**Classify SMS Text messages Using Natural Language Processing**

**Objective:**

This project aims in implementing methods for spam filtering in Python, on a dataset containing SMS data. This project uses Python 3 and Jupyter Notebook.

It uses dataset of 5,572 labeled SMS, and a Random Forest Model for classification. It is developed in Jupyter notebook. Code for this project can be found in 'SMS classification (NLP).ipynb'.

An overall accuracy of 99% has been achieved.

# Required Software:

We will need scikit-learn; pandas and their dependencies, already installed using pip. The coding part is done using Jupyter notebook and all the necessary modules needed for this project are available there.

# Detailed Outline:

Reading a text based dataset into pandas

Representing text as numerical data

Vectorizing dataset

Model building in scikit-learn

Evaluating a model

Comparing models

# References:

<https://github.com/udacity/machinelearning/blob/master/projects/capstone/capstone_report_template.md>

<https://github.com/udacity/machine-learning/blob/master/projects/capstone/report-example-1.pdf>

<http://gim.unmc.edu/dxtests/ROC2.htm>

<http://scikit-learn.org/stable/modules/pipeline.html>

<https://en.wikipedia.org/wiki/Natural_language_processing>

<https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words>

<http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html>

<https://github.com/udacity/machinelearning/blob/master/projects/capstone/capstone_proposal_template.md>

<http://www.nltk.org/book/>

<https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection>