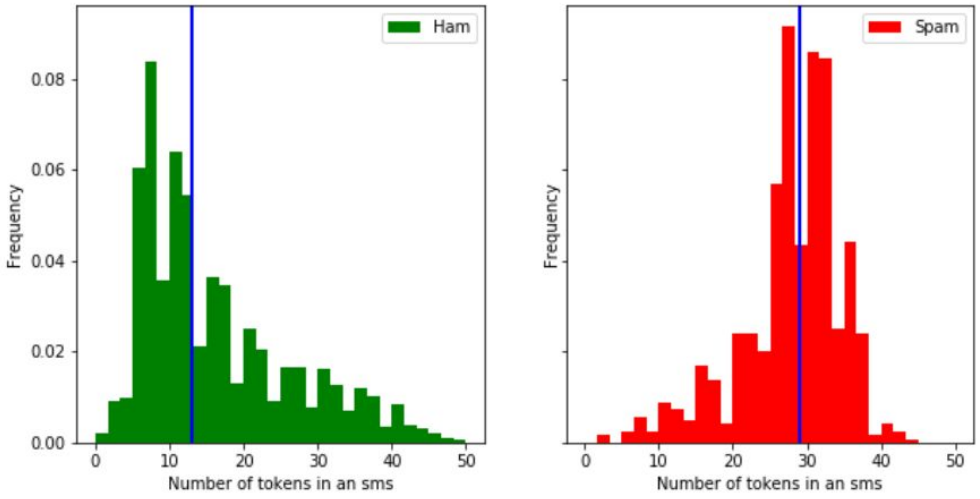
Capstone Project 1: Data Story
Spring Board Data Science Career Track
~ By Zach Nguyen

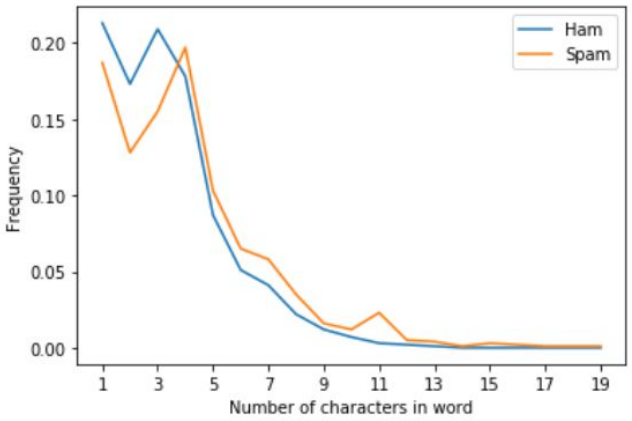
A. Classical feature extraction:

Exploratory Data Analysis was conducted to further get to know the data as well as to extract discriminative features. The following features were analyzed:

- **Length of the SMS**: The number of tokens in each sms was calculated to create the feature n_token for each document. Overall, spams seem to have a higher token count for each document. The following histograms help visualize the difference between ham and spam in terms of this variable. The blue line represent the median token count.

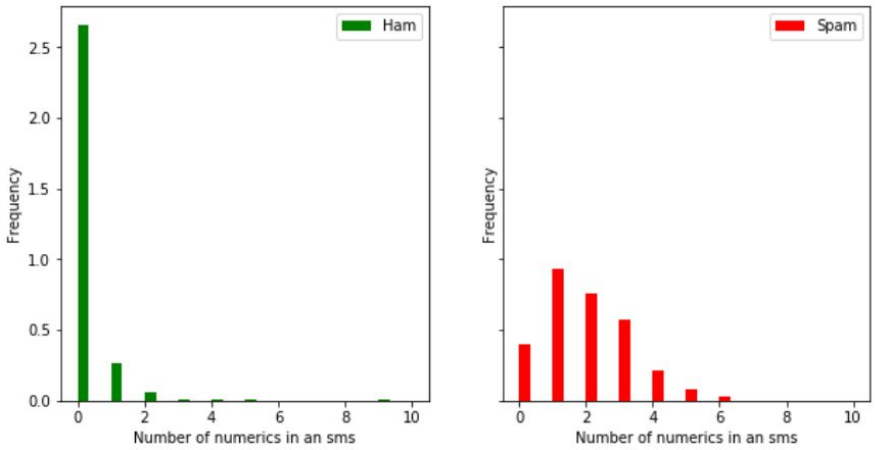


- **Length of sms words**: Length of sms words seem to be discriminative, although not by much as we can see in the following visualization. It seems both ham and spam undergo an exponential decay in character length (which is very typical since shorter words are much more frequent). However, permutation test shows that there is a statistically significant test with p-value of 0 for 10,000 permutation replicates between these two sets. Thus, we can extract n_token as one of our features.

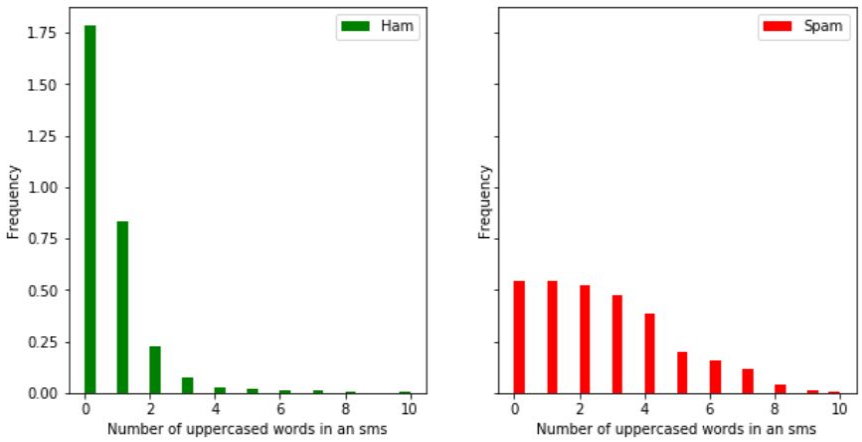


In addition to Length of sms words, maximum and minimum word length was also experimented and visualized. However, these features were shown to be non-discriminative.

- **Number of numerics**: The number of numeric characters within an sms seems to play an important role. It seems that most spam sms has numbers whereas most ham sms don't as visualized below by the number of numerics in a ham or spam document. Since we want to minimize dimensionality, we will extract has_num as a binary feature which describes whether the set has numbers.

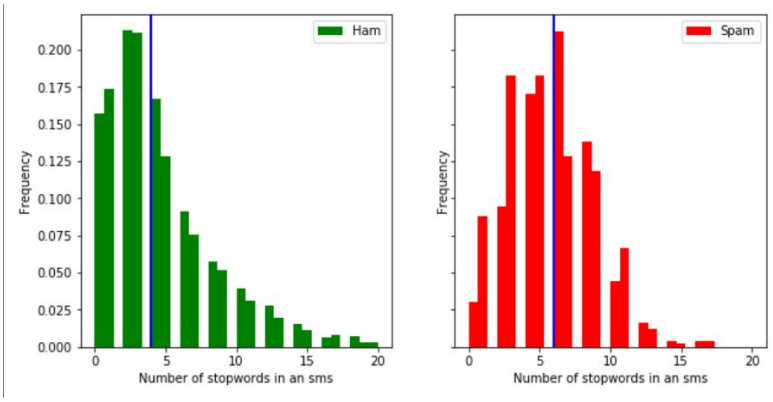


- **Number of uppercase:** In general, ham text has fewer uppercase tokens than spam text. However, the relationship is not entirely discriminative since there is still a sizeable proportion of spam messages with few uppercase as shown in the histogram below:



We will, thus keep the feature as n_upper.

- **Number of stopwords:** The distribution of stopwords in spam and ham is different with a different center (median) as shown below:



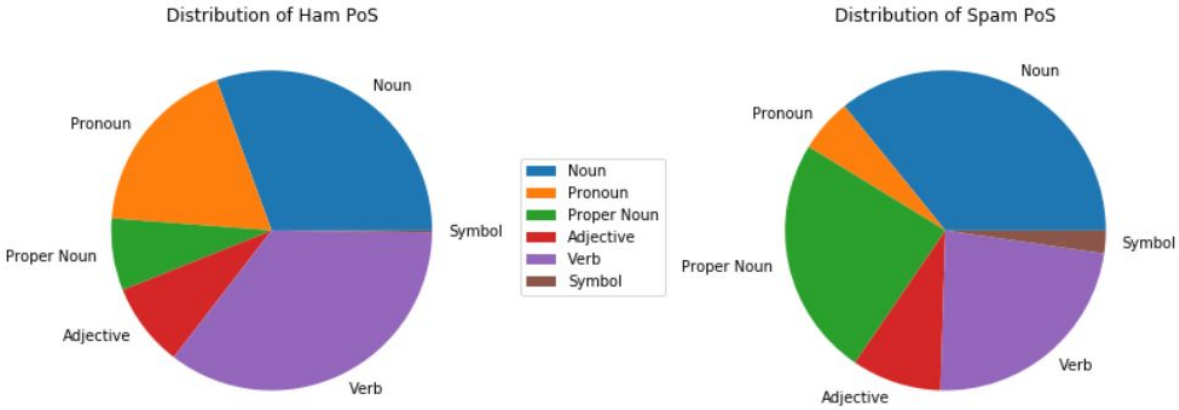
We will then keep the feature (as n_stop) as it has some predictive power.

- **Discriminative symbols:** Analyze non-verbal information such as the appearance of email addresses, currencies, phone numbers and urls. The frequency of those within ham and spam is tabulated in the chart below. It is concluded that these factors can be disciminative in classifying ham and spam.

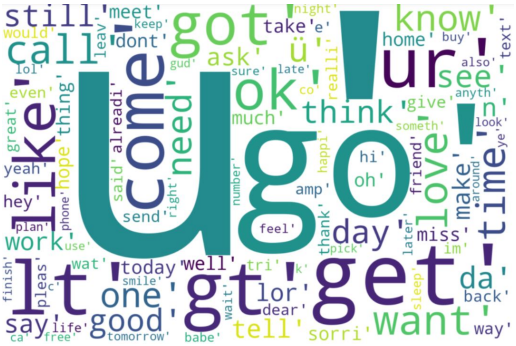
label	has_email	has_money	has_phone	has_url
ham	3	19	1	0
spam	18	256	405	21

B. Other visualizations

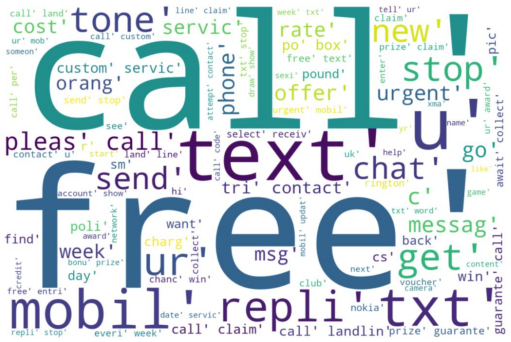
In addition to, other NLP and visualization methods such as TF-IDF analysis of the most relatively frequent words, word clouds, Part of Speech with Spacy, and Topic Modelling with gensim was also employed to get a sense of the data. The following Part of Speech Frequency shows that part of speech can be discriminative in ham/spam sms, particularly Symbols, Pronouns, Proper Nouns and Verb.



The following word clouds of lemmatized tokens show that ham messages are often conversational and informal, with an emphasis on quick and short words.



Whereas the spam messages can easily be flagged with certain words or phrases.



Finally, a visualization of LDA topic modelling shows that hammy and spammy words can be distinguish accross the distance of Principle Component 1 (topic 2, 6 and 15 contain spammy words):