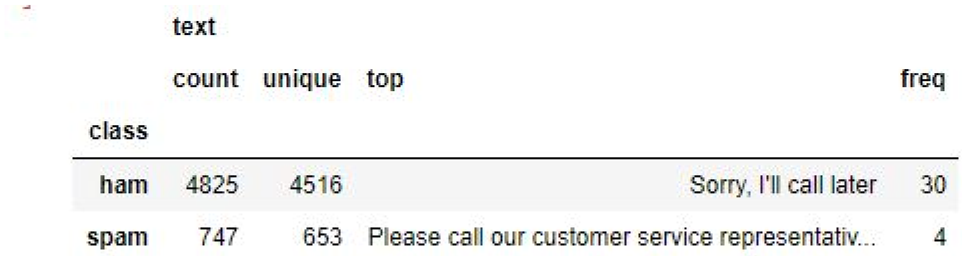
## Data Collection

The data has two columns of Ham and Spam messages. There are total 5572 rows.
https://www.kaggle.com/astandrik/simple-spam-filter-using-naive-bayes/data



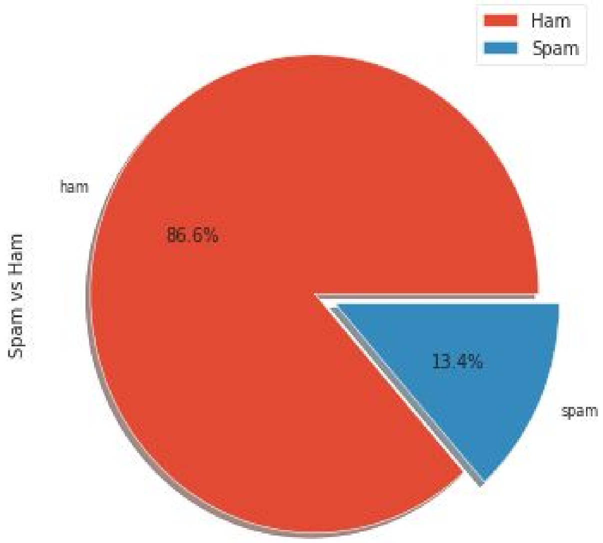
|  | text |  |  |  |
|---|---|---|---|---|
|  | count | unique | top | freq |
| class |  |  |  |  |
| ham | 4825 | 4516 | Sorry, I'll call later | 30 |
| spam | 747 | 653 | Please call our customer service representativ... | 4 |

From the above information it can be determined that about only 15.48% of messages are classified as spam. There are some also some duplicate messages since the number of unique value is lower than the count value.

## Data Preparation

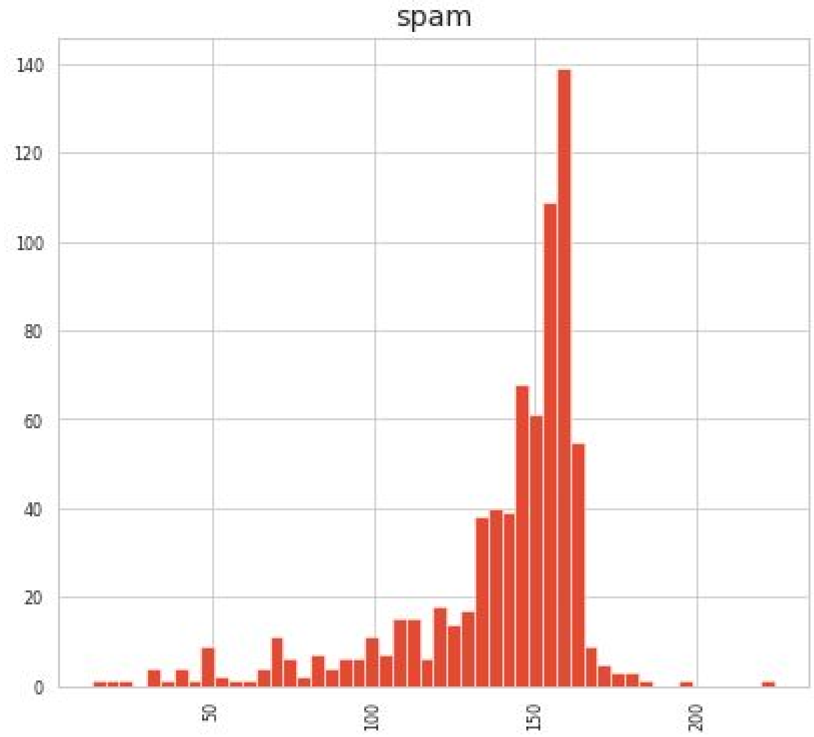
In cleaning process punctuation and stop words are removed by the help of CountVectorizer. The CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document.

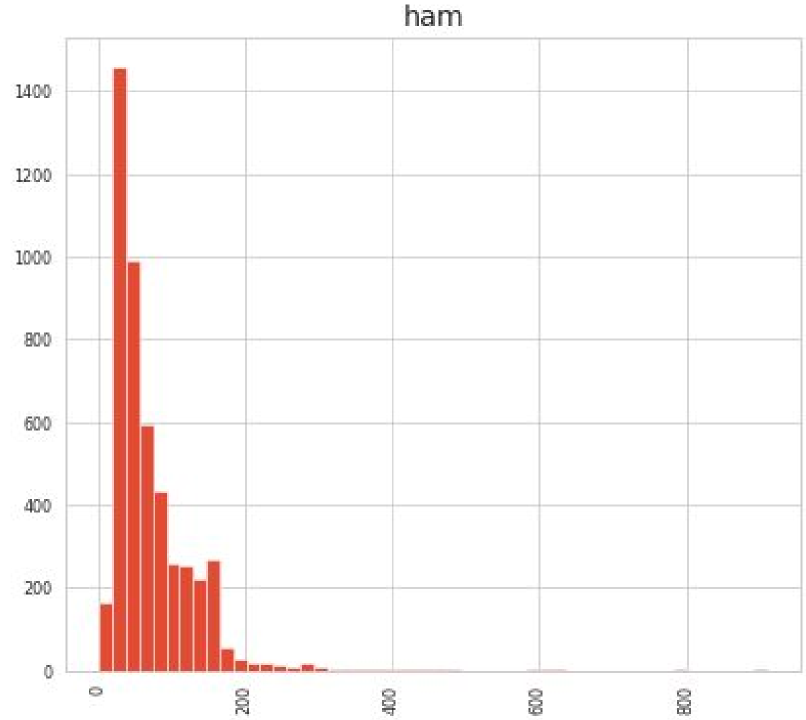
## Exploratory Data Analysis



According to the pie chart there is 86.6% of ham messages and 13.4% of spam messages.

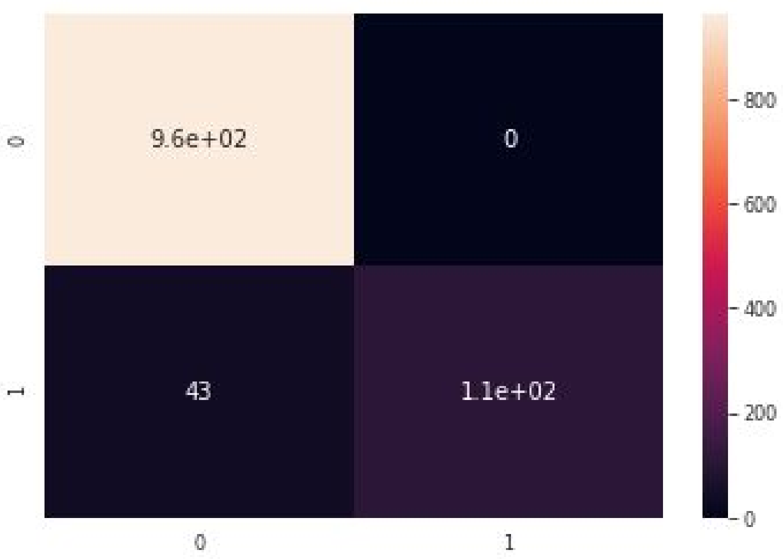
Histogram for the text length of spam and ham messages.





# Evaluation

Following confusion matrix is been plotted after the results.


0
1
9.6e+02
0
43
1.1e+02
0
1
800
600
400
200
0