Nico Paik
MIS3640-01
Professor Li Zhi

Assignment 2: Text Mining & Analysis
SMS Spam Collection Data

Project Overview

The data sourced explored in this project is a text file collection provided by http://www.dt.fee.unicamp.br/~tiago/smsspamcollection, which is made up of 5,574 English, real and non-encoded messages. These messages are labeled accordingly as legitimate as ham or non-legitimate as spam. The goal of this project is to analyze the data through word frequencies to identify which words are commonly used in spam and ham messages. These word frequencies are visually displayed through WordCloud and matplotlib.pyplot libraries in Python in order to characterize word frequencies as ham or spam in texts and messages.

Implementation

The text mining and analysis implementation is separated into two parts: data processing and data visualization. For the data processing portion, the data was imported into a pandas data frame and processed with the objective of creating two lists of words: corresponding to spam and ham. After lists of words are defined, the lists are inputted into the WordCloud function, which generates a word cloud. The words increase in font size as the frequency of the word increases. Finally, Matplotlib was used to output the word cloud.

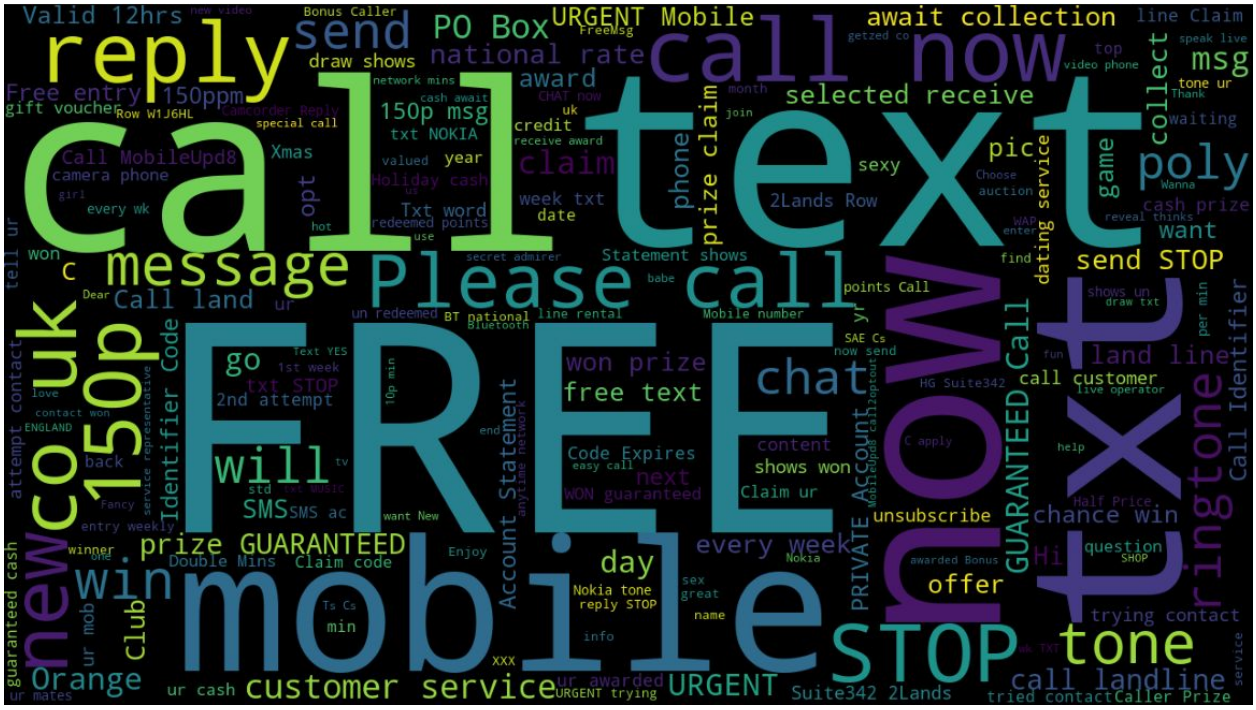
To characterize word frequencies in this data, a wordcloud was utilized to visually demonstrate which words were frequently used in ham or spam messages. The wordcloud is helpful in displaying a long list of words explored in the texts; most importantly, it helps the reader quickly perceive which words are the most prominent in a ham or spam message. The text size of the words allows the reader to get a sense of how frequent the word appears in the text; as the size increases, the more commonly the word is used in a message. When choosing the wordcloud design as a visual demonstration of the data's characterization of word frequencies, other options such as a plot word frequency bar graph, frequency term-document matrix, or word frequency scatter plot could also be utilized. However, the wordcloud design was chosen because it efficiently allows the reader to perceive the relativity of the prominence of each word to another in ham or spam messages through text sizes and colors. On the other hand,

if the reader would prefer to specifically identify the exact numerical values of the frequency of words in the messages, it is a better option to choose a plot word frequency bar graph or frequency term-document matrix as it demonstrates them clearly.

Results

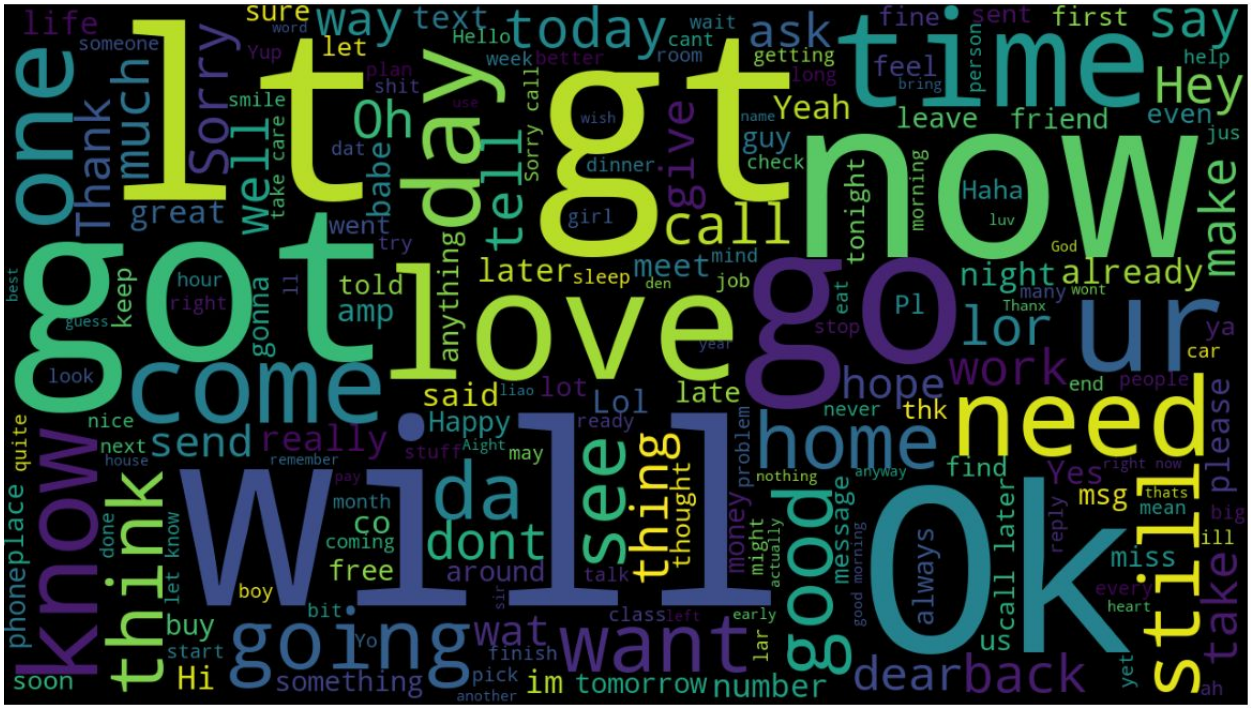
Through the wordcloud graphic, the reader is able to view the frequency of words in the data provided to differentiate which words are more prominent in ham or spam messages. In the first wordcloud visualization, the biggest texts in size and frequency were “free”, “call”, “text”, “now”, “mobile, “txt”, and “reply”. These findings were interesting because I was expecting the results shown in the wordcloud visualization. In my own personal experience when receiving “spam” messages, I frequently get messages that urge me to quickly call, message, or reply back now to an unknown number. Apart from that, the “spam messages” received seek to sell me something for “free”, which is a common scam and spam tactic used to deceive receivers into replying to spam. My expectations of what type of words would frequently appear in “spam” messages are parallel to the results shown in the wordcloud because they are along the lines of “reply”, “text”, “free”, “call”, and “now”. The wordcloud visualization for frequent words in “spam” text messages in the data provided is shown in Exhibit 1.

Exhibit 1



In the second wordcloud visualization, the reader is able to perceive which words are most prominent in ham messages, being recognized as legitimate and non-spam. Unlike the first wordcloud visualization, which demonstrated which words were most prominent in “spam” messages, it was challenging to expect what kind of words would frequently appear in “ham” messages. This is mainly because spam messages have common characteristics in using words such as “free”, “reply”, “call”, and text” to urge the receiver to respond back to the spam message. On the other hand, “ham” messages are words that, put simply, are not “spam” and do not contain those common words that are recognized as “spam”. The results of the second wordcloud visualization showed the biggest and most frequent words being “ will”, “it”, “gt”, “now”, “ok”, and “got”. The words shown in the second wordcloud visualization helps the reader understand which words are most commonly used in “everyday” messages because they are legitimate and not recognized as “spam” messages. However, these findings were interesting to analyze because the word “now” was also a prominent word used in the first wordcloud visualization for “spam” messages. The difference in the commonality of the frequent word “now” in “spam” and “ham” messages lies in the context of the message and deciphering if the overall message is recognized as “spam” or not. The wordcloud visualization for frequent words in “ham” text messages in the data provided is shown in Exhibit 2.

Exhibit 2



Reflection

This project successfully explored a method that allows the reader perceive the prominence and frequency of words in “ham” or “spam” messages efficiently. The project was appropriately scope because it utilized a large sample size of 5,574 English, real and non-decoded messages; however, it could be improved by ensuring a larger sample size from the same area of the messages sent/received. This project has expanded my knowledge on data mining and analysis in Python by exploring methods to visually demonstrate word frequency and prominence in given texts. I was able to discover a visual representation of text frequency through wordclouds and in the future would like to discover using other text frequency visualization methods such as a frequency bar graph, frequency term-document matrix, or word frequency scatter plot in different projects. I would also like to learn to manipulate not only the size of the texts in wordclouds, but also colors, which I was not able to explore. For instance, providing a sensitivity analysis of the words in the wordcloud by identifying certain colors, such as red shades, as harsher/negative connotations, and blue shades, as softer/positive connotations.