Final project: spam detection

You are provided with a data set spam.csv. You should use this data set to train three classifiers: a Bayesian Classifier (BC) and two versions of Neural Network (NN). The data set is a group of SMS messages[1] classified into two categories: spam (undesired messages) and ham (useful messages). Below are the steps of your work.

I. **Pre-processing:** this step consists of doing a statistical analysis of the data and data preparation to help the classifier do an optimal classification work.

1. **Stop Words:** you should keep the most informative words and remove the less informative ones (called stop words). Stop words are words that are believed to have a small contribution to the meaning of the text like *the, is, he,* etc. A simple way to remove stop words from a text consists of using the NLTK library. More details on how to do that are provided in the link below[2].

2. **Word Stemming:** stemming consists of removing morphological extensions of the words such as the marks of plural or verb ending (e.g. stem of books is book, stem of ending is end). The goal of stemming is to reduce the variation of the words and find similarity between them. For example, for the computer, the words *thinks*, and *think* are two different words: by stemming these two words, we can have the computer see that they are the same word. To stem the words of the text you can also use the *word_tokenize* function from NLTK library[3].

3. **Word Count:** you should write a program that can provide the count of every word in the data set and their rank (the word that is most frequent has the rank 1, the second most frequent word's rank is 2, etc.).

4. **Frequent Words Identification:** Identify the 50 most frequent words, after the stemming and stop word removal, that are present in the spam class only as well as the most frequent words that are present in the ham class only.

5. **Outliers Identification:** using the statistical method and the word frequencies, identify the outlier words. Try to explain why some words have much higher frequency or much lower one than the

[1] https://www.kaggle.com/c/spam-detection/
[2] https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
[3] https://pythonspot.com/nltk-stemming/

others do. You should use the entire word list before removing stop words to get the words counts.

II. **Bayesian classification:** you should clean your text from stop words and stem all the words to reduce diversity. See slides 29-51 Bayesian classification.

III. **Neural Network classification:** you should build two neural networks. The first one takes as input the 50 most common words in the spam category (input layer size is 50). Your input is a vector of binary cells. The value of the cell is 1 if the word is present in the text and 0 if the word is not. You should define the architecture of the NN accordingly (among others, you need to decide how may neurons you should have in the hidden layer and how many in the output layer). What you need to do in this case is to scan your text searching for the words that are in the top 50 spam words and populate your input vector. The second NN takes in addition to the top 50 spam words a list of 50 most common words not in the spam category (that were only used in the ham category). Please note that state of the art NN approaches to text classification use vector representation of the words called *word embedding* and use more advanced NN such as convolutional NN. The approach adopted here is just for practicing basic NN with the task of text classification.

IV. **Evaluation:** you should split the data set into 70-30 and 80-20 for training and testing respectively and provide the results of three implemented approaches (Bayes classification and the two Neural Networks). The evaluation consists of counting the number of correct classifications over the total number of classifications. You should conduct two types of evaluations: one using the entire data set (training + testing), second test data only. Visualize your results using an appropriate visualization chart type (e.g. bars, pie, line). Use pyplotlib preferably. Excel is accepted.

V. **Report:** write a report describing your work. It should has the following parts.

1. **Introduction:** the problem of spam detection, its importance and how it was addressed and how you plan to address it with BC and NN (brief presentation). You may do some background search to get some more information.
2. **"Our System":** You describe thoroughly your implemented algorithms (BC and the two NN versions). Some diagrams are welcome to explain your system and the NN architectures.
3. **Evaluation and Results:** you describe the evaluation methodology adopted and present your results. You should provide a brief discussion of your results (e.g. which technique does a better

job and why). Your measure is the number of correct classification divided by the total number of cases.

4. **Conclusion:** where you summarize your work and results and you describe what you could do in the future to improve your work.

VI. **Oral presentation:** you should present your work orally Friday May 1st and Monday 4th (We will discuss which teams go first).

VII. Friday 24th I will have a progress checking meetings with you. I will have meetings of about brief 10 minutes with every team to see where you are and answer your questions about the project.