**Text Analysis Lab 1**

* Using a labelled text dataset, do the following exercises and answer the following questions:

NOTE: for each of the new methods that you need to use to complete the lab, you need to check out its documentation.

NOTE: for each question, you can type your answer in the actual script then comment it out or, if you are using jupyter notebook, you can use markdown.

1. Create a new script and import re, string, numpy, pandas, nltk, and matplotlib.pyplot.
2. Download the spam\_short.csv file and load it into your script.
3. Retrieve and output the first 5 instances/rows and the last 10 instances: use head() and tail(); retrieve also and output 7 random instances (use the sample method). Retrieve and output the following rows: 34, 57, 109.
4. Use the info method
   1. What is the type of object that stores the data?
   2. How many instances/rows does the dataframe have?
   3. How many columns/attributes does it have and what are their names?
5. Use the describe method to obtain the overall statistics. Interpret them.
6. Remove the duplicate rows and output the new stats. Use them to compute the percentages of spam and ham after removing duplicates, then use matplotlib.pyplot.bar to plot them; what does it suggest?
7. Retrieve the value of the second attribute of row 28 and assign it to a variable called *message*:
   1. what is its data type? How many characters does message have?
   2. Using a regex, replace all the punctuation in *message* with a space; to get all punctuation you can use the string module; check that the punctuation has been replaced in *message*.
8. Filter all the rows that are spam and put them into another dataframe called *spam\_df* ; do the same with ham.
   1. Why do you think we might want to isolate the spam data?
9. Retrieve the values from the second column of the spam\_df and put them into a separate list using the list built-in function; do the same with ham.
10. Replace using a regex any digit sequence from each message with the word “DIGITS”.
    1. Why do you think we might want to do so?
    2. Create a function that returns the percentage of the occurrence of a given token/word in a list of texts. Use the nltk.word\_tokenize to break down each message into words.
    3. Apply the function with ‘DIGITS’ for both lists of emails, obtain the percentage of frequency of DIGITS in each, and then plot them together using a bar chart. What do you observe?
    4. Do the same as above but for ‘FREE’.