SPAM Filter Task

Andrew Zieffler

2018-07-15

Dr. Zieffler is tired of getting so much SPAM in his email inbox. He is hiring you to design a SPAM filter that will determine whether or not each email message he receives is SPAM or a legitimate message based on the subject line of the email message. As you design the SPAM filter, keep in mind that: (1) different people have different criteria for what constitutes SPAM; and (2) a SPAM filter will be specific to a given individual—the SPAM filter you develop needs to work only for Dr. Zieffler's emails.

Classification Rules

In order to develop the SPAM filter, you will be given a data set that contains subject lines from a sample of both SPAM and non-SPAM email messages that Dr. Zieffler has received. Use these data to develop the SPAM filter. Even though SPAM filters may use a variety of information (email address of sender, reply to email address, subject line, body of the text, etc.), the data set consists only of the subject line for each email. Your task is to develop a set of rules that will identify SPAM and NON-SPAM emails using only the subject line. Your group needs to come to consensus on the set of rules that comprise the SPAM filter.

Measuring Effectiveness of the Filter

Once you have a SPAM filter that you believe works well, you will be given another sample of subject lines from Dr. Zieffler's emails to test your set of classification rules. Based on this test data set, you will need to produce a numerical measure of how effective the rules were in detecting SPAM.

Presentation

Your group will then put together a presentation for Dr. Zieffler that includes the following:

1. A description of the rules you used to determine if an email was SPAM or not.
2. A description of the numerical measure you used to determine how well the SPAM filter works.
3. A description of how you might adapt the rules based on results from the test data.