Lab report Sentiment analysis in textual movie reviews

Author: François Amat

Contact : francois.amat@telecom-paristech.fr

Abstract

In this lab we want to dertermine if a textual movie review is positive or not.
I will implement a naive bayesien classifier, test it and compare it with the classifiers in the scikitLearn library
In order to test these classifiers, i will use the same data with differents alterations seen in the course (Pos tagging, remove stop word, stemmer) To train the classifiers.

Implementation of the classifier

Explain how positive and negative classes have been assigned to movie reviews

The positive and negative classes have been assigned from scraping webpages. Only the explicit rating have been kept.
The positive and negative classes are determined from the rating system that the website have used.
If the score of the movie is above 80%, 3.5/5, 3/4 or B or plus , the review is considered as positive.
If the score of the movie is below 2/5, 1.5/4 or c- or below, the review is considered as negative.

Evaluate the performance of your classifier in cross-validation 5-folds.

Using a cross-validaiton 5-folds I obtain a score of `0.52`, which is very low in comparison of a classifier that give a random answer.

Change the count_words function to ignore the “stop words” in the file data/english.stop. Are the performances improved ?

By filtering the stop words contained in the file given, we got a slightly improvement : `0.5275` against `0.52` before.

Thoughts about the implemented classifier

I was very disapointed to get only a classifier with a 0.52 score,
I think that the prior does not really help with a dataSet with half positive and half negative reviews.
I suspect a problem in the Laplace smoothing but i did not find an error.
But I think that the fact that my score improved with the stop words filter is encouraging.

Scikit-learn use

1. Compare your implementation with scikitLearn
   Scikitlearn get better results in a fraction of the time taken with my algorithm.
   The score is `0.8` and the computation time is around `3 minutes` against a score of `0.52` and a computation time of `15 minutes`.

2. I obtain these scores with differents classifiers:

Logistic R	SVC	NB
0.8675	0.85	0.825

3. I obtain theses scores by using a stemmer from the nltk library:

Logistic R	SVC	NB
0.8625	0.85	0.8325

4. I obtain theses scores by using a POS tagger from the nltk library:

Logistic R	SVC	NB

0.805	0.7975	0.7825

Thoughts On the scikitLearn implementation

As expected, The algorithms from scikitLearn perform better and faster,
The score is improving after using a stremmer. But using a Pos tagger and limit the text to the verbs, adverbs, nouns and adjective have reduce the score of all classifiers.