

Text Classification and Clustering with WEKA

A guided example by
Sergio Jiménez



The Task

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



Sentiment Polarity Dataset Version 2.0

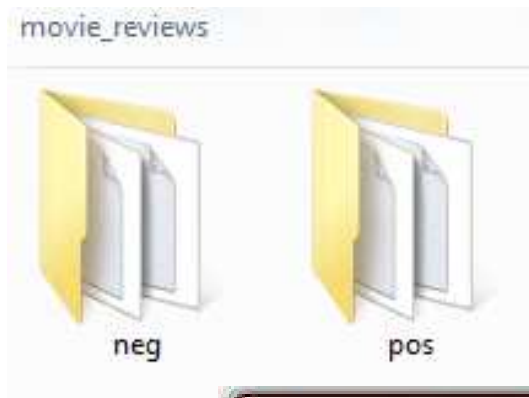
1000 positive movie review and 1000 negative review texts from:

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/>

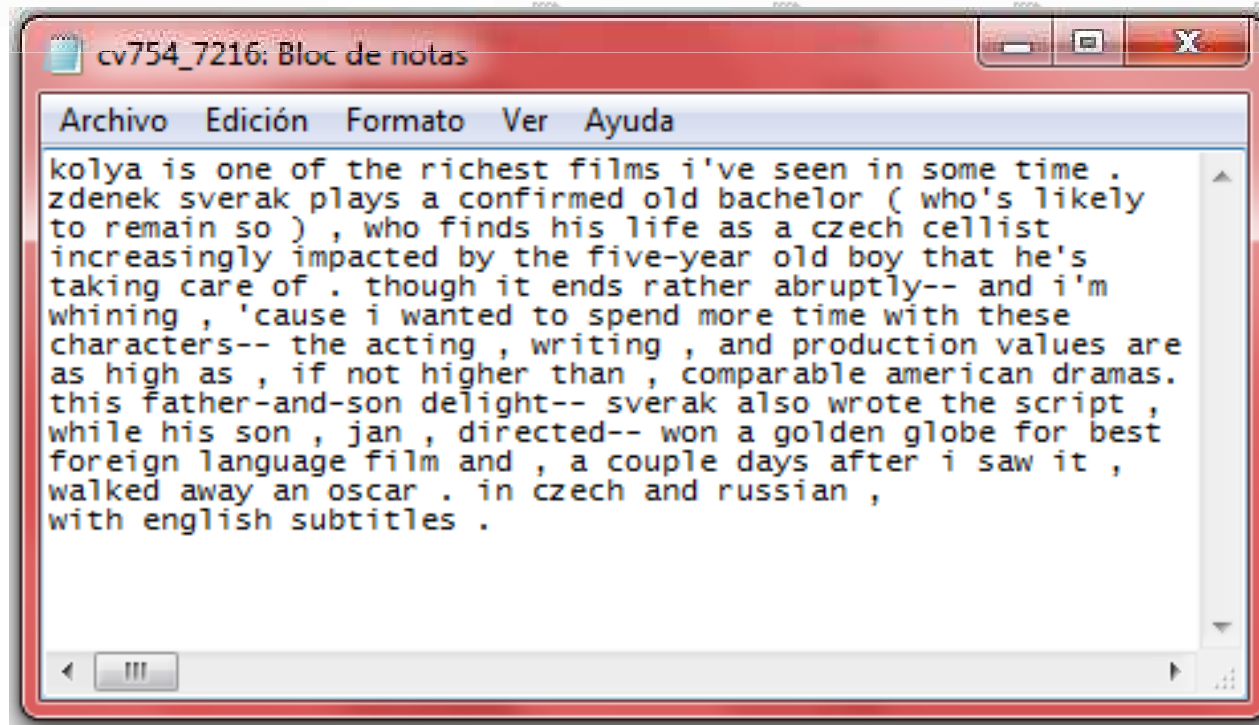
The Data (1/2)



Biblioteca Documentos

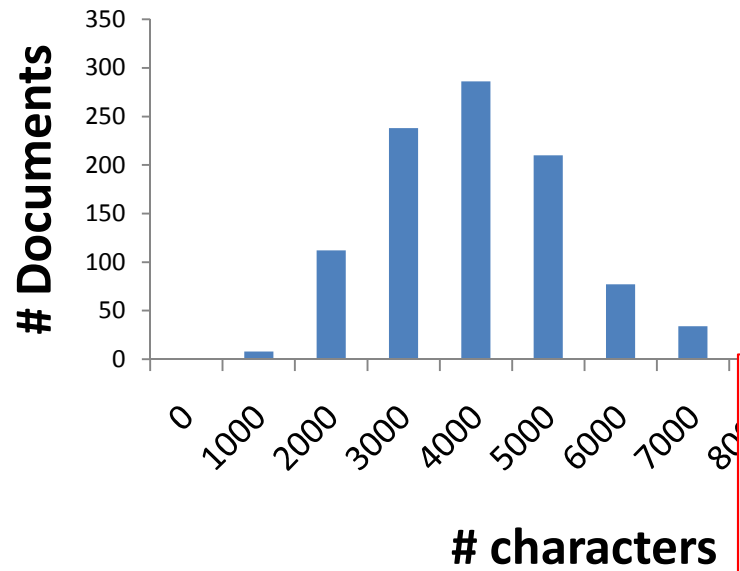
pos

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

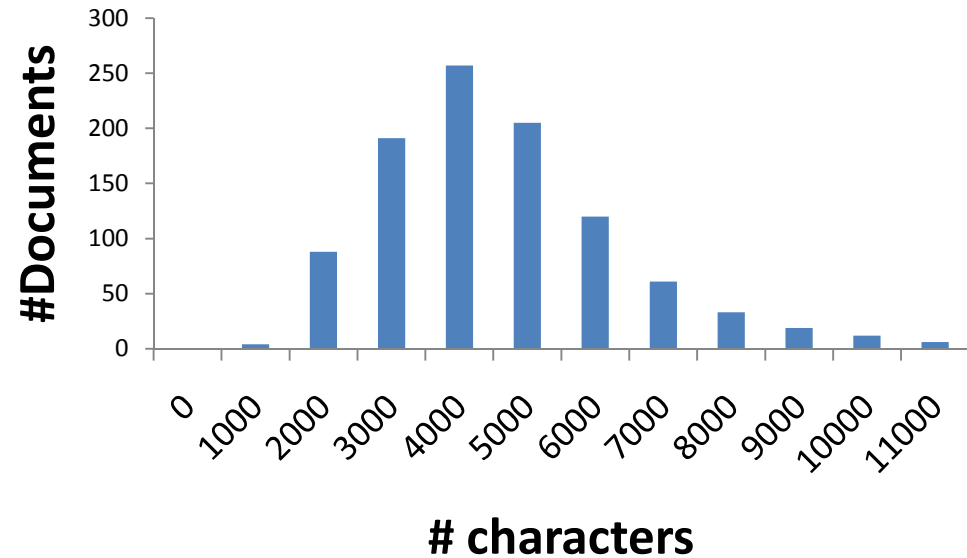


The Data (2/2)

1000 negative revisions histogram



1000 positive revisions histogram



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?



WEKA



Rechercher

Environ 969 000 résultats (0,15 secondes)

[Recherche avancée](#)



Tout



Images



Vidéos



Plus

Recherche sur le Web

[Rechercher les pages en français](#)

Date indifférente

[2 derniers jours](#)



Plus d'outils

► [Weka 3 - Data Mining with Open Source Machine Learning Software in ...](#) ☆ - 4

visites - 15:23 - [[Traduire cette page](#)]

Collection of machine learning algorithms for solving data mining problems implemented in Java and open sourced under the GPL.

www.cs.waikato.ac.nz/~weka/ - En cache - [Pages similaires](#)

[Éditions Weka](#) ☆

Protection sociale des personnels médicaux et hospitaliers www.weka.fr. Maîtrisez les subtilités de chaque situation :... Rémunération et paie des personnels ...

[Marchés Publics](#) - [RH Publiques](#) - [Action sociale](#)

www.weka.fr/ - En cache - [Pages similaires](#)

[Weka---Machine Learning Software in Java | Download Weka---Machine ...](#) ☆

- [[Traduire cette page](#)]

Get **Weka**---Machine Learning Software in Java at SourceForge.net. Fast, secure and free downloads from the largest Open Source applications and software ...

[sourceforge.net > Projects](http://sourceforge.net/projects/) - En cache - [Pages similaires](#)

Getting WEKA

The image shows a sequence of steps to download WEKA. It features two overlapping Mozilla Firefox browser windows and a Windows file dialog box.

The top browser window displays the WEKA website homepage. The navigation menu includes: Home, Project, Software, Book, Publications, People, and Resources. The left sidebar contains links: Home, Getting started, Requirements, **Download**, Documentation, FAQ, Citing Weka, Further information, Datasets, and Terminé. The **Download** link in the sidebar is circled in red.

The bottom browser window shows a page with a list of download links. The link "Click here to download the Java VM (weka-3-6-3jre.exe)" is circled in red.

The foreground dialog box, titled "Ouverture de weka-3-6-3jre.exe", displays the following information:

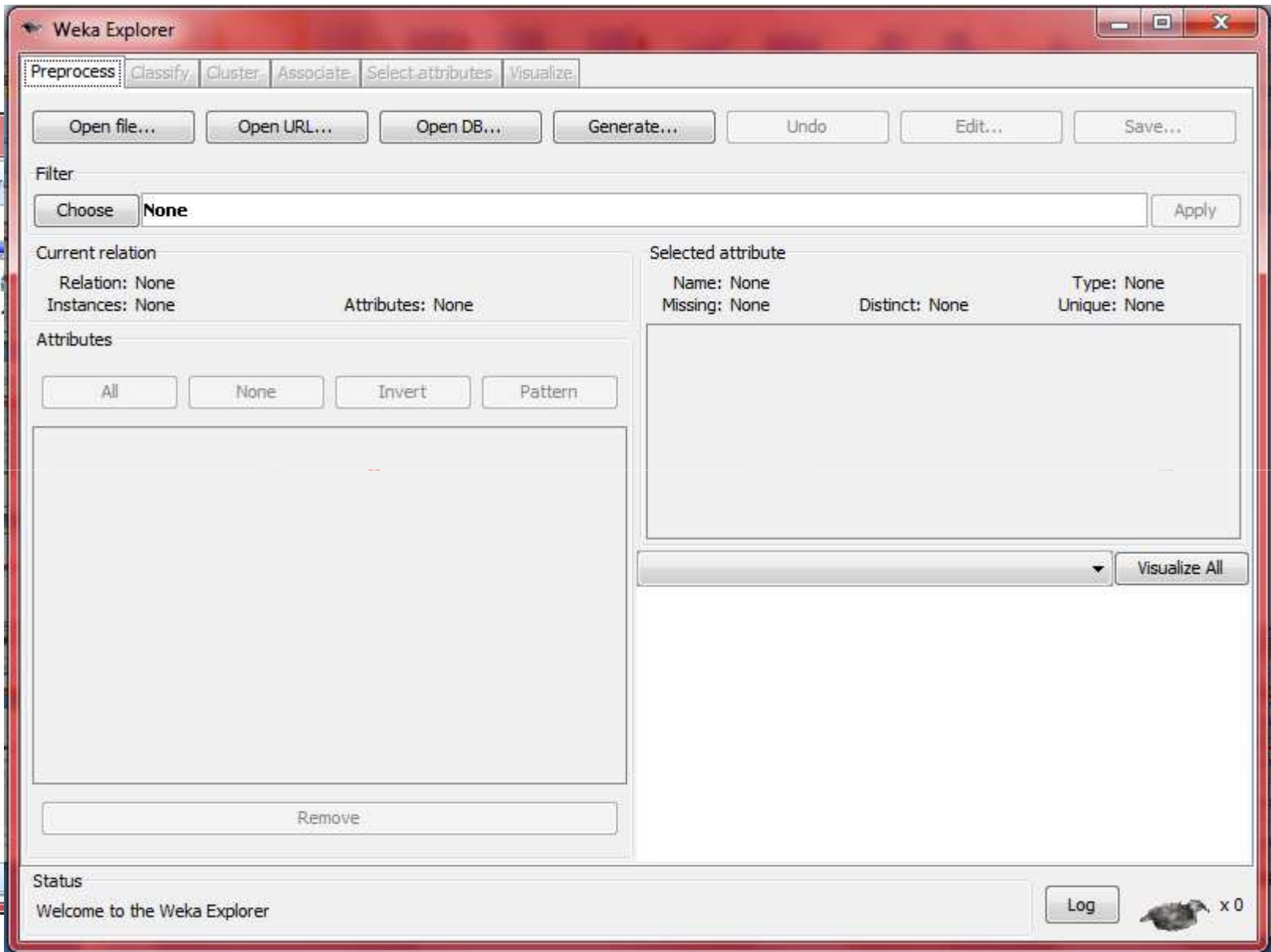
- Vous avez choisi d'ouvrir
- weka-3-6-3jre.exe**
- qui est un fichier de type : Binary File
- à partir de : http://iweb.dl.sourceforge.net
- Que doit faire Firefox avec ce fichier ?
- Orbit Downloader
- Enregistrer le fichier**
- Toujours effectuer cette action pour ce type de fichier.
- Buttons: **Enregistrer le fichier** and **Annuler**

Before Running WEKA

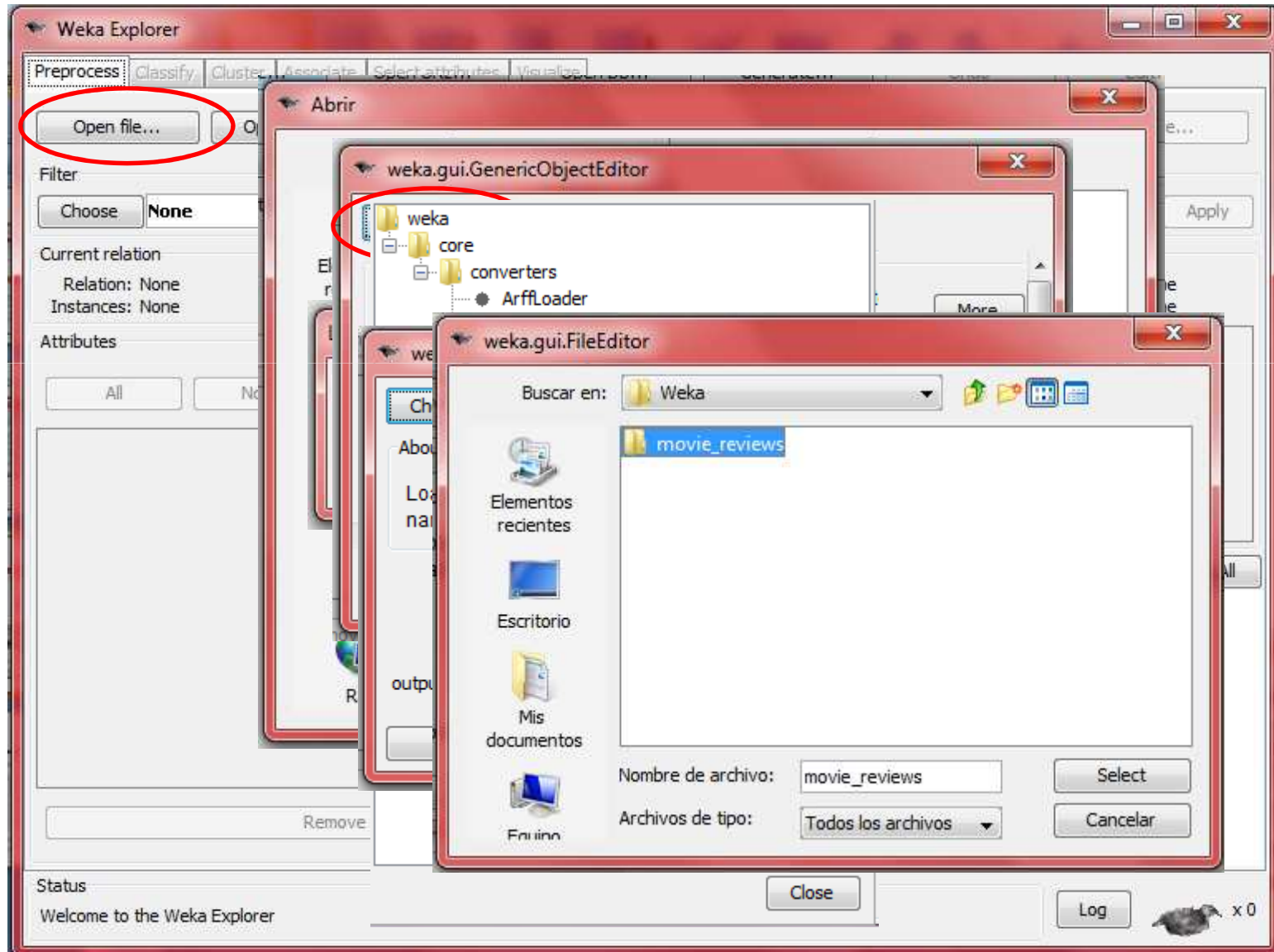
Increasing available memory for Java in RunWeka.ini

The image shows a Windows file explorer window with the 'RunWeka' folder selected. A Notepad window titled 'RunWeka: Bloc de notas' is open, displaying the contents of the 'RunWeka.ini' file. The file contains configuration for running WEKA, including a 'maxheap' setting. A yellow callout bubble with a red border points to the 'maxheap=256m' line, with the text 'Change maxheap=256m to maxheap=1024m' written inside it.

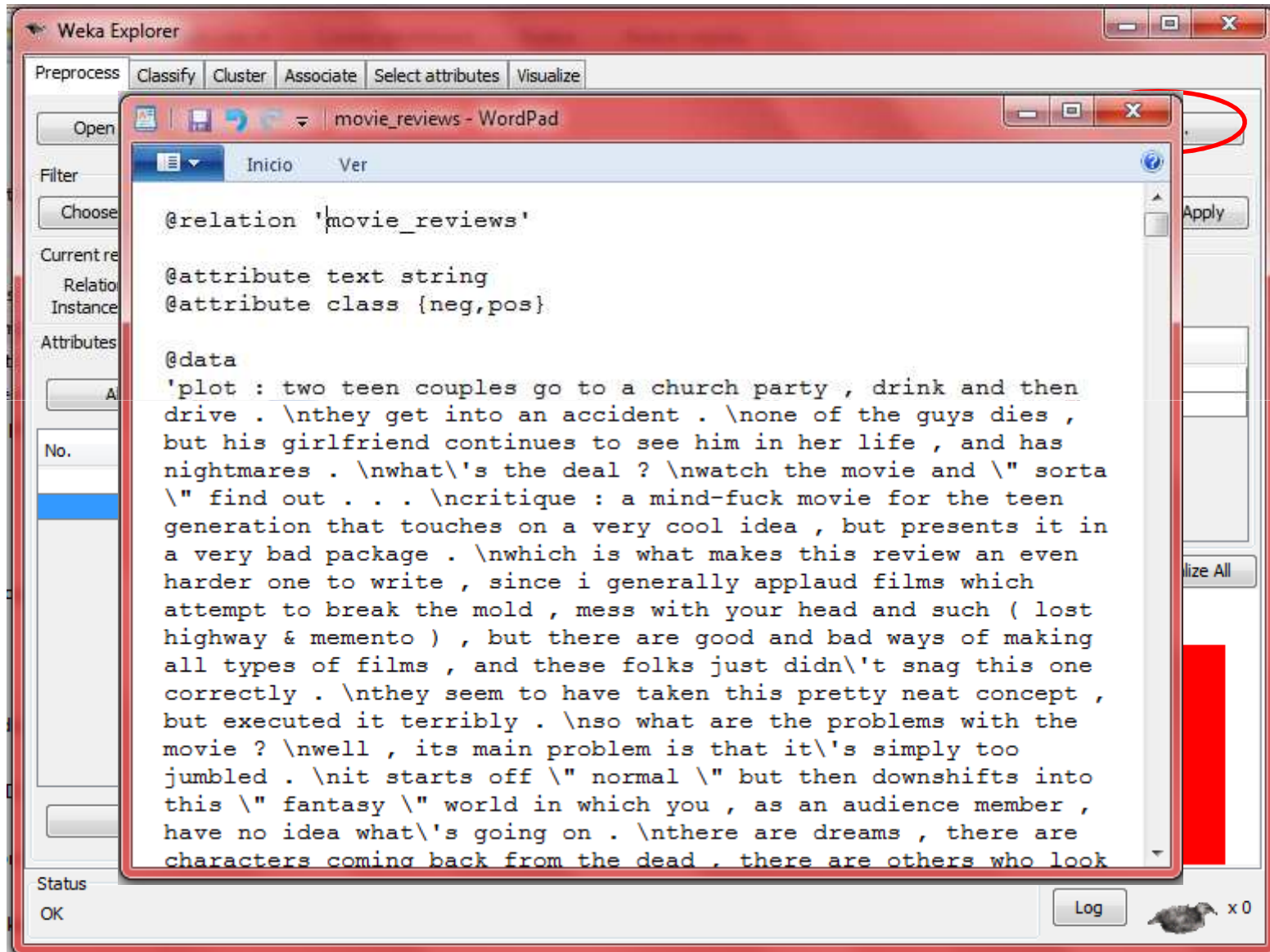
```
# Contains the commands for running weka either with a con
# ("cmd_console") or without the command prompt ("cmd_defa
# One can also define custom commands, which can be used v
# launcher "RunWeka.class". E.g., to run the launcher with
# "custom1", you only need to specify a key "cmd_custom1"
# command specification.
#
# Notes:
# - This file is not a DOS ini file, but a Java properties
# - The settings listed here are key-value pairs, separate
#   key can only be listed ONCE.
#
# Author  FracPete (fracpete at waikato dot ac dot nz)
# Version $Revision: 1.3 $
#
# setups (prefixed with "cmd_")
cmd_default=javaw -Dfile.encoding=#fileEncoding# -Xmx#max#
cmd_console=cmd.exe /K start cmd.exe /K "java -Dfile.encod
cmd_explorer=javaw -Dfile.encoding=#fileEncoding# -Xmx#max
#
# placeholders ("#bla#" in command gets replaced with cont
# Note: "#wekajar#" gets replaced by the launcher class, s
# provided as parameter
maxheap=256m
```



Creating a .arff dataset



Saving the .arff dataset



The screenshot shows the Weka Explorer interface with a WordPad window open. The WordPad window contains the following ARFF dataset definition and text:

```
@relation 'movie_reviews'

@attribute text string
@attribute class {neg,pos}

@data
'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 see him in her life , and has
nightmares . \nwhat\'s the deal ? \nwatch the movie and \" sorta
\" find out . . . \ncritique : a mind-fuck movie for the teen
generation that touches on a very cool idea , but presents it in
a very bad package . \nwhich is what makes this review an even
harder one to write , since i generally applaud films which
attempt to break the mold , mess with your head and such ( lost
highway & memento ) , but there are good and bad ways of making
all types of films , and these folks just didn\'t snag this one
correctly . \nthey seem to have taken this pretty neat concept ,
but executed it terribly . \nso what are the problems with the
movie ? \nwell , its main problem is that it\'s simply too
jumbled . \nit starts off \" normal \" but then downshifts into
this \" fantasy \" world in which you , as an audience member ,
have no idea what\'s going on . \nthere are dreams , there are
characters coming back from the dead , there are others who look
```

The WordPad window title is "movie_reviews - WordPad". The Weka Explorer interface shows the "Preprocess" tab selected, and the "Save" button in the top right corner of the WordPad window is circled in red. The status bar at the bottom of the Weka Explorer window shows "OK" and a "Log" button.

From text to vectors

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

review₁ = “great movie” review₃ = “worst film ever”

review₂ = “excellent film” review₄ = “sucks”

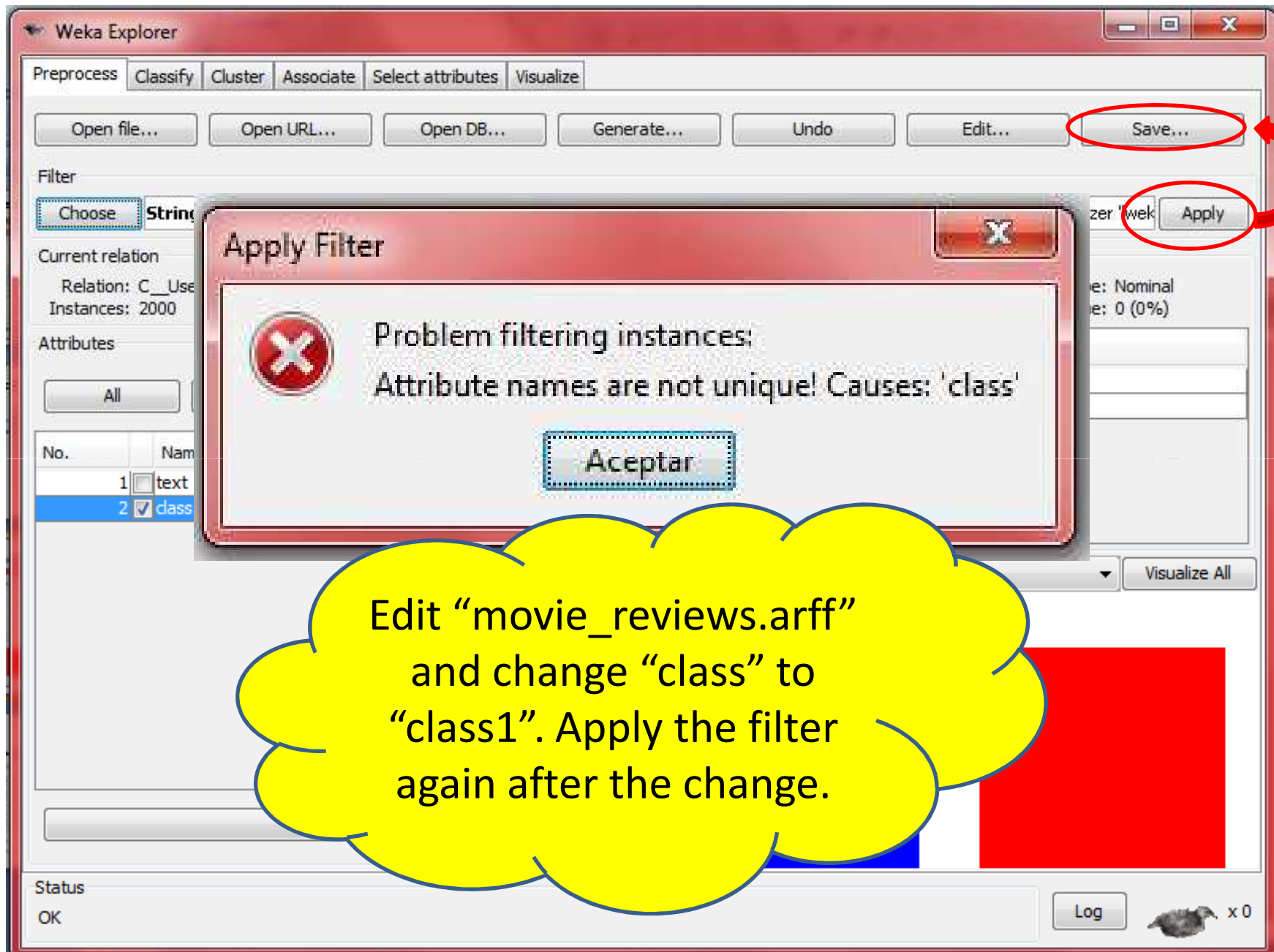
ever
excellent
film
great
movie
sucks
worst

$$V_1 = [0, 0, 0, 1, 1, 0, 0, +]$$


$$V_2 = [0, 1, 1, 0, 0, 0, 0, +]$$

$$V_3 = [1, 0, 1, 0, 0, 0, 1, -]$$

$$V_4 = [0, 0, 0, 0, 0, 1, 0, -]$$



Apply Filter

 Problem filtering instances:
Attribute names are not unique! Causes: 'class'

Acceptar

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

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | **Save...**

Filter

Choose | String

Current relation
Relation: C_Use
Instances: 2000

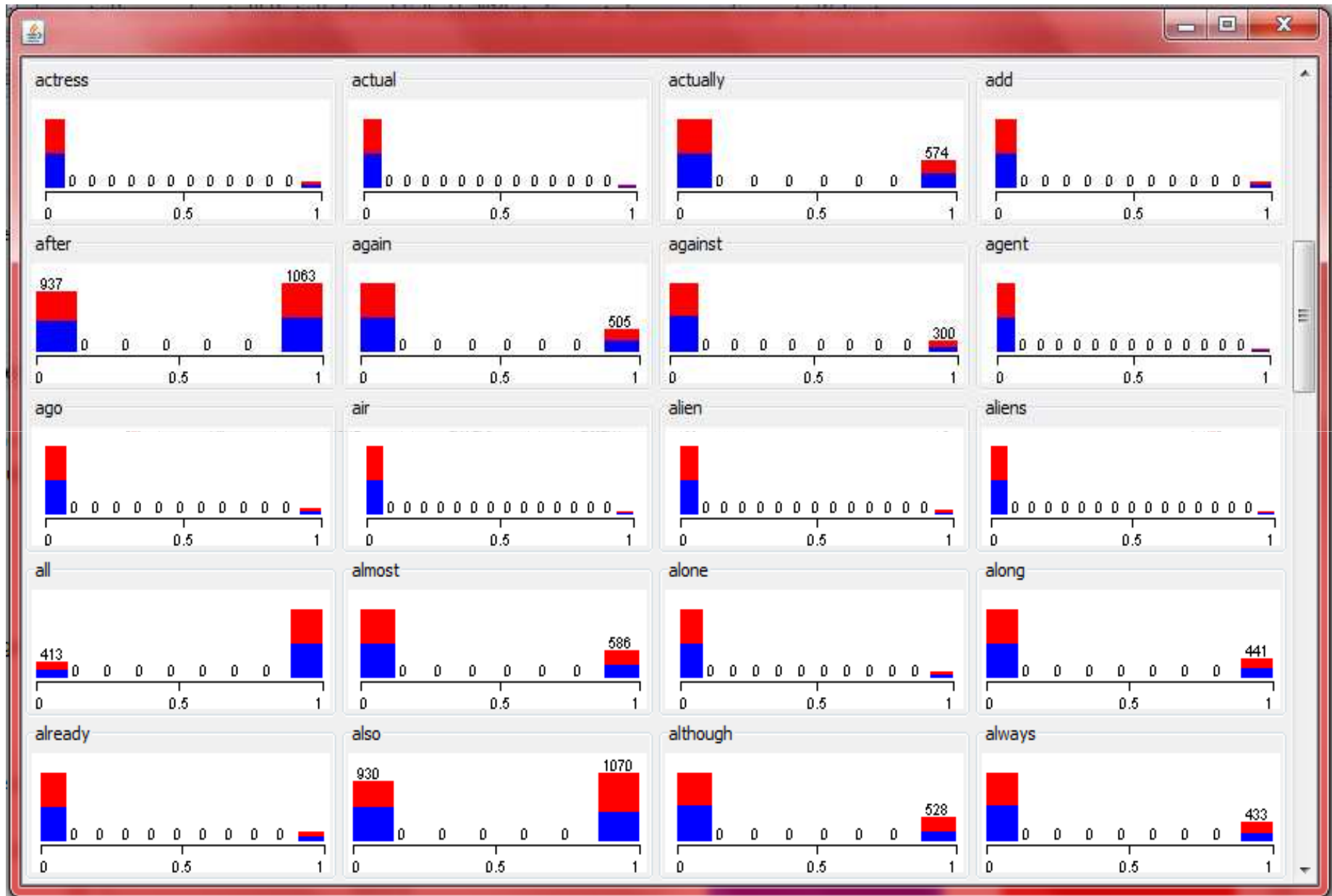
Attributes
All

No.	Name
1	<input type="checkbox"/> text
2	<input checked="" type="checkbox"/> class

Apply

Status
OK

Log  x 0



StringToWordVector

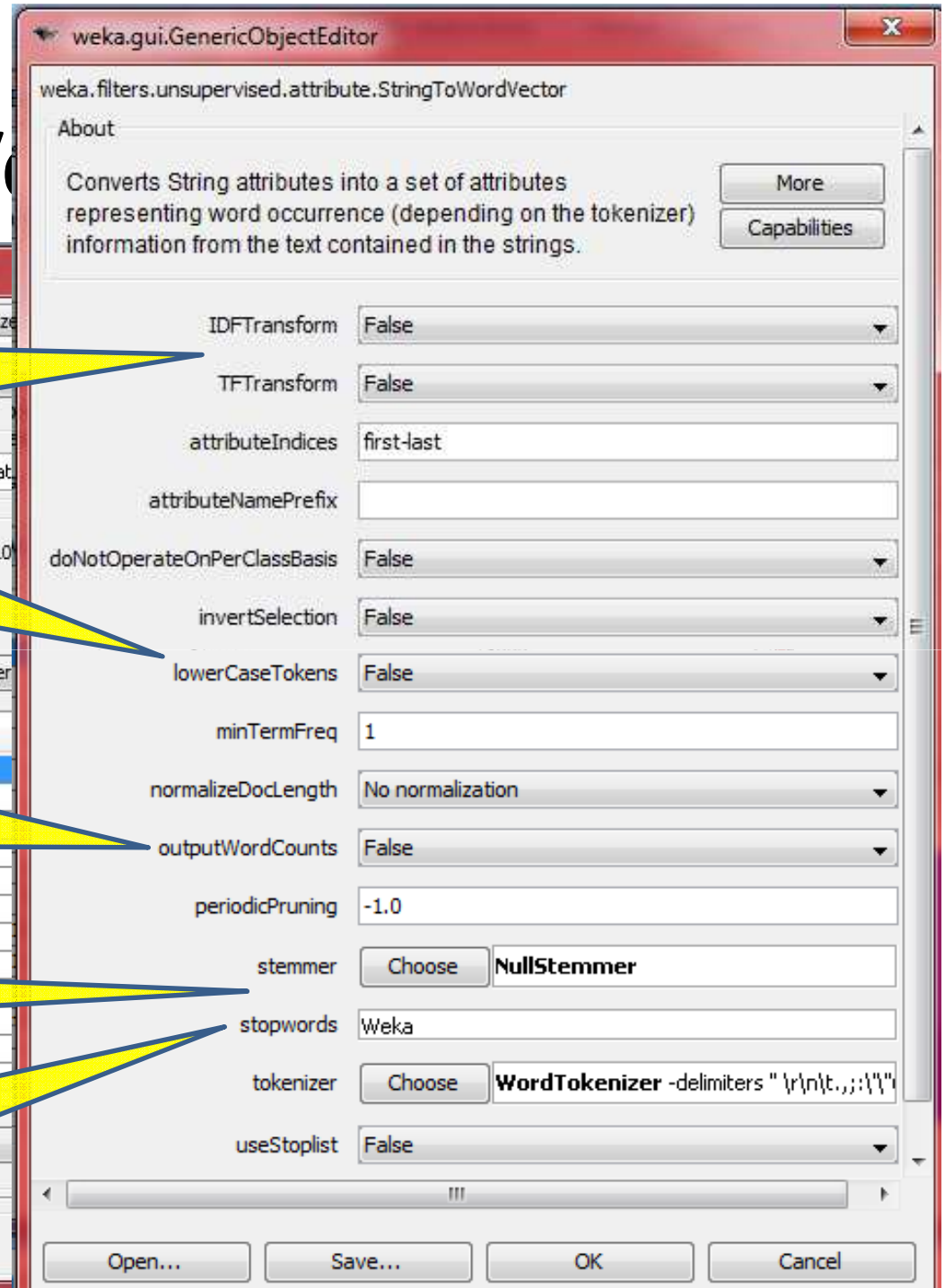
TF-IDF weigthing

lowerCase conversion

Use frequencies instead of single presence

Stemming

Stopwords removal using a list of words in a file



Generating datasets for experiments

<i>dataset file name</i>	<i>Stopwords</i>	<i>Stemming</i>	<i>Presence or freq.</i>
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
- Cross-validation Folds
- Percentage split %

(Nom) class1


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...

 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

<i>dataset name</i>	<i>Stopwords</i>	<i>Stemming</i>	<i>Presence or freq.</i>	<i>Naive Bayes 3-fold</i>	<i>NaiveBayes Multinomial 3-fold</i>
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) Selection

The screenshot displays the Weka Explorer application window. The 'Attribute Evaluator' is set to 'CfsSubsetEval' and the 'Search Method' is 'BestFirst -D 1 -N 5'. Under 'Attribute Selection Mode', 'Use full training set' is selected. The 'Folds' are set to 10 and the 'Seed' is 1. The dataset is '(Nom) class1'. The 'Attribute selection output' pane shows the following information:

```
=== Run information ===  
  
Evaluator:   weka.attributeSelection.CfsSubsetEval  
Search:     weka.attributeSelection.BestFirst -D 1 -N 5  
Relation:   C:_Users_Ser_Documents_PhD_Visita_prof_Gelbukh_2010_Curso_Expo  
Instances:  2000  
Attributes: 1166  
            [list of attributes omitted]  
Evaluation mode: evaluate on all training data
```

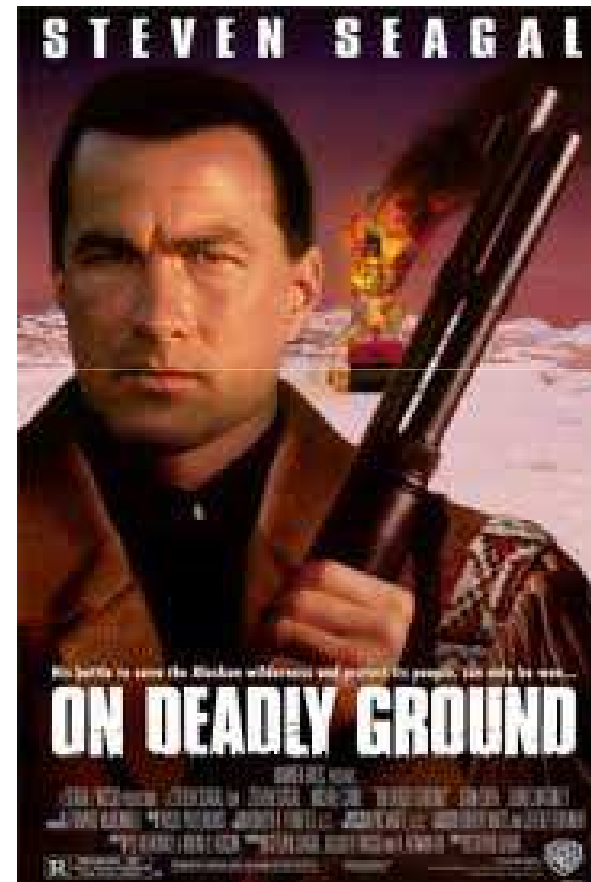
The 'Result list' shows a single entry: '20:44:28 - BestFirst + CfsSubsetEval'. The status bar at the bottom indicates 'Evaluating on training data...' and includes a 'Log' button and a small bird icon with 'x 1' next to it.

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

The screenshot shows the Weka Explorer interface for the 'Pruned movie_reviews_1.arff' dataset. The 'Current relation' is 'C:_Users_Ser_Documents_PhD_Visita_prof_Gelbukh_2010_C...', with 2000 instances and 51 attributes. The 'Attributes' list includes 'class1', 'also', 'awful', 'bad', 'boring', 'both', 'dull', 'fails', 'great', 'joke', 'lame', 'life', and 'many'. The 'Selected attribute' summary shows 'class1' is a Nominal type with 0 missing values, 2 distinct values, and 0 unique values. The 'Visualize All' button is active, and the visualization shows two bars: a blue bar for 'neg' (1000 instances) and a red bar for 'pos' (1000 instances).

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: C:_Users_Ser_Documents_PhD_Visita_prof_Gelbukh_2010_C...
Instances: 2000 | Attributes: 51

Attributes: All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> class1
2	<input type="checkbox"/> also
3	<input type="checkbox"/> awful
4	<input type="checkbox"/> bad
5	<input type="checkbox"/> boring
6	<input type="checkbox"/> both
7	<input type="checkbox"/> dull
8	<input type="checkbox"/> fails
9	<input type="checkbox"/> great
10	<input type="checkbox"/> joke
11	<input type="checkbox"/> lame
12	<input type="checkbox"/> life
13	<input type="checkbox"/> many

Remove

Selected attribute
Name: class1 | Type: Nominal
Missing: 0 (0%) | Distinct: 2 | Unique: 0 (0%)

No.	Label	Count
1	neg	1000
2	pos	1000

Class: class1 (Nom) [Visualize All]

1000 [Blue Bar] | 1000 [Red Bar]

Status: OK | Log [Turtle icon] x 0

Naïve Bayes with the pruned dataset

The screenshot shows the Weka Explorer interface with the Naive Bayes classifier selected. The 'Test options' section is set to 'Cross-validation' with 3 folds. The classifier output is displayed in a text area, showing a summary of performance metrics and a detailed accuracy by class table.

Classifier output Summary

Correctly Classified Instances	1621	81.05 %
Incorrectly Classified Instances	379	18.95 %
Kappa statistic	0.621	
Mean absolute error	0.2157	
Root mean squared error	0.3814	81.05 %
Relative absolute error	43.1437 %	
Root relative squared error	76.276 %	18.95 %
Total Number of Instances	2000	

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Weighted Avg.	0.804	0.183	0.815	0.804	0.809	0.888
	0.817	0.196	0.807	0.817	0.812	0.888
	0.811	0.19	0.811	0.811	0.81	0.888

Confusion Matrix

```
a  b  <-- classified as
804 196 | a = neg
183 817 | b = pos
```

Result list (right-click for options)

- 20:23:48 - bayes.NaiveBayes
- 21:05:56 - bayes.NaiveBayes
- 21:08:11 - bayes.NaiveBayesMultinomial
- 21:10:00 - bayes.NaiveBayes

Status: OK

Clustering

Weka Explorer

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

Clusterer

Choose EM -I 100 -N 2 -M 1.0E-6 -S 100

Cluster mode

Use training set

Supplied test set: Set..

Percentage split: % 66

Classes to clusters evaluation

(Nom) class1

Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

21:15:57 - EM

21:16:19 - EM

Clusterer output

Clustered Instances

0	1481 (74%)
1	519 (26%)

Log likelihood: 0.15075

Class attribute: class1

Classes to Clusters:

0	1	<-- assigned to cluster
893	107	neg
588	412	pos

Incorrectly clustered instances : 695.0 34.75 %

Correctly clustered instances: 65.25%

Status OK Log x 0

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