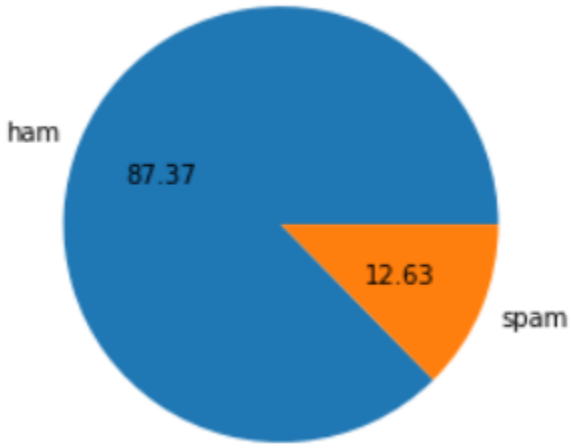
Email/SMS Spam Detection

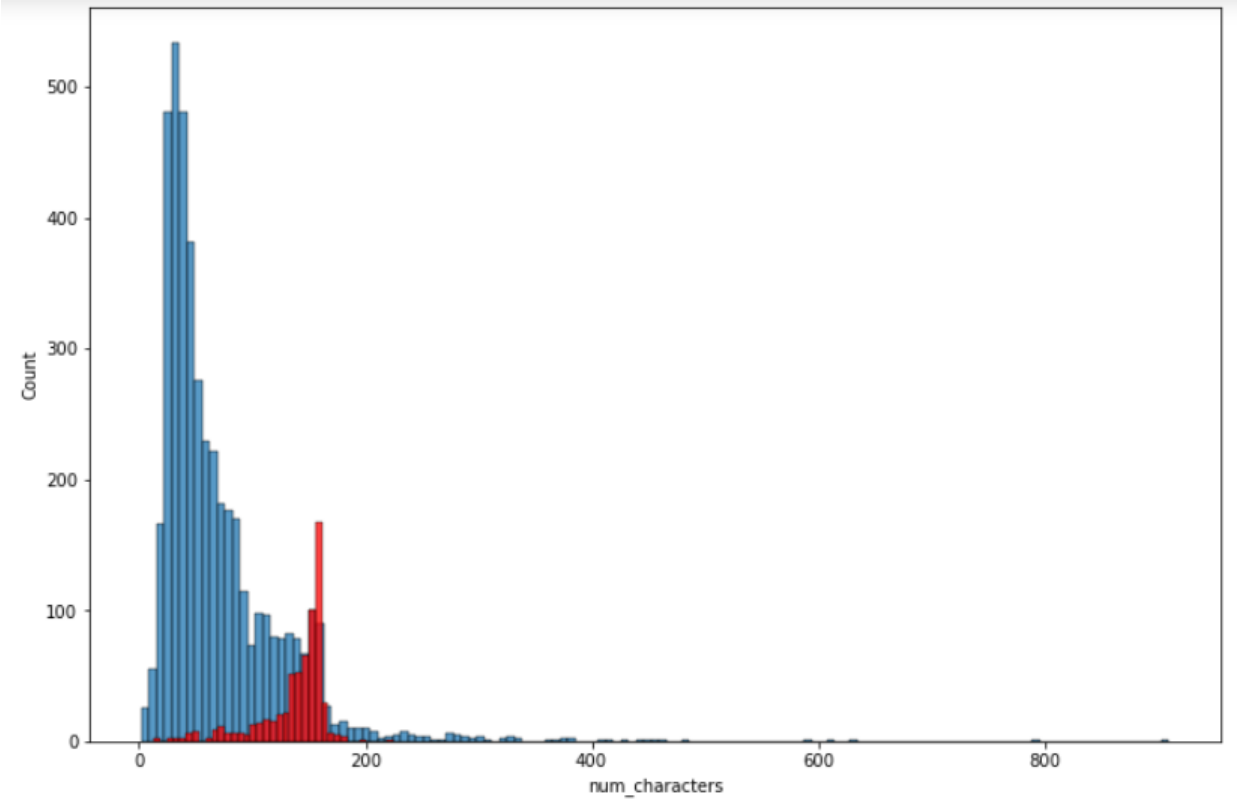
Data Analysis Report



Pie chart for ham and spam messages (target column)

Observation:-

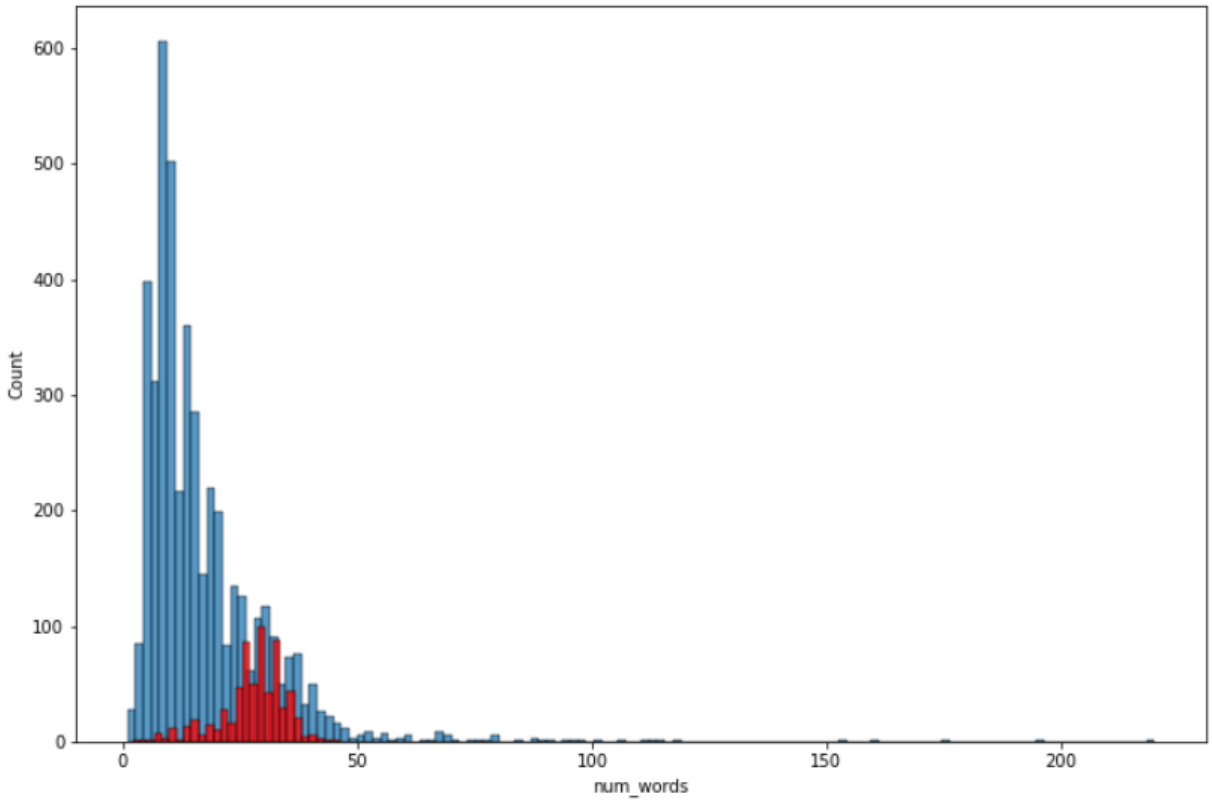
- It is seen that the data is imbalanced.
- Ham messages are 87.37% and spam messages are 12.63%.



Histogram for num_characters

Observations:-

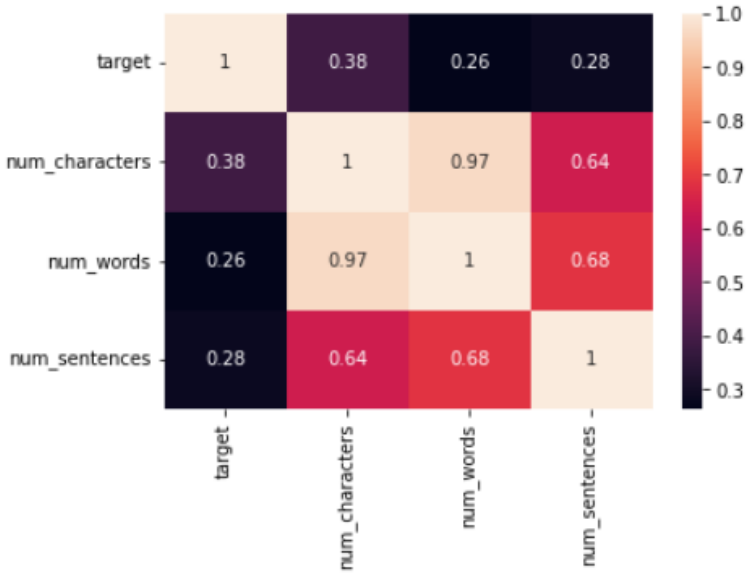
- Majority of ham messages are made up of small number of characters.
- Majority of spam messages are made up of large number of characters.
- Ouliers are also present in ham messages with more than 900 characters.



Histogram for num_words

Observations:-

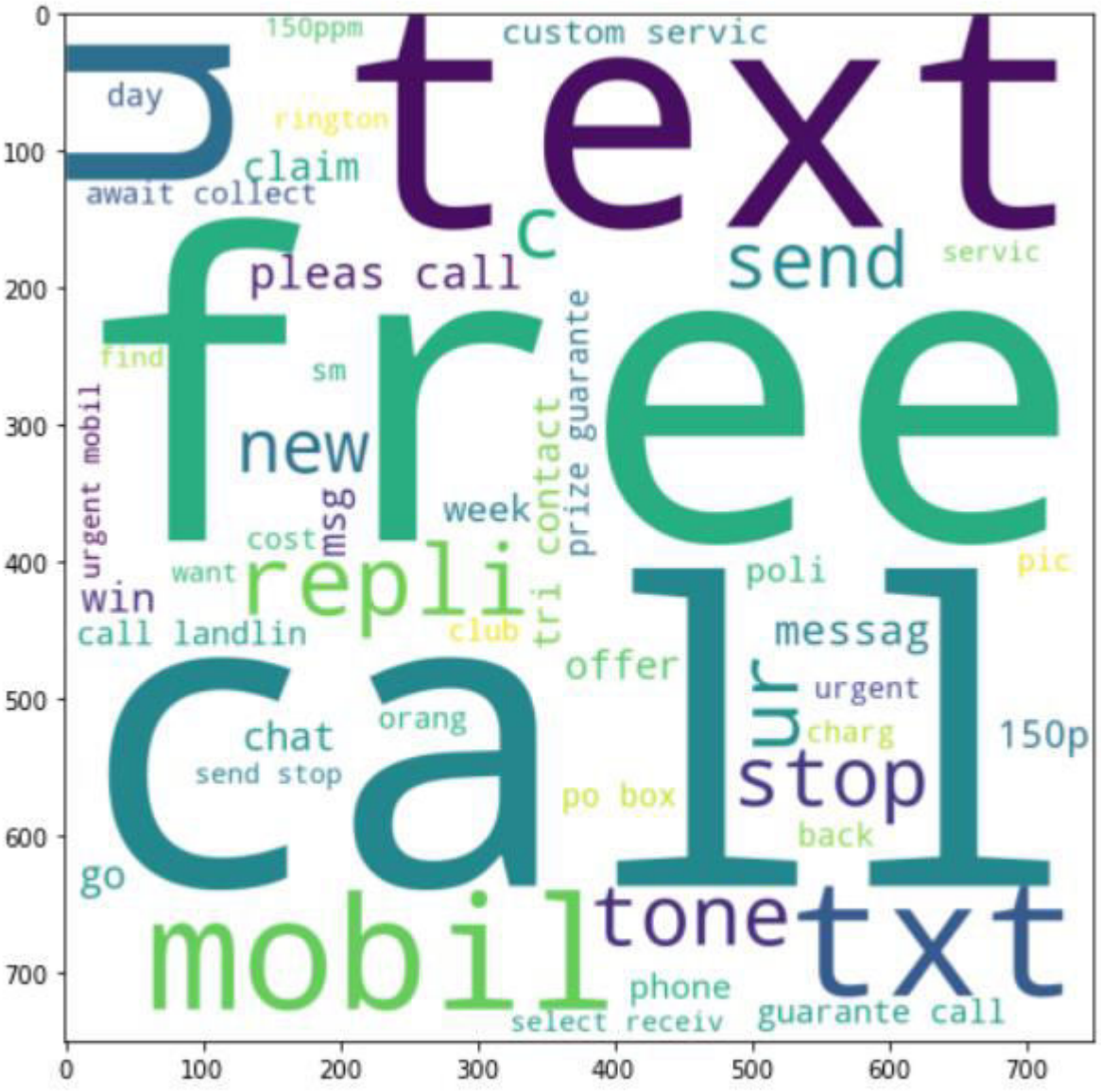
- Majority of ham messages are made up of small number of characters.
- Majority of spam messages are made up of large number of characters.
- Ouliers are also present in ham messages with more than 900 characters.



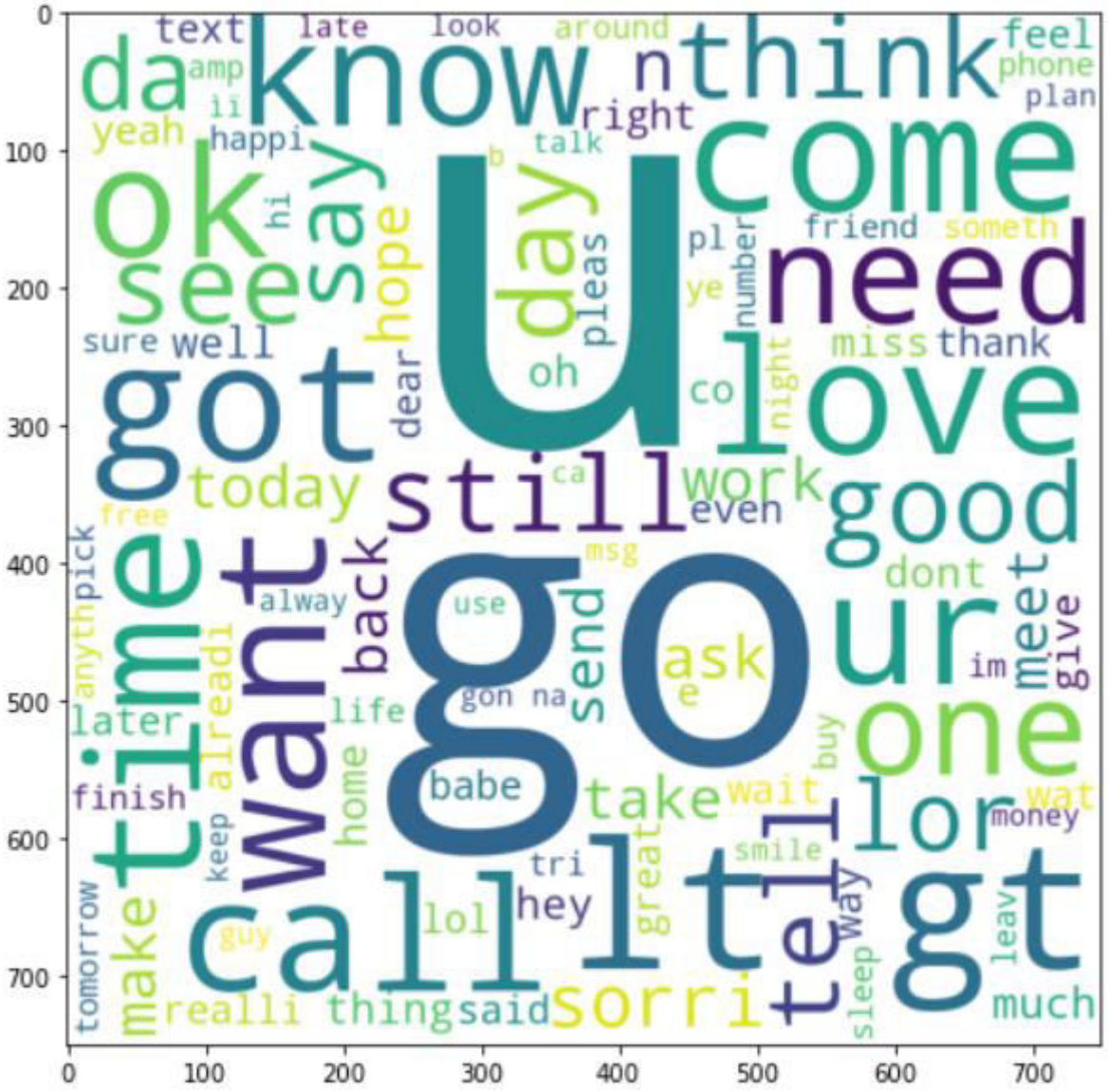
Heatmap showing correlation between features

Observation:-

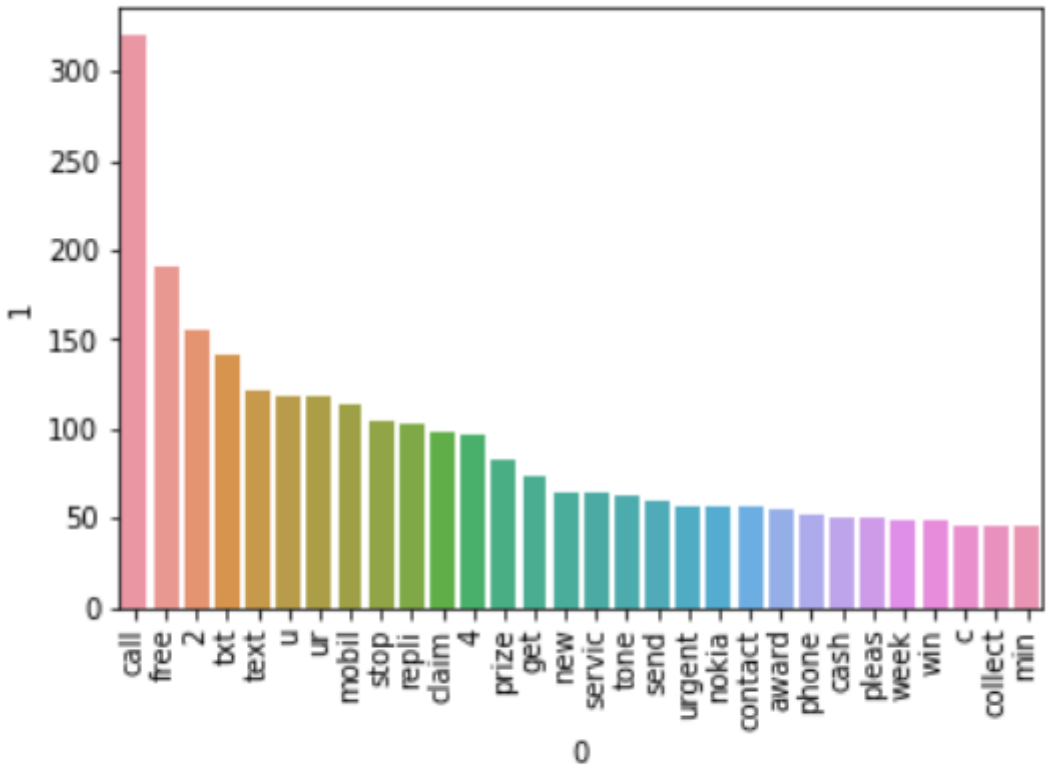
- Target has good correlated with num_characters, as compare to num_words, num_sentences.
- num_characters has very high correlation with num_words, and good correlation with num_sentences.
- num_words and num_sentences also has good correlation.
- Multicoliniarity exist between the columns.
- Keep num_characters column only as it has good correlation with target column.



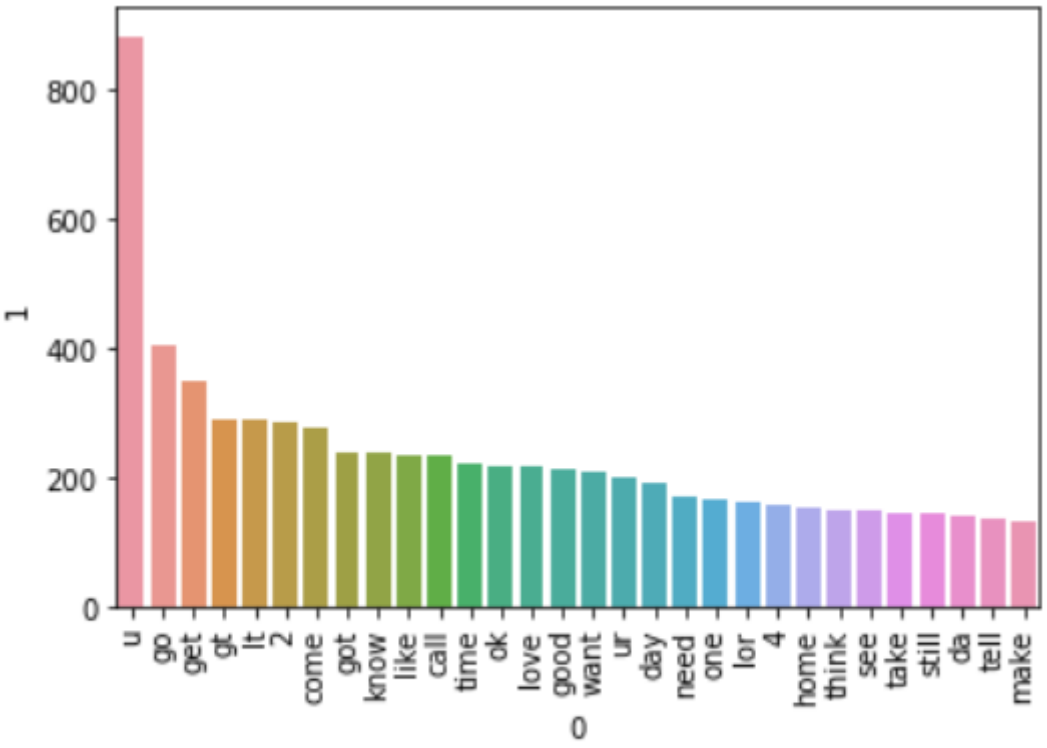
Wordcloud for spam message



Wordcloud for ham message



Top 30 words in spam messages



Top 30 words in ham messages