LING 165 Lab 1: Spam detection using a naive Bayes classifier

Build a naive Bayes classifier that determines whether an email is spam or not (a.k.a. *ham*) based on words in its subject line.

Data

We have two files containing email subject lines from the SpamAssassin public corpus in `/home/ling165/lab1/` on the gray server:

(1) `spam_assassin.train`

(2) `spam_assassin.test`

Each line in a file specifies a data point in the following format:

```
class \t subject-line
```

`class` is `1` for spam and `0` for ham. `\t` denotes tab-space.

For example,

```
1       we pay cash now
0       asteroids anyone
```

Task

Build a naive Bayes classifier from scratch using (1) and report its performance on (2). More specifically,

(3) Assume there are two generative models: one for spams ($c = 1$) and one for hams ($c = 0$).

(4) Each model generates a subject-line one word at a time by sampling from a *bag of words* with replacement. So for example,

$$P(\text{we pay cash now}|c = 1) = P(\text{we}|c = 1) \cdot P(\text{pay}|c = 1) \cdot P(\text{cash}|c = 1) \cdot P(\text{now}|c = 1)$$

(5) Apply add-one smoothing to estimate the probability of choosing a word from the bag. Assume that the vocabulary for each model consists of all words relevant to the model in (1) plus a dummy word reserved for any unknown word that the model may later encounter. So for example,

$$P(\text{we}|c = 1) = \frac{freq(\text{we}, c = 1) + 1}{N_1 + |V_1|}$$

where $freq(\text{we}, c = 1)$ denotes how often the word `we` occurred in data points labeled `1`, N_1 denotes the total number of word tokens in data points labeled `1`, and $|V_1|$ equals the number of different words in data points labeled `1` plus one (for the dummy word).

(6) Report performance on (2) in terms of precision and recall, where

- precision: the proportion of true spams that your classifier detected out of those that your classifier thought were spam
- recall: the proportion of true spams that your classifier detected out of those that should have been detected

(7) Email me the performance scores and let me know where I can see your work.