Homework 3

Warmup:

Given the following short movie reviews, each labeled with a genre, either comedy or action:

1. fun, couple, love, love **comedy**
2. fast, furious, shoot **action**
3. couple, fly, fast, fun, fun **comedy**
4. furious, shoot, shoot, fun **action**
5. fly, fast, shoot, love **action**

and a new document D:

fast, couple, shoot, fly

compute the most likely class for D. Assume a naive Bayes classifier and use add-1 smoothing for the likelihoods.

Assignment:

Build a naive Bayes sentiment classifier that will assign reviews of an application as either **positive**, **neutral**, or **negative**.

- You will need to do some basic preprocessing on the documents (normalization, etc).
- Do not use a stop word list.
- Ignore any Out-Of-Vocabulary (OOV) terms when classfying

You are provided a small set of pre-classified training data to build your model. The data is formatted such that each line of text contains a document (the title of a review). The first token of each line will be the classification of that review, either POS, NEU, or NEG. Below is a sample document:

```
POS The program was quite helpful with creating websites.
```

An example output of your system may look something like this:

```
The program does what it should do. : POSITIVE
It functions adequately. : NEUTRAL
The program sucks. : NEGATIVE
This thing runs like a pregnant cow. : NEGATIVE
It was a little slow, but not too bad. : NEUTRAL
Slow. Slow. SLOW! : NEGATIVE
Great software! : POSITIVE
Worth the trouble to install. : NEUTRAL
```

Once the model has been built, feed in the provided test documents and write a report detailing your results. In the report, address the following:

- How accurate was the classifier? What was the Precision and Recall? The F-measure?
- Choose one incorrectly classified document.
  - Manually calculate the sentiment probabilities for the document (you can use your classifier to generate the likelihoods and prior probabilities, but do the classifying on paper)
  - What is the difference of the probability sums of the correct class and the class assigned by the system?
  - Identify the term or terms that caused the system to misclassify the document.
  - Build a document (or documents) to add to the training set that would allow the system to correctly classify the document.
    - Show the mathematical reasoning for your choice of words in the document.
    - Rerun the tests with the additional information.
    - Did adding the additional information change any other document classification? If so, how? Did it improve the overall accuracy of your system or make it worse?
  - Add the MPQA Subjectivity Cues Lexicon to your system and run the tests again and report the results.
    - Choose a document that was classified differently after adding the lexicon. Was it correctly or incorrectly classified? Discuss why.
  - Finally, use the provided collection of Amazon reviews from 2007 to train your classifier. Run the associated tests and report the Precision, Recall, and F-measure.
  - Briefly discuss what you learned from this assignment, what you liked or disliked about the assignment and, optionally, anything you would like to see changed or added to improve the assignment.