Text Mining. Session 3

In this home-exercise you will be learning how to build a sentiment analysis classifier. Here you have an orientation of how to proceed:

1. Download and unzip the file review_polarity.tar about film reviews and if they are positive or negative.
2. Find a way to import all the sub-files in a pandas dataframe, with the corresponding label
3. Split the data into a training and test
4. Clean and preprocess the data using some of the tools that we saw on Session 2 that you consider they may help to increase performance. Lemmatization, Stemming, Stopwords, N-grams (how to? Maybe not yet?), etc.
5. Once you have a tokenized data structure for each of the texts, explore what the function CountVectorizer does, and why do we need it. The final object of this step is called bag of words, and you can choose from a wide variety of parameters that help you adapting your process to specific scenarios.
6. Explore the concepts term frequency and inverse term frequency, and convert the bag of words into one of these. Sklearn has it implemented.
7. Use and tune one or more models to predict the sentiment of each of the texts in the testing set, and show measures like the accuracy.
8. (optional) Plot a (nice) confusion matrix
9. (optional) Plot a word cloud for each of the cells of the matrix, where you can check how reasonable it is to “fail” trying to classify those items.