Spam Detection for Text Messages

Text Message (SMS) Spam are unsolicited text messages (SMS), especially advertising, directed at mobile phones or smartphones. As the popularity of mobile phones surged in the early 2000s, frequent users of text messaging began to see an increase in the number of unsolicited (and generally unwanted) commercial advertisements being sent to their telephones through text messaging (SMS). This can be particularly annoying for the recipient because, unlike in email, some recipients may be charged a fee for every message received, including the spam messages. Hence, the problem.

In this assignment, we ask you to complete the analysis of what type of text messages are likely to be spam. In particular, we ask you to apply the tools of machine learning to predict which messages in a corpus are spam. You may treat it as a classic case-study of Binary Classification on the dataset of SMS Messages (Texts) provided to you with this assignment.

Problem Statement : Classification

Use as training set the labeled (good/spam) text messages available in the corresponding "**smsdata.txt**" file to build a robust tree-based binary classifier that is capable of distinguishing spam text messages from regular ones. The tree-based binary classifier that you build may be a single decision tree or an ensemble (forest), whichever is better in this case.

Dataset : Labeled Text Messages

Note that the training set of labeled text messages in the corresponding "**smsdata.txt**" file is structured as follows, where the first element is the label, either good or spam, and then the text message is posted in the raw text format.

```
good Go until jurong point, crazy.. Available only in bugis n great world la e buffet...
good Ok lar... Joking wif u oni...
spam Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005.
good U dun say so early hor... U c already then say...
good Nah I don't think he goes to usf, he lives around here though
spam FreeMsg Hey there darling it's been 3 week's now and no word back!
good Even my brother is not like to speak with me. They treat me like aids patent.
```

Primary Challenge: Note that the raw text format of the Text Messages (SMS) in the dataset have no specific "features" or "variables" for you to use as predictors. Thus, your primary challenge is to engineer important "features" in this case. Try extracting as many "features" from the dataset as you like and use them to classify the response "good" or "spam". You may feel free to talk to your classmates or the TAs or the Instructor regarding your choice of "appropriate" features.

Workflow and Submission Guide

1. Download the data file "**smsdata.txt**" posted corresponding to this assignment and store it in a dedicated folder
2. Download the starter file "**starterfile.ipynb**" posted corresponding to this assignment and store in same folder
3. Check the code written in the starter file to get an idea on how to work with text documents like "**smsdata.txt**"
4. Create a new Jupyter Notebook, name it "assignment_MatID.ipynb", where MatID is your NTU Matriculation ID
5. Solve the Problem posted above by writing code and corresponding comments within "assignment_MatID.ipynb"
6. Submit ONLY the finished notebook "assignment_MatID.ipynb" to NTU Learn, as you do for any other assignment
7. Mention all online/offline sources you referred to, and list all online reference links at the end of your notebook