Machine Learning for Language Technology
(Autumn 2016)

Assignment 2: Naïve Bayes & Sentiment Classification

CHANGELOG: 4 NOV 2016, 30 NOV, 12 DEC 2016

Individual Home Assignment (Undergraduate Students in Language Technology)
SUBMISSION DEADLINE: SUNDAY 18 DECEMBER 2016, 23:59
Make sure that your answers are thorough and comprehensive.

Assignments' Deadlines
18 Dec 2016: Ass1 and Ass2
15 Jan 2017: Ass 3
24 Feb 2017: Final submission date for all assignments.

Acknowledgements: this assignment is the result of the adaptation of several sources, namely home assignments, exercises and tutorials on Weka that were retrieved from the web in 2016.

Learning objectives

In this assignment you are going to:

- use the Naïve Bayes algorithm as implemented in Weka;
- explore how attributes/features affect classification results.
- work with tweets;
- classify sentiment polarity (positive, negative);

Data: The Edinburgh Twitter corpus

In this assignment we are going to use a dataset of tweets. The dataset is extracted from the so-called "Edinburgh Twitter corpus". The corpus is described in Petrovic et al., 2010[1]). Tweets are widely used for sentiment analysis, as pointed out by the corpus creators: "The microblogging service Twitter has become a popular tool for expressing opinions, broadcasting news, and simply communicating with friends. People often comment on events in real time, with several hundred micro-blogs (tweets) posted each second for significant events." (Petrovic et al., 2010).

In this assignment we are going to use a small dataset of tweets that has been created for educational purposes from a larger dataset in ARFF format[2]. The "educational" dataset contains 100 positive tweets and 100 negative tweets.
The dataset includes two attributes. One attribute is the class. The class is nominal and can be either positive (pos) or negative (neg). The attribute "tweet_body" is to type "string" and contains the text of each tweet. See Fig 1.

---

[1] Sasa Petrovic, Miles Osborne and Victor Lavrenko. *The Edinburgh Twitter Corpus. Computational Linguistics in a World of Social Media* (Workshop at NAACL), Los Angeles, USA. June 2010. (In August 2010, Twitter have asked us to stop distributing this data. However, tweets dataset in ARFF format is still available).

[2] The original version of the tweets dataset in ARFF format is available here:
< https://drive.google.com/file/d/0B1pvkpCwTsiSd1pyTFZkdWVRdEs5Q1NiQW1mRmF1Zw/view >.

```
@relation tweets

@attribute tweet_body string
@attribute sentiment {pos,neg}

@data
'anyone feel motivated the fri afternoon prior to a holiday? wanted to get lots do
&lt;3 her ',pos
'seriously, do you have to rub it in maggie!!!!  ',pos
'if i\'m not wrong.. Alt is when image can\'t be displayed.. Tooltip is the \'titl
'I don\'t like social karma  much. Would rather skip it, but can\'t afford to piss
'I\'d be happy to review the Iomega if EMC send me one!!! ',pos
'Something  I have wanted to make for a while now... finally done   URL',pos
'you\'re so sweet!  proving me right again...  the Dutch are the Best!   ',pos
'Look for us on the back of your Pepsi can! Our Pepsi can offer hit\nstores today!
'well rob i have to admit that you have to admit that you feel cool for being on t
'big skype call at 1010! msg me if you want in ',pos
```

Fig 1. Screenshot of the tweets dataset in ARFF format.

Download the educational tweet dataset from here:
< http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/100pos_100neg_tweets.arff >

Tip

You might find convenient to summarize the performance results on this dataset in a summary table, as in the example shown below:

Comparison of results

#	Stemming	Stop-words removal	Attribute selection	Naive Bayes	Nearest neighbor	3-NN	5-NN
1	No	Yes	No	**81,1%**	51,7 %	51,9%	54,6 %
2	No	Yes	Yes	**79,5%**	66,3%	70,8%	70,55%
3	Yes	Yes	Yes	**80,45%**	65,4%	**73,95%**	73,9%
4	No	No	Yes	**81,45%**	65,35%	70,9%	72,35%

Source: http://www.stefanoscerra.it/movie-reviews-classification-weka-data-mining/

G tasks

Task 1: Theoretical question

Q1: What is Naive Bayes inductive bias?

Task 2:

Start Weka and choose the Explorer interface. Open the dataset in Weka (Preprocess --> Open File). You should see the screen in Fig 2 when the dataset is correctly uploaded.

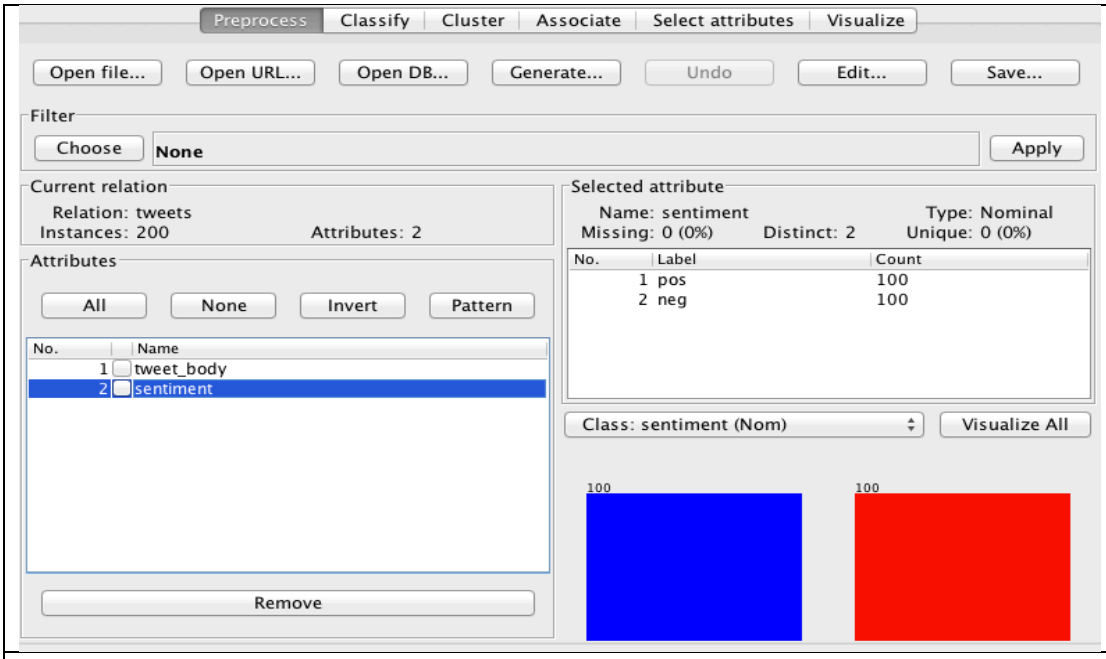

Fig. 2. Uploaded dataset and distribution of the ”sentiment” class.

Go to the Classify tab. Click the Choose button and then select Meta-->FilteredClassifier. Click the FilteredClassifier name (see Fig. 3) and a window will appear.

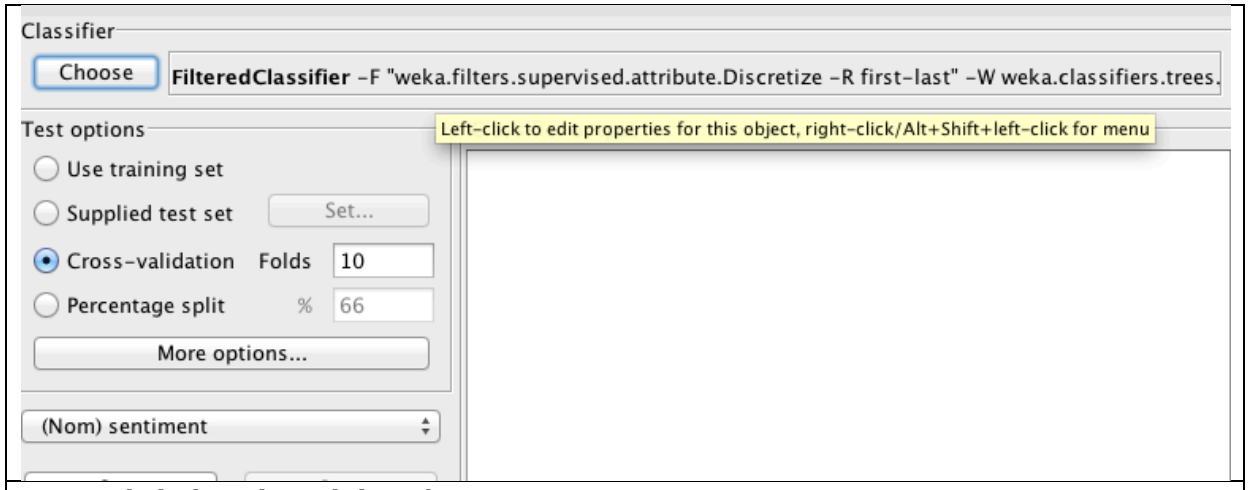

Fig. 3. Click the FilteredClassifier name.

In that window, you can choose the classifier name and the filter. Choose Naive Bayes (classifier→bayes→naive bayes) as classifier, and StringToWordVector as a filter (filter→unsupervised→StringToWordVector) See Fig. 4.). The filter **StringToWordVector** converts the strings (ie the tweets' body) to vectors of words.

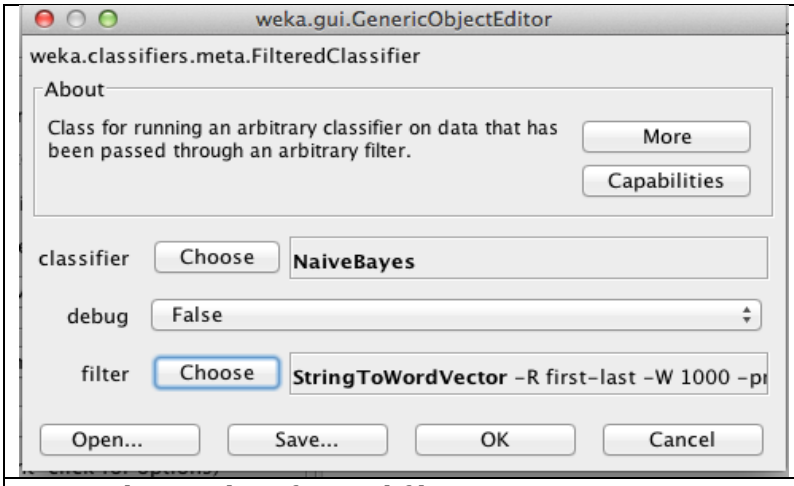

Fig. 4. Choose classifier and filter.

Click OK. You will be presented the screen in Fig 5. Click Start. Read carefully the classifier output (RHS) and answer the questions below.

Q2: The StringToWordVector filter converts the strings into numeric attributes (@attribute) How many numeric attributes do you count in the classifier aoutput?
Q3: Is the class attribute (ie ”sentiment” of each tweet) affected by the filter?
Q4: Report accuracy, TP Rate FP Rate Precision Recall F-Measure and examine the confusion matrix. All in all, how does the classifier perform? Are you happy with the performance? Why?

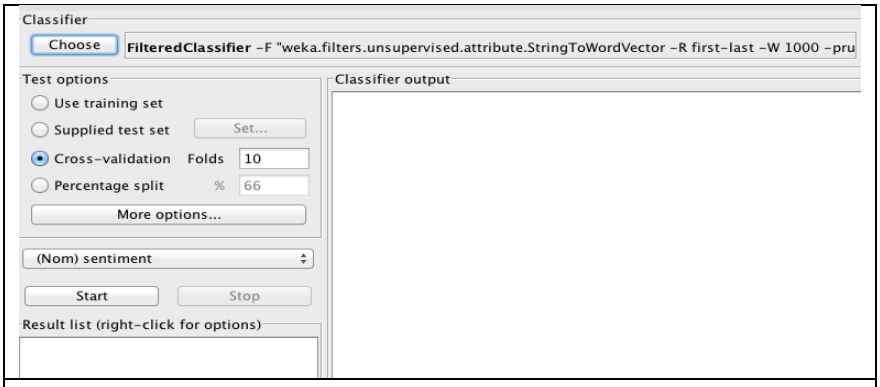

Fig. 5. Classification ready to start.

Click StringToWordVector. A window containing several options. (see Fig 6) These options are parameters that affect the behaviour of the filter and the classifier as a whole will appear. Click More and read the behaviour of the parameters. When you have read all the paramenters, focus your attention on the following parameter: **minTermFreq**. Modify the value of this parameter. First set it to 5. Run the classifier again, analyse the output and report accuracy, TP Rate FP Rate Precision Recall F-Measure and examine the confusion matrix.
Q5: How does the classifier perform?

Then set the parameter 10. Run the classifier again, analyse the output and report accuracy, TP Rate FP Rate Precision Recall F-Measure and examine the confusion matrix.
Q6: How does the classifier perform? Can you explain the behavior of this parameter through the way it affects the performance?

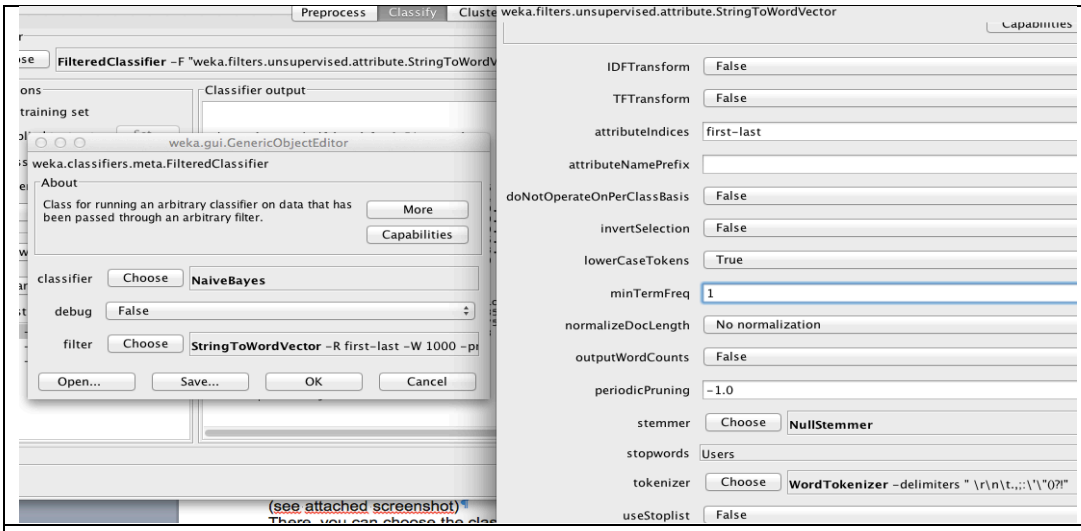

Fig. 6. StringToWordVector: parameters' window.

VG tasks

Task 3: Theoretical question

Q7: What is Bayes theorem (aka Bayes Law)? How does Bayes theorem work?

Task 4:

Reset the MinTermFreq to 1. Download the stopwords file on to your computer from here: < http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/stopwords_eng.txt >
Modify the following paramenter: **useStoplist**. Set useStoplist. to True and specify the following file in the **stopwords** field: *stopwords_eng.txt*. Read carefully the classifier output: How many numeric attributes do you count in the classifier output?
Report accuracy, TP Rate FP Rate Precision Recall F-Measure and examine the confusion matrix.
Q8; How does the classifier perform in comparison to the performance in Part A?
How would you increase the impact of the stopword list on classification? Make some suggestions (eg. adding more tweets-related word to the stopwordlist file, or reducing the number of words in the stopword file, removing/adding/processing negations, etc).

Try and use one parameter of your choice from the filter's parameters window. Choose a parameter that we have not used before. Describe the parameter and motivate your choice. Report accuracy, TP Rate FP Rate Precision Recall F-Measure and examine the confusion matrix.
Q9: How does the classifier perform with the parameter setting you have decided? Compare with the previous runs.

To be submitted

A written report (at least 2 pages) containing the **reasoned** answers to the tasks and questions above and a short section, that you can call "*Conclusions*", where you summarize your experience and your reflections.
Warning: Cutting and pasting Weka's results page into the report without commenting or explaining the whys and wherefores is not enough to get a pass on the assignment.

Submit the report in **PDF** format to santinim@stp.lingfil.uu.se no later than **18 December 2016, 23:59**. **Please, write this phrase in the subject line of your email: "ML4LT 2016 – Ass2: *your name*".** Attach any additional material that you think is important to fully understand your report.

Naming conventions

Please, name your pdf report in this way (it will be easier for me to organize and archive them): surname_name_*ass2report*.pdf (ex: *santini_marina_ ass2report.pdf*).

--the-end--