

CSCI 4511 writing 3

Zirui Chen

March 2022

Description

With the rapid development of the internet comes with rapid increases of spam messages. Spam message, as defined by Merriam Webster dictionary, is an unsolicited commercial message (such as email, text message, or Internet posting) sent to a large number of recipients or posted in a large number of places. It is estimated that over 70 percent of email sent worldwide can be categorized as spam.[1] It has been a challenging topic since the content of spam messages can vary, and it would be hard to identify them from usual correspondences. There are different types of spam filters, namely content filter, header filter, block filter and rules-based filter.[1] Each one of these filters target a specific part of the email. For the purpose of this project, I will be focusing on content spam filtering for SMS messages, which also takes a significant portion of spam messages

Approach

There are many exciting approaches to solve this problem. The approach I am planning on taking is to use NLP and machine learning. Data set that consists of 5000 messages(about 1000 of which are spam messages) has been downloaded. The NLTK language library will be used to clean and tokenize data. Machine learning concepts will be used to classify ham or spam.[2] Data set will be randomly divided into two sets(80 percent training and 20 percent testing). After the model is trained, test data will be used to evaluate the model. Accuracy will be reported.[3] Implemented the program with uni-gram, bi-gram and trig-ram. Test whether N-gram and smoothing will affect the accuracy of such a program.

Preliminary result

Firstly, data set is divided into two groups(spam or ham). Word occurrence are calculated and plotted using matplotlib. the bigger the word is,

by total number of spam messages (false negative included). Accuracy is the overall chance of predictions is correct.

Evaluation

Results will be evaluated by four scores, namely Precision, F-score, recall and accuracy. Results generated by programs using different method will be compared and analysed.[5]

Time Frame

This project has several due dates. This proposal is due March 18. Literature review is due April 1 and final report is due May 2. Since I would like to produce Preliminary results prior the due date for writing 4, the program must be up and running before April 1. Before April 15, programs with different N-gram and smoothing should be implemented. Model and results should be generated and analysed.

References

- [1] unknown. What is spam filtering?
- [2] Aleksander Kolcz Yang Song and C.Lee Giles. Better naive bayes classification for high-precision spam detection. 2009.
- [3] Tejan Karmali. Spam classifier in python from scratch, 2017.
- [4] Anirudha Simha. Understanding tf-idf for machine learning, 2021.
- [5] Thomas Wood. What is the f-score?