Homework 2

For this assignment, you will implement your own version of a Naive Bayes Classifier. In case you need a review of this classification method, please read one (or both) of the following documents:

- https://web.stanford.edu/~jurafsky/slp3/
- http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

For the dataset, you can either continue using the movie review data from the previous homework (see if you can build a better model now!):

https://github.com/dennybritz/cnn-text-classification-tf/tree/master/data/rt-polaritydata

or pick any other document classification datasets. Here are some standard datasets to use for this purpose (but you are welcome to pick something else):

- The Reuters dataset
- The 20 newsgroups dataset
- The IMDB movies dataset
- Hate speech data https://hatespeechdata.com

I am not providing the links for some of these datasets intentionally; research skills are extrimely valuable and now is the time to start developing them! Note that, if you pick a dataset that has more than two classes, it is acceptable for this assignment to find a way to convert it to a binary classifcation task.

Concretely, please complete the following steps:

1. Split your data into training (70%), development (15%), and test (15%) sets. (10 points).

2. Implement and train a Naive Bayes classifier on the training data and tune it on the development set. Report classifier performance. **Important:** The classifier has to be written from scratch rather than using an existing implementation. If you have trouble implementing the algorithm it is useful to debug it using the "chinese/japanese example" from the lecture slides. (50 points).

3. Train your best model on the concatenation of training and development sets and evaluate it on the test set. Report classifier performance. (20 points).

4. Find several examples from the test set where the classifier is very confident and very uncertain. Include these examples in your report. Can you explain why the classifier was confident or uncertain? (10 points).

5. Based on your best model, what were the most useful features for each class? Justify your answer. (10 points).

When you are working on (2), there are a few variables that you can tweak to improve the training speed and classification accuracy, e.g.:

- Consider discarding infrequent words (e.g. you might want to keep only 1000 most frequent words and/or remove all words that occur only in a single example).
- It may or may not be useful to remove certain stop words from the data, such as the ones listed here:

http://www.ranks.nl/stopwords

- You could also try training your classifier using only the words that appear on these lists:

http://ptrckprry.com/course/ssd/data/negative-words.txt http://ptrckprry.com/course/ssd/data/positive-words.txt

The ability to summarize your findings is extremely important when doing research or working as a data scientist. Please summarize your findings in a 1-2 page report. Include the details on what kind of pre-processing you performed, key implementation choices that you made, and comment on the questions asked in (1-5) above. This assignments will be evaluated primarily based on your report.