**Spam Filter**

**Spamicity of a word**

The formula used by our model to determine the probability that an email is a spam given a word present in the email, is derived from Bayes’ theorem.

\Pr(S|W) = \frac{\Pr(W|S) \cdot \Pr(S)}{\Pr(W|S) \cdot \Pr(S) + \Pr(W|H) \cdot \Pr(H)}

Where,

* \Pr(S)is the overall probability that any given message is spam
* \Pr(W|S)is the probability that the word "W" appears in spam messages
* \Pr(H)is the overall probability that any given message is not spam (is "ham")
* \Pr(W|H) is the probability that the word "W" appears in ham messages.

**Spam probability**

The overall probability that a given mail is spam is calculated by the formula:

 p = \frac{1}{1 + e^\eta} 

Where,

\eta = \sum_{i=1}^N \left[ \ln(1-p_i) -\ln p_i \right] 

With pi being probability that an email is a spam knowing that it contains the ith word.

**Smoothing:**

We have used corrected probability to not consider words that occur only a few times in the training set. This is done as these words are unreliable.

\Pr'(S|W) = \frac{s \cdot \Pr(S) + n \cdot \Pr(S|W)}{s + n }

where:

* \Pr'(S|W)is the corrected probability for the message to be spam, knowing that it contains a given word ;
* sis the *strength* we give to background information about incoming spam, we consider s = 3;
* \Pr(S)is the probability of any incoming message to be spam ;
* nis the number of occurrences of this word during the learning phase ;
* \Pr(S|W)is the spamicity of this word.

Among 1000 emails from the test set, our model predicted 672 as spams though only 580 of them were actually spams. Thus our model wrongly classifies 92 of the total mails as spams. At the same time, our model predicted 328 mails as hams though 420 of them are actual hams.