Homework #4

Sentiment 2

Please present a report with all your answers, explanations, and sample images or plots. Submit also a soft copy of the source code and binaries used to generate these results. Please note that copying of any results or source code will result in ZERO credit for the whole homework.

In this homework you will be trying several ideas to improve upon the performance of your sentiment analysis from the previous homework. You should explore these ideas and comment on your experiments.

Ideas

- Using a Boolean vs normal Naive Bayes i.e. using counts for words in a document or just the occurrences.
- Using an SVM for classification instead of Naive Bayes. You can use the SVM implementation of the package **`sklearn`** included with Python(x,y).
- Implementing simple negation features. For example, detect some negation words (use Not, n't, and never) and add NOT_ to each word until the next punctuation (e.g. period, comma, question mark, … etc.). You will need to write regular expressions to do that.
- [BONUS] Use a sentiment lexicon instead of using all words in the documents (e.g. the MPQA lexicon). You can try using the lexicon in two ways:
  - Use it as the only vocabulary.
  - Add features for the presence of these words i.e. add it to the existing vocabulary as a Boolean feature.

Requirements

You are required to explore the ideas above. Explore these ideas independently using the best settings from homework 3. For each idea, implement it and comment on whether it improves the performance of homework 3, and by how much.

Please submit your code and report in one zip file, named **CMP462.HW04.bench#.firstname.lastname.zip**. For example, if your name is Mohamed Aly and your bench number is 26, your file should be named **CMP462.HW04.26.Mohamed.Aly.zip**.

Grading

2 pts: implementation of Boolean Naive Bayes
3 pts: implementation of SVM classifier

3 pts: implementation of negation features
2 pts: implementation of sentiment lexicon
2 pts: report and submission file name
2 pts: bonus