CIS391 HW6 Write-Up

Shadia Al-Shafei and Josh Rojas

Features

1.) Bigrams

This is what we used for our final implementation that was submitted. It did not do much better than unigrams that was implemented in the previous homework.

2.) Word length

This did significantly worse than unigram probabilities

3.) All Caps

This did significantly worse than unigram probabilities

4.) Using the sender, calculate probabilities of the username, website and domain

This did significantly worse than unigram probabilities

5.) Multiple punctuation in a row (number of times multiple punctuation in a row appears in an email)

This did significantly worse than unigram probabilities

All features beside Bigrams did worse than Unigrams. We also tried combining features but that did not help. So we decided to just submit Bigrams. We also added features to the tokenizer.

Tokenizing

1.) Only extract the tokens that come after Subject

2.) change numbers to JSNUM6, JSNUM3, or just JSNUM

3.) change websites into the word website (if they start with http or www)

4.) if we see a dollar sign change to JSMONEY

5.) separate symbols from chars but not from other symbols