Belol Nessar
CSCI420 - Machine Learning
December 2015

Bayesian Filter Report

Features of the spam filter:
When determining ham or spam, the program counts the occurrences of **spammy punctuation marks**, e.g. exclamation marks, questions marks, and dollar signs.
When determining ham or spam, the program only considers the **100 most common words** that it has observed from studying the ham and spam archives. In the common case where multiple words have the same number of occurrences, all of those words are added. So the actual number of words considered tends to be somewhere between 130 and 180.
Actually, there is another list of words that the program keeps track of, despite their not (necessarily) being in the top 100. It includes what I consider to be **inherently spammy words**, such as “free”, “sale”, “singles”, “weight”, and “natural”.
The program keeps a list of **“trash words”** that should be ignored when determining ham or spam. These words are mostly “noise” words that appear frequently in both ham *and* spam, e.g. html tags; common English filler words (“the”, “that”, “it”); and the commonly-occurring names/corresponding entries in e-mail headers (font names, months, days of the week, and the contents of assets/header_field_names.txt).
The program only considers words which, when lowercased, can be found in the **American dictionary**. I make an exception for **titlecased words**, which may be useful proper nouns.

CONFUSION TABLE (**using “full”**):

	POSITIVE	**NEGATIVE**
TRUE	`0.336582863007`	`0.560800410467`
FALSE	`0.0297588506927`	`0.0728578758338`

CONFUSION TABLE (**using “decap”**):

	POSITIVE	**NEGATIVE**
TRUE	`0.352488455618`	`0.503848127245`
FALSE	`0.0867111339148`	`0.0569522832222`

Note: I got the idea to use only the top 100-ish words from the following website:

http://www.paulgraham.com/better.html