**Project 1. Email Spam Classification Haejong Dong & Marko Haataja**

Today, we are all living in a period of human history known as the Information Age. The phenomenon of the Information Age is that the digital industry creates a knowledge-based society surrounded by a high-tech global economy. The world progressed to the Information Age by growing information storage, information transmission and computation capabilities. [1]

One of the most important factors in progressing to the Information Age is the founding of the Internet. The Internet has enabled new forms of social interaction, activities and social associations. Internet users have steadily risen from about 10% in 1997 to about 80% in 2014. [2] One of the earliest uses, and reasons why it was created, of the Internet was email. Email preceded the Internet, reaching back to the early ARPANET. [3]

Email was the early connector of people, and it is still used today very much. However, soon after the email went mainstream unsolicited messages were being received by most email users. This junk mail is called spam. Spam messages can be commercial but unsolicited or malicious that tries to get user’s credentials or credit card number or something else that can bring them money.



Figure 1: Email spam

In 2010, it was estimated that the proportion of spam email was around 80% of all email messages sent, with Microsoft estimating unwanted messages being up to 97% of all messages on the Internet. This amount of unwanted messages cost estimated 21 billion US dollars of lost productivity in 2004. The amount of lost productivity and user inconvenience begged for a solution, and email spam message filtering was one of the tools used. [5]



Figure 2: Email spam filtering

To combat the problem of spam, email filtering was used to process the received emails and to organize it according to specific criteria, e.g. spam and ham (legitimate messages). In some cases, this technique reduces the amount of spam received significantly. This does not come without problems, though. Often legitimate messages, or ham, are erroneously categorized into spam, and most likely it is never read by the recipient. [6]

This is where data classification and Natural Language Processing (NLP) steps in. ( By using advanced methods of data classification, email classifiers and filters have become so powerful, that the amount of spam received has dropped to record low.) [4]

In Data and Text Mining course work, our objective was to study Email Spam Classification, and test several implementations and compare the results.

The project work first familiarized us with basic email classification and testing, using the techniques that we learned earlier in the course. After that, several other weighing schemes needed to be tested and results reported. In addition to other weighing schemes, different ways of choosing the vocabulary were tested. After testing all the variations of different classifiers, weighing schemes and ways of choosing vocabulary, the data set needed to change. We wrote a Python script to extract spam and ham messages from our email folders to be used as a data set. The earlier experiments were then re-run with the new dataset. The results of all the tests can be found in the document “classifier\_report”.

[1] <https://en.wikipedia.org/wiki/Information_Age>

[2] <https://en.wikipedia.org/wiki/International_Telecommunication_Union>

[3] <https://en.wikipedia.org/wiki/Email>

[4] <http://www.themonitordaily.com/email-spam-rates-reaching-all-time-low-might-be-the-calm-before-the-malware-storm/24025/> “Spam hits all time low”

[5] <https://en.wikipedia.org/wiki/Email_spam>

[6] https://en.wikipedia.org/wiki/Email\_filtering