Homework 1

Author: Nicholas M. Synovic

Table of Contents

About

The homework assignment description can be found in hw1.pdf. The `hw1.py` script is the executable code to run to generate results.

Methodology

The code takes a 60-40 split of the data (60% training, 40% test) from both the positive and negative sentiment datasets. It then tokenizes each training data component seperately and returns the types. Stop words and punctuation is not removed as tests found that better performance could be found by keeping them included. Finally, the testing splits are evaluated against the negative **and** positive types.

The decision of whether a document is of positive or negative sentiment is determined by the number of types within the testing document that intersect with the types of the training type set. An example would be a testing document that has 8 words in the positive types, and 2 words in the negative types. Since there are more words in the positive types, the testing document is returned as positive. If there is a tie, it is considered to be the inverse of the dataset it originated from by defualt.

Results

Out of the 4,266 testing documents, 46.156% were classified accurately. I would not recommend this *rule based classifier* as a sentiment analysis tool.

The breakdown of the accuracy can be found in the table below.

True Positive	True Negative	False Positive	False Negative
45.429%	46.882%	54.571%	53.118%