Homework 6 Report
Arjun Raj Jain

General Design – 99% Accuracy

Overall, the general design of the spam filter was very similar to the previous homework. The main changes were in the loading of the tokens and the calculation of the log probabilities. First of all, Instead of just considering all tokens delimited by white space, I used three lists of tokens: Headers, Unigrams, Email Addresses. I then respectively calculated the log probabilities of each word within each list of tokens. This made sure that the weight of special tokens such as emails, headers, were given extra weight. I made my code modular in way that now if I think of more tokenization features, it can be easily added to my program without significant changes.

Experimented Features

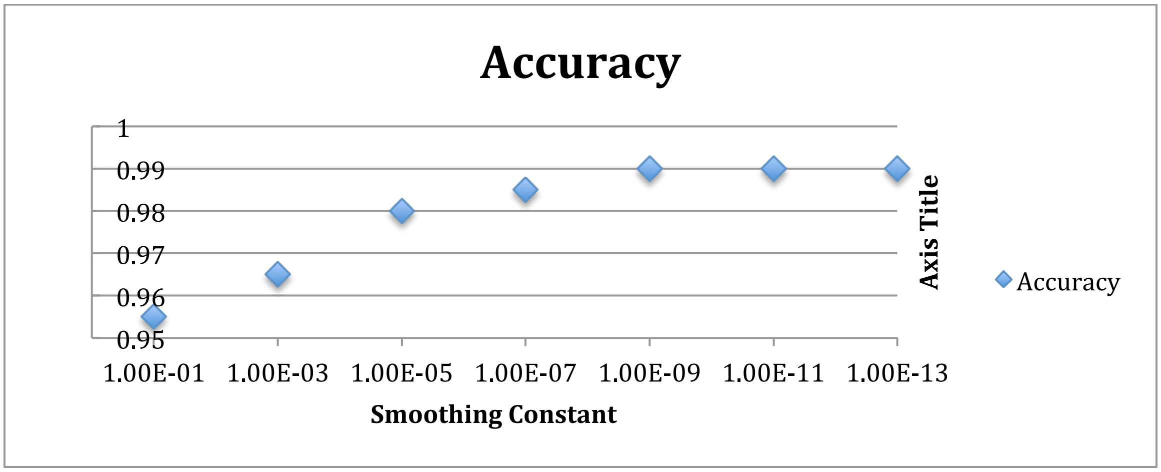
- Features related to number, punctuation counting
- Features related to length of token
- Features related to email headers
- Features derived from email addresses

Final Pre-processing + Tokenization Techniques

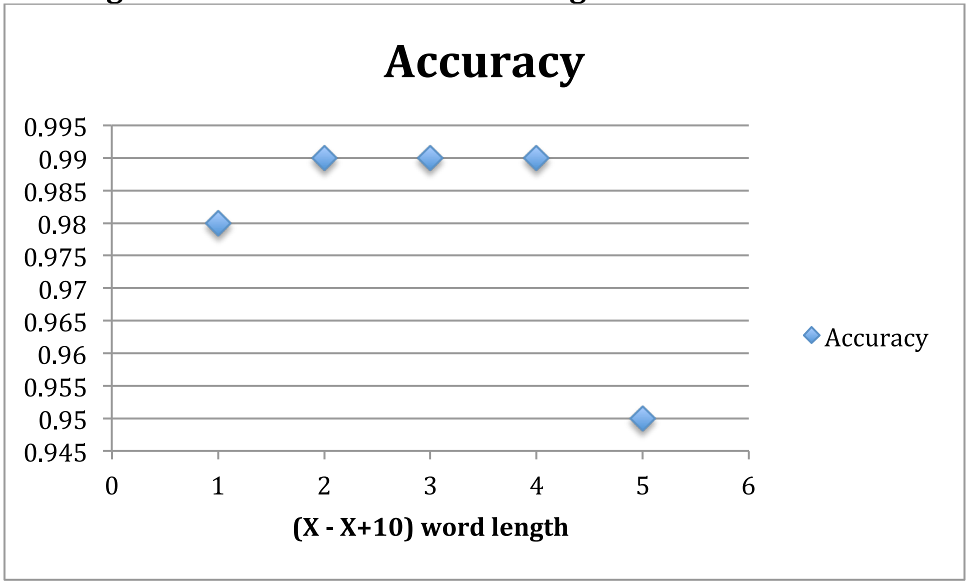
- Consider all content-types in the header
- Consider all words from 3-12
  - Based on research - http://arxiv.org/pdf/1207.2334.pdf
  - Based on tests, I ultimately chose 3-12 as it gave the best accuracy in determining spam and ham emails.
- Consider email addresses
  - Spam Email Addresses are similar as are Ham ones

Tests

- **Smoothing Constant**



- **Length of words to consider for Unigrams**



Analysis of Errors

Interestingly enough the two emails that were actually spam but counted as ham were either very long in length or very short in length. I think the former email simply didn’t have enough tokens for which it could make a decision, and the latter email had too many proper words for it to be diagnosed as spam.