Spam Filtering

The concept of spam filtering is simple - detect spam emails from authentic (non-spam/ham) emails. To do this, the goal would be to get a measure of how 'spammy' an incoming email is. The extended form of Bayes' Rule comes into play here.

With Bayes' Rule, we want to find the probability an email is spam, given it contains certain words. We do this by finding the probability that **each word in the email is spam**, and then **multiply these probabilities together** to get the overall email spam metric to be used in classification.

The the probability of an email being spam S given a certain word W appears is defined by the left hand side of the above equation $Pr(S|W)$.

The right hand side gives the formula to compute this probability. This is:

- the probability the word occurs in the email given it is a spam email $Pr(W|S)$ multiplied by the probability of an email being spam $Pr(S)$,
- divided the probability the word occurs in the email given it is a spam email multiplied by the probability of an email being spam,
- plus the probability the word occurs in the email given it is a **non-spam** email $Pr(W|\neg S)$ multiplied by the probability of an email being **non-spam** $Pr(\neg S)$.

Probabilities can range between 0 and 1. For this spam filter, we will define that any email with a total 'spaminess' metric of over 0.5 (50%) will be deemed a spam email.

When the $Pr(S|W)$ has been found for each word in the email, they are multiplied together to give the overall probability that the email is spam. If this probability is over the 'spam threshold' of **0.5**, the email is classified as a spam email.