

Naive Bayes Spam Filtering

Jamin Thornsberry

Start out with 2 training sets of emails

SPAM

bad email

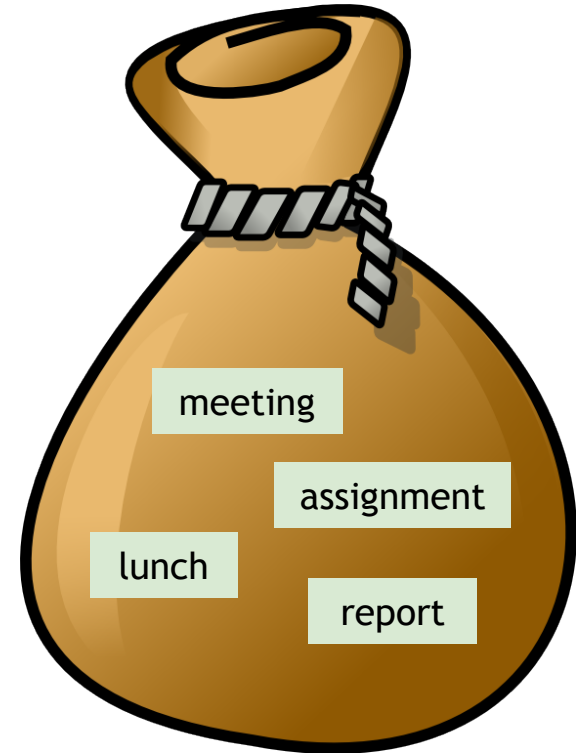
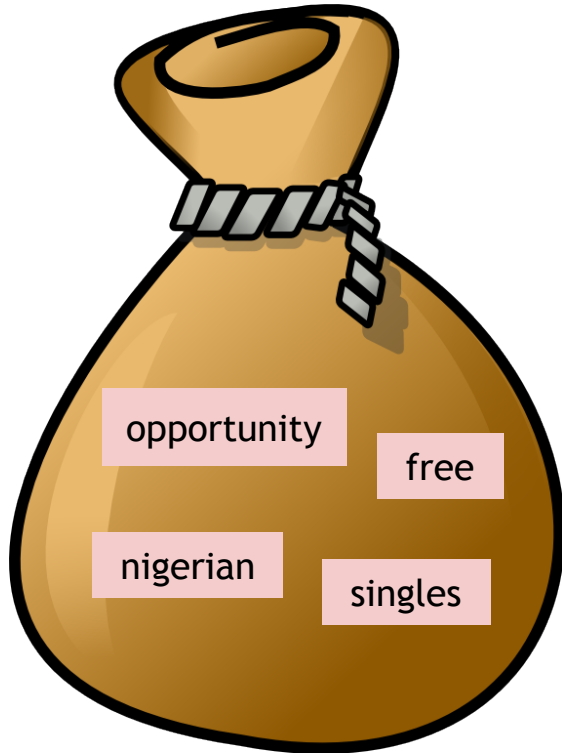


HAM

good email



Create 2 bags of words



The “spamcity” of a word is determined using Bayes’ theorem, taking into account the frequency a word appears in each of the spam emails vs. the ham emails

$$P(S|W) = \frac{P(W|S) * P(S)}{P(W|S) * P(S) + P(W|H) * P(H)}$$

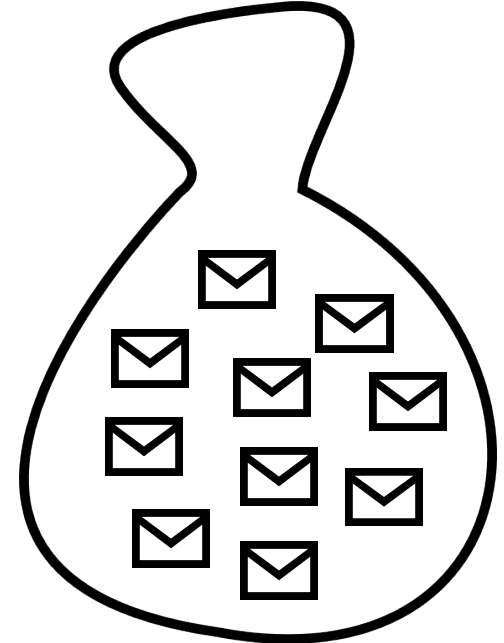
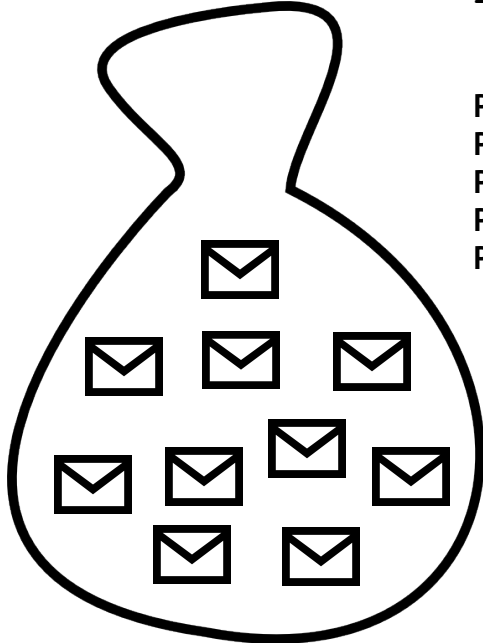
$P(S|W)$ is probability email containing word is spam

$P(W|S)$ is probability word appears in spam emails

$P(W|H)$ is probability word appears in ham emails

$P(S)$ is initial probability email is spam (assume 0.5)

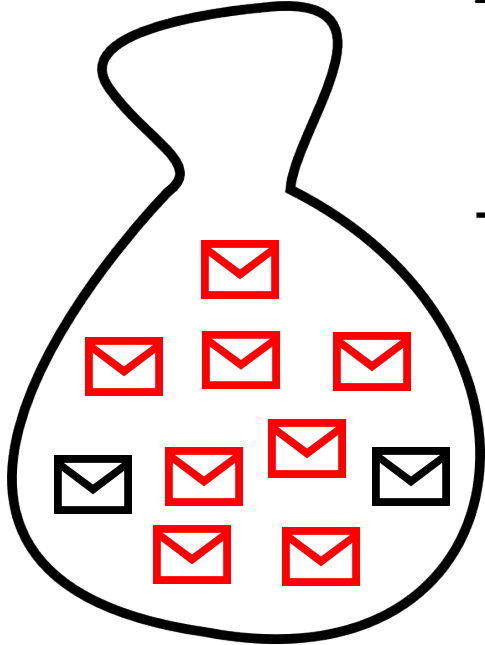
$P(H)$ is initial probability email is ham (assume 0.5)



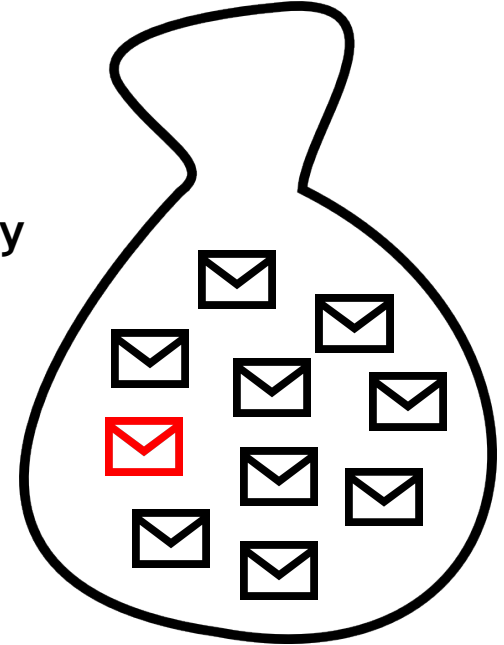
The “spamicity” of a word is determined using Bayes’ theorem, taking into account the frequency a word appears in each of the spam emails vs. the ham emails

$$P(S|winner) = \frac{0.8 * 0.5}{0.8 * 0.5 + 0.1 * 0.5}$$
$$\approx 0.8889$$

The word “winner” has a spamicity of about 0.8889



“winner” appears
in 8 out of the 10
spam emails.



“winner” appears
in 1 out of the 10
ham emails.

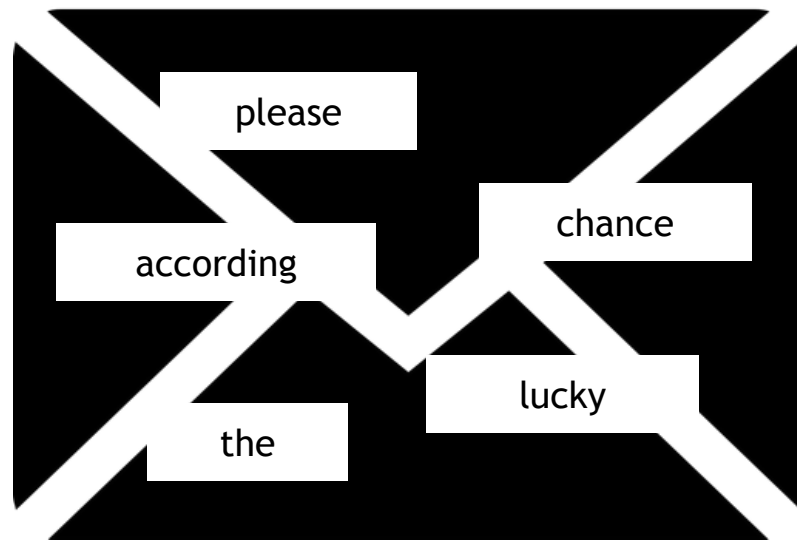
To find the probability that a certain email is spam, we use Bayes' theorem again, this time using the product of the spamicities of all the words in the email.

$$p = \frac{p_1 p_2 \dots p_N}{p_1 p_2 \dots p_N + (1 - p_1)(1 - p_2) \dots (1 - p_N)}$$

p is the probability that a given email is spam

$p_1 \dots p_N$ are the spamicities of the words in the email

$(1 - p_1) \dots (1 - p_N)$ are the inverse spamicities of the words



To find the probability that a certain email is spam, we use Bayes' theorem again, this time using the product of the spamicities of all the words in the email.

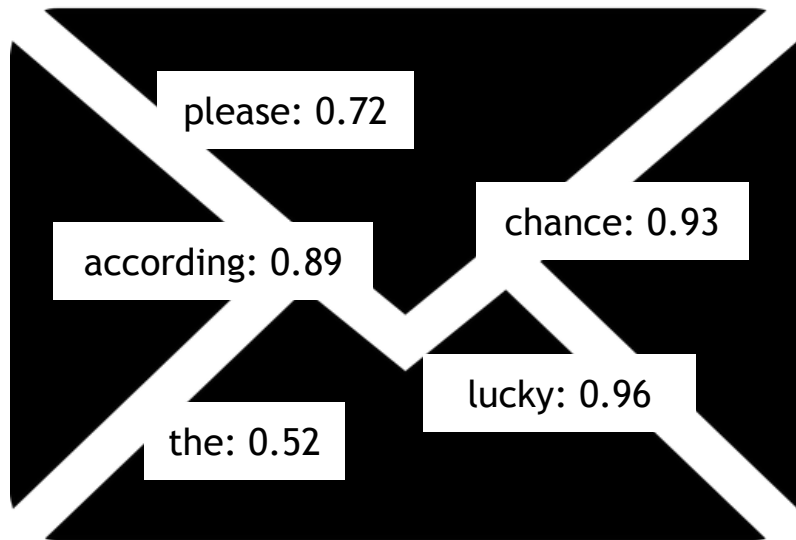
$$\frac{0.72}{(1 - 0.72)} * \frac{0.93}{\dots} * \frac{0.96}{\dots} * \frac{0.89}{\dots} * \frac{0.52}{(1 - 0.52)} = \frac{0.2975}{0.00004}$$

$$p = \frac{0.2975}{0.2975 + 0.00004} \approx 0.9999$$

This message is

SPAM!

(probably)



My Implementation

- Spam success rate (percentage of correctly labelled spam):
 - Around **96%**
- Ham success rate (percentage of correctly labelled ham):
 - Around **99%**

A success! ✓

Optimizations

- Only consider the top N words, sorted by the most “interesting” spamicities (farthest from 0.5)
- Ignore words that don’t occur often (not a trustworthy source of information)