

CSE 5525 Homework 1: Text Classification

Alan Ritter

In this assignment you will implement the naïve bayes algorithm for sentiment classification. You will train your models on a (provided) dataset of positive and negative movie reviews and report prediction accuracy on a test set.

We provide you with starter Python code to help read in the data and evaluate the results of your model's predictions. You are *strongly* encouraged to make use of the provided code. If you prefer to implement everything from scratch, please talk to the instructor first. Your submitted code should run on the command line in a unix-like environment (e.g. Linux, OSX, Cygwin or Windows Supsystem for Linux).

Depending on the efficiency of your implementation the experiments required to complete the assignment may take some time to run, so it is a good idea to start early.

Naïve Bayes

First, implement a naïve bayes classifier. The provided code in `imdb.py` reads the data into a document-term matrix using scipy's `csr_matrix` format (See http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.sparse.csr_matrix.html#scipy.sparse.csr_matrix for details). We recommend working with log-probabilities using addition instead of directly multiplying probabilities to avoid the possibility of floating point underflow (see: https://en.wikipedia.org/wiki/List_of_logarithmic_identities).

You can run the sample code like so:

```
python NaiveBayes.py aclImdb_small 1.0
```

The two methods you will need to implement are `NaiveBayes.Train` and `NaiveBayes.Predict`. Before you do this, the classifier always predicts +1 (positive). Once you have implemented these methods, the code will print out

accuracy. Try running with different values of the smoothing hyperparameter (**ALPHA**) (suggested values to try: 0.1, 0.5, 1.0, 10.0), and record the results for your report.

What to Turn In

Please turn in the following to the dropbox on Carmen:

1. Your code
2. A brief writeup that includes the numbers / evaluation requested above (text file format is fine).