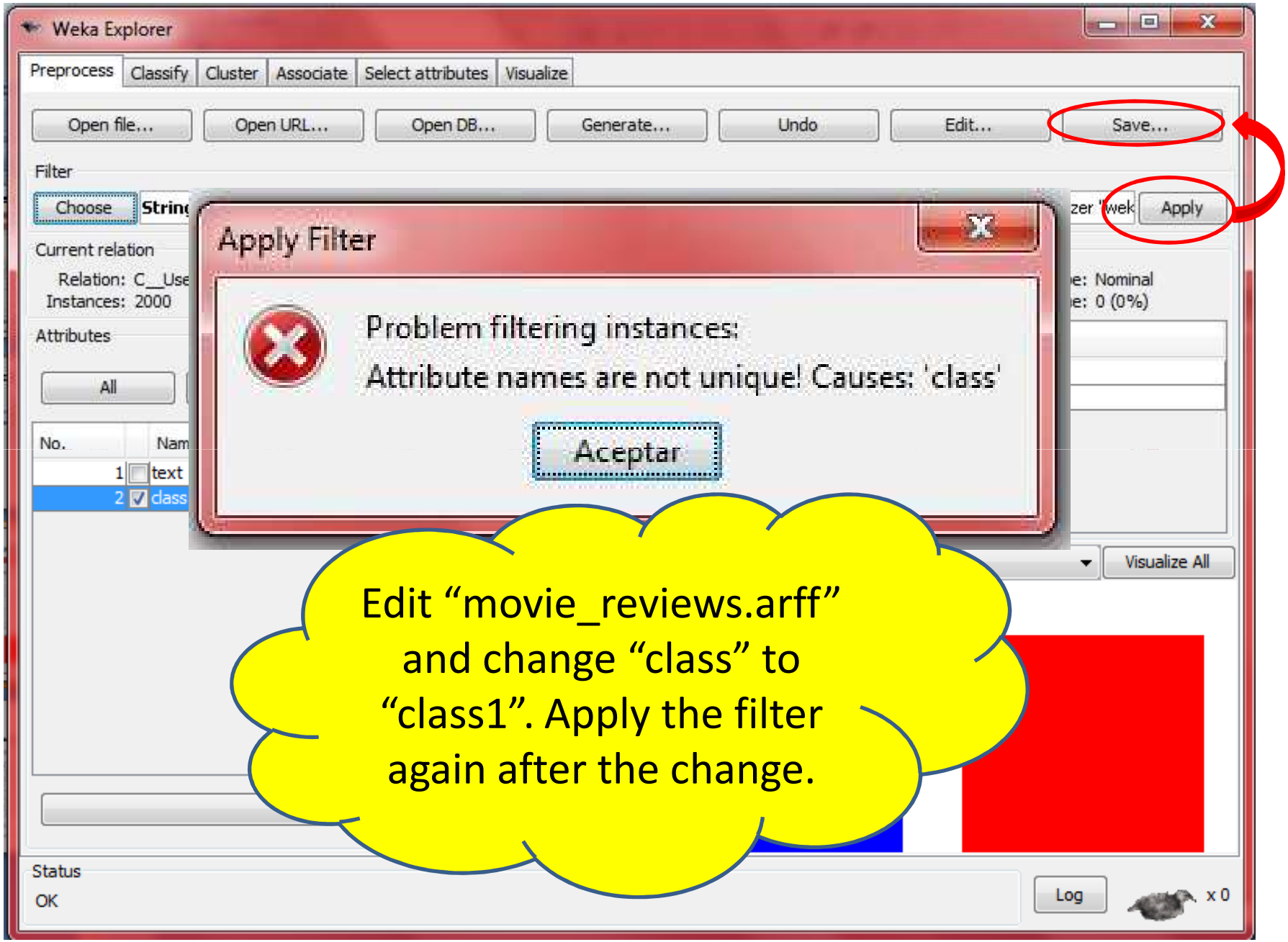
From text to vectors

$$V = [v_1, v_2, v_3, \cdots, v_n, class]$$

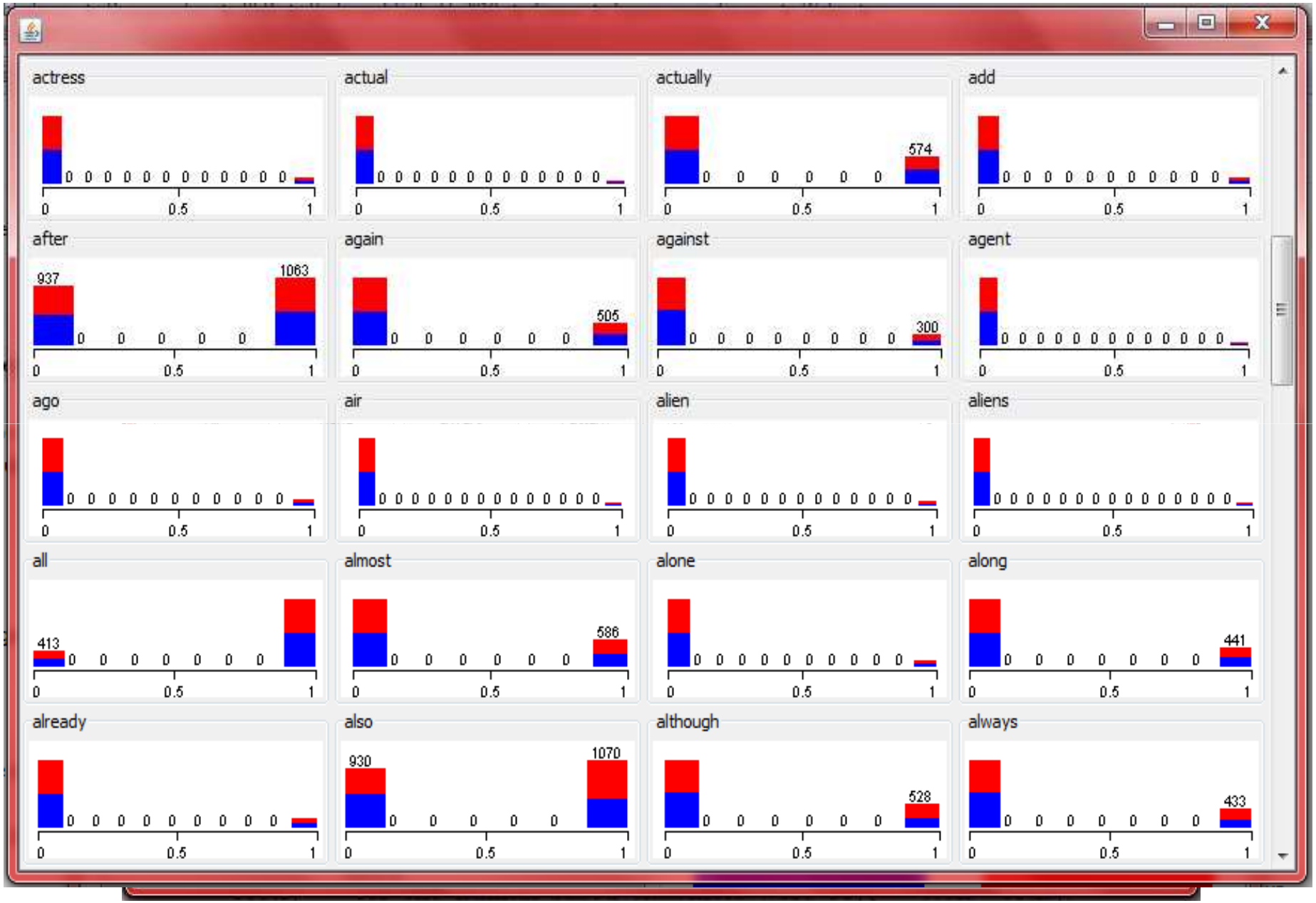
review$_1$="great movie" review$_3$="worst film ever"

review$_2$="excellent film" review$_4$="sucks"

	ever	excellent	film	great	movie	sucks	worst	
$V_1 =$	0,	0,	0,	1,	1,	0,	0,	+]
$V_2 =$	0,	1,	1,	0,	0,	0,	0,	+]
$V_3 =$	1,	0,	1,	0,	0,	0,	1,	−]
$V_4 =$	0,	0,	0,	0,	0,	1,	0,	−]



Edit “movie_reviews.arff” and change “class” to “class1”. Apply the filter again after the change.


actress
actual
actually
574
add
after
937
1063
again
505
against
300
agent
ago
air
alien
aliens
all
413
almost
586
alone
along
441
already
also
930
1070
although
528
always
433
0
0.5
1

StringToWordV

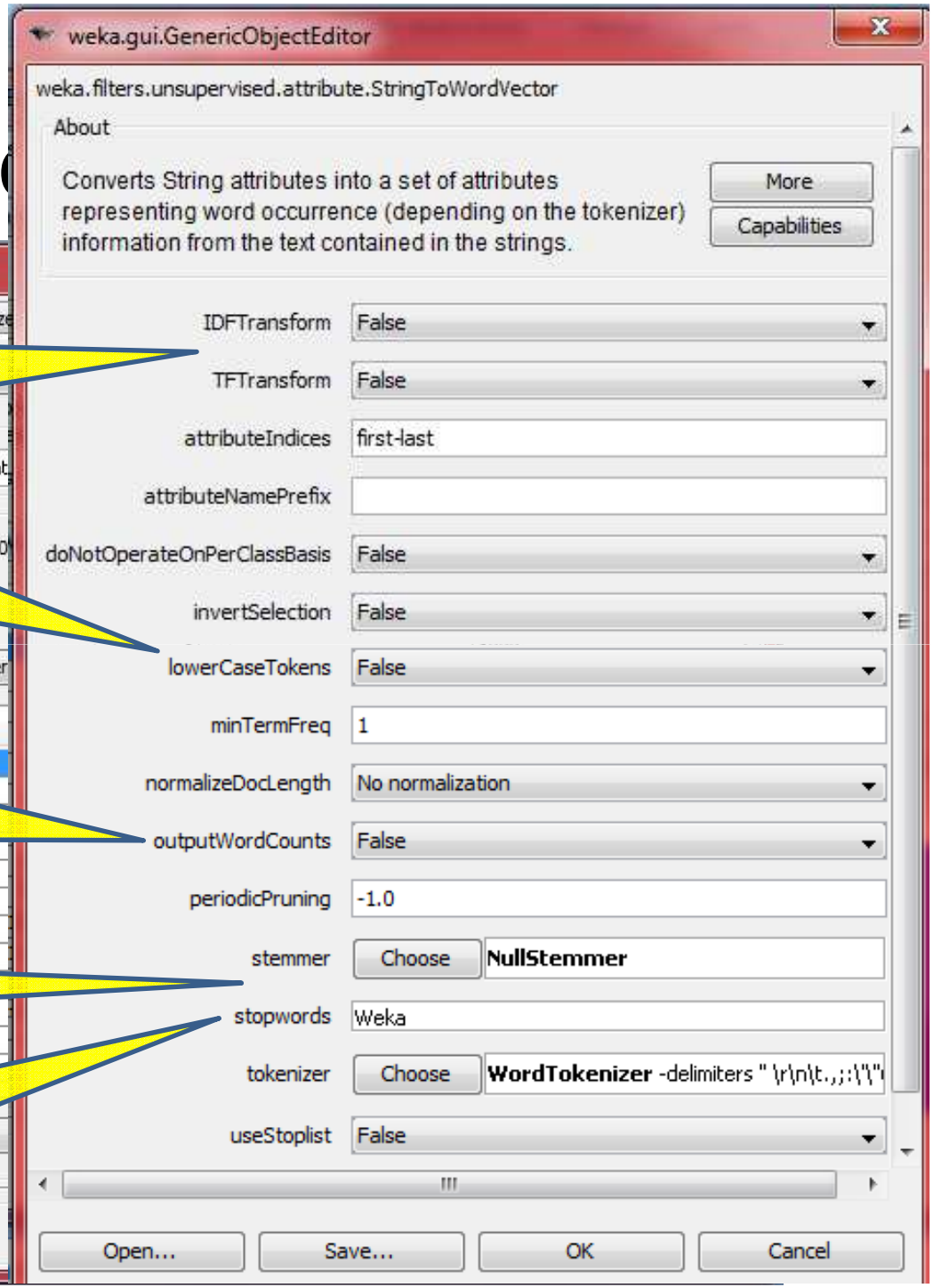
TF-IDF weigthing

lowerCase convertion

Use frequencies instead of single presence

Stemming

Stopwords removal using a list of words in a file



Generating datasets for experiments

dataset file name	***Stopwords***	***Stemming***	***Presence or freq.***
movie_reviews_1.arff		no	presence
movie_reviews_2.arff		no	frequency
movie_reviews_3.arff		yes	presence
movie_reviews_4.arff		yes	frequency
movie_reviews_5.arff	removed	no	presence
movie_reviews_6.arff	removed	no	frequency
movie_reviews_7.arff	removed	yes	presence
movie_reviews_8.arff	removed	yes	frequency

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **NaiveBayes**

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) class1

Start | Stop

Result list (right-click for options)

20:23:48 - bayes.NaiveBayes

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     C__Users_Ser_Documents_PhD_Visita prof Gelbukh 2010_Curso_Expos
Instances:    2000
Attributes:   1166
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

Status

Building model on training data...

Log | x 1

Results

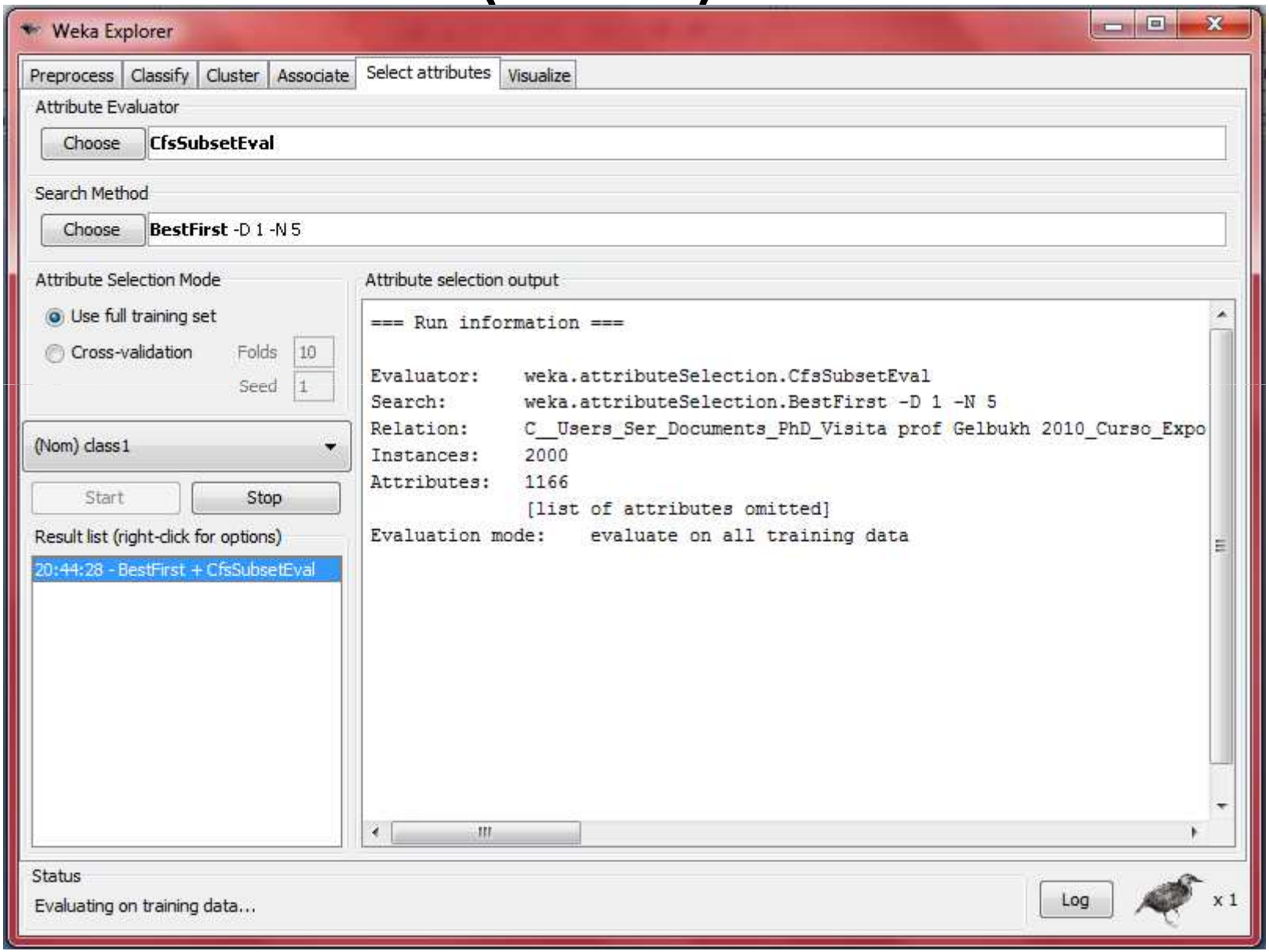
```
Correctly Classified Instances        1616               80.8     %
Incorrectly Classified Instances       384               19.2     %
Kappa statistic                          0.616
Mean absolute error                      0.1918
Root mean squared error                  0.4111
Relative absolute error                 38.3507 %
Root relative squared error             82.2217 %
Total Number of Instances             2000
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.832     0.216      0.794     0.832     0.813      0.897
                 0.784     0.168      0.824     0.784     0.803      0.897
Weighted Avg.    0.808     0.192      0.809     0.808     0.808      0.897

=== Confusion Matrix ===

   a   b   <-- classified as
 832 168 |   a = neg
 216 784 |   b = pos
```

Results Correctly Classified Reviews

dataset name	*Stopwords*	*Stemming*	*Presence or freq.*	*Naive Bayes 3-fold*	*NaiveBayes Multinomial 3-fold*
movie_reviews_1.arff		no	presence	80.65%	83.80%
movie_reviews_2.arff		no	frequency	69.30%	78.65%
movie_reviews_3.arff		yes	presence	79.40%	82.15%
movie_reviews_4.arff		yes	frequency	68.10%	79.70%
movie_reviews_5.arff	removed	no	presence	**81.80%**	**84.35%**
movie_reviews_6.arff	removed	no	frequency	69.40%	81.75%
movie_reviews_7.arff	removed	yes	presence	78.90%	82.40%
movie_reviews_8.arff	removed	yes	frequency	68.30%	80.50%

Attribute (word) Selecction

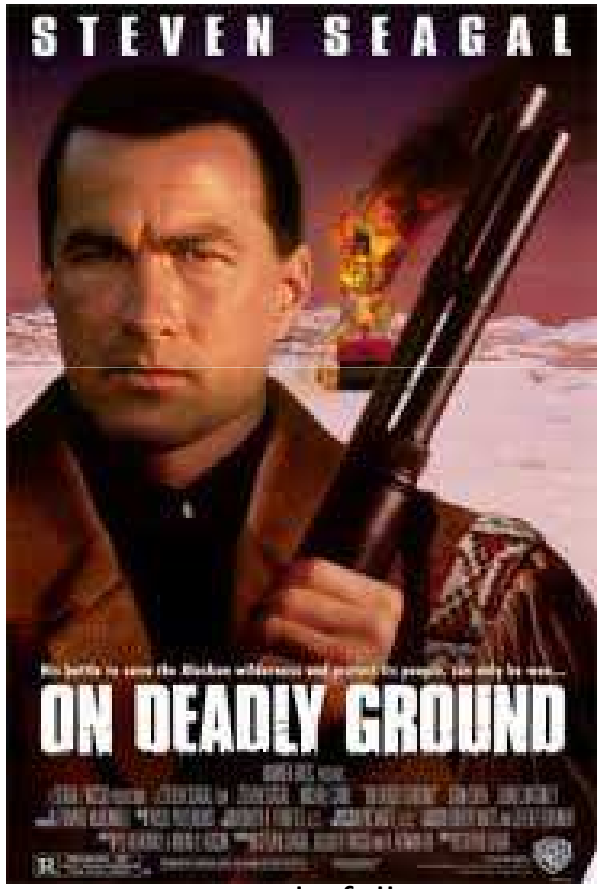


Selected Attributes (words)

also
awful
bad
boring
both
dull
fails
great
joke
lame
life
many
maybe
mess
nothing
others
perfect
performances

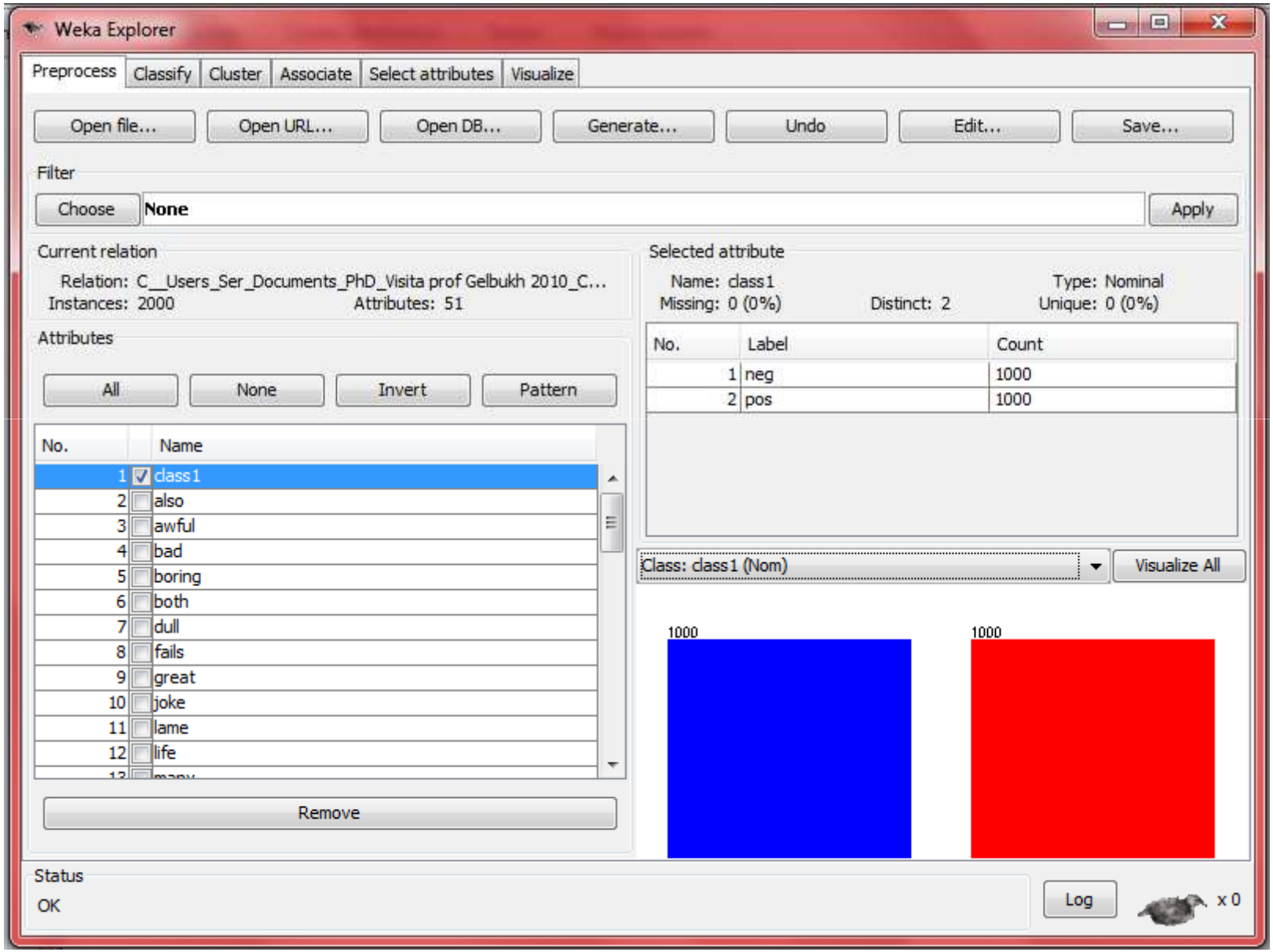
pointless
poor
ridiculous
script
seagal
sometimes
stupid
tale
terrible
true
visual
waste
wasted
world
worst
animation
definitely

deserves

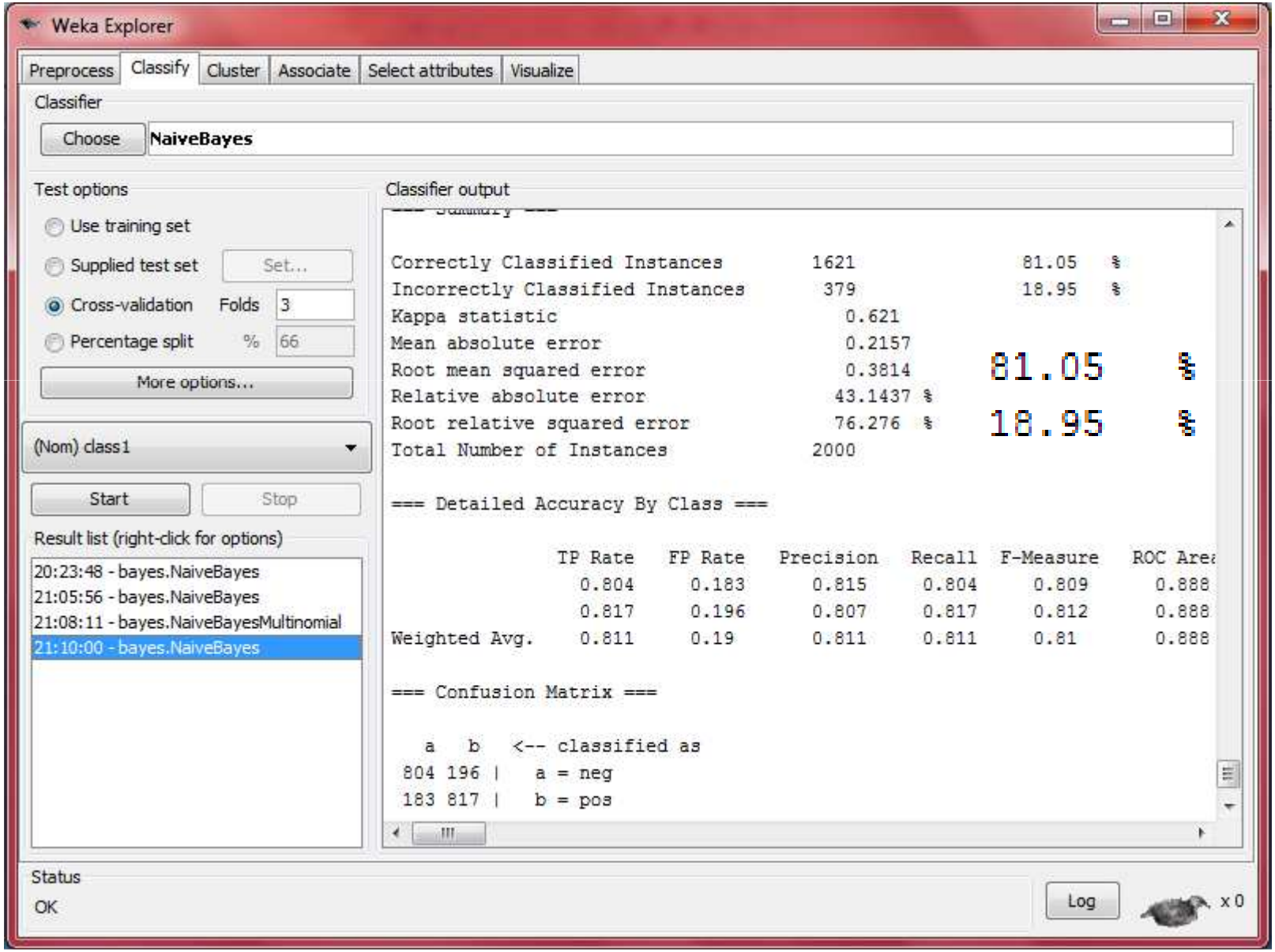


wonderfully

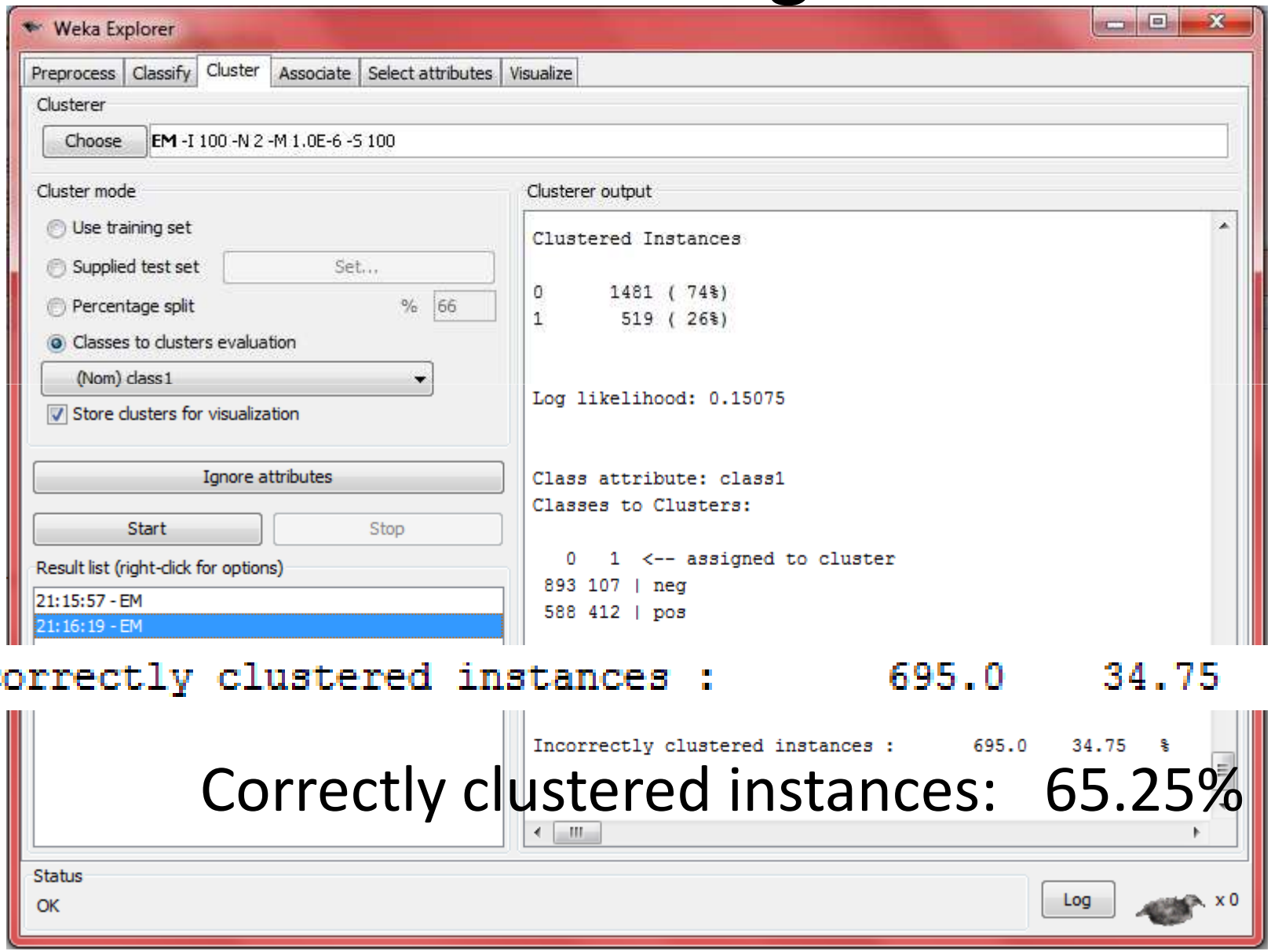
Pruned movie_reviews_1.arff dataset



Naïve Bayes with the pruned dataset



Clustering



Incorrectly clustered instances : 695.0 34.75 %

Correctly clustered instances: 65.25%

Other results

Results of Pang et al. (2002) with version 1.0 of the dataset with 700+ and 700-

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

No stemming or stoplists were used.

are the average three-fold cross-validation results

Thanks