The data set to be analysed includes several SMS messages classified with the ham/spam tag. The goal is to build a classifier that, given the evidence provided by all the words contained in the message, classifies it as spam or ham.

1. Start by loading the data set (sms_spam.csv).
2. Explore the data set.
   a. View some ham/spam messages.
   b. View the sms with the highest number of words.
   c. Check the distribution of the target attribute.
   d. Make a histogram with the distribution of the number of words by SMS.

Text Preprocessing

3. Develop a function that reduces the words contained in a string to its radical. The function must:
   a. convert the letters to small letters
   b. remove all digits
   c. remove special characters
   d. remove stop-words
   e. reduce words to their radical using the Porter Stemmer.

4. Separate the ham/spam messages and apply the previously defined function to both.
5. View the most common words in each ham/spam message type using a word cloud[1].

The final step of text processing is to divide messages into individual components through a process called **tokenization**. The tokenization function takes a corpus and creates a matrix with the occurrences of the tokens. The matrix lines indicate the documents (SMS messages) and the columns indicate tokens (words).

6. With the count vectorizer function (sklearn), convert the corpus into a token/occurrence matrix.

1 In a word cloud, the words that appear most often in the text are displayed in a larger font, while less common words are shown in smaller fonts.

The sparse matrix includes all the words that appear in at least one SMS message. It is unlikely that all of these words will be useful for classifying messages. Thus, it is advisable to delete the terms with more and fewer occurrences.

7. Visualise quantitatively and graphically the occurrence of terms in all SMS and the number of terms in each SMS and define the terms and SMS to maintain.

Model Training

8. Create the training and test set (70% / 30%) stratified, with the original distribution of the target attribute (spam/ham) in both sets.
9. Develop spam/ham forecasting models using the cross-validation method (10 folds) with the following algorithms:
   a. Logistics Regression
   b. Decision Tree
   c. Naive Bayes
   d. KNN
   e. SVM
10. Finally, select the best algorithm to make the spam/ham forecast, rebuild the model with the training set, and evaluate it with the test set.