**Text Mining and Sentiment Analysis using Python**

1. Many mobile applications provide a way to screen incoming SMS as spam. With 95% accuracy the company is able to classify sms as spam or ham. And there is always space for improvement. Your task is to use any of the machine learning models to improve the accuracy of the model to above 96%
   1. Use encoding=’cp1252’ while reading the data (spam.csv) using pandas, since this data set contains other languages
   2. Make sure that you split the data in to train (70%) and test (30%). Use random\_state=100, while using train\_test\_split function to set the seed
   3. Print the following metrics
      1. Accuracy, sensitivity, specificity, F1 score

Product companies are always interested in identifying negative feedback and responding back to customers immediately to create a positive impact. But the problem is to identify such negative reviews from many reviews. They receive close to 5000 reviews each day in twitter. It takes lot of manual effort to manually identify the sentiment of each tweet. Your job is to help them using the following set of analysis

1. Bag of word analysis
   1. Read the labelled dataset (sentiment\_amazon.csv). Positive sentiment are encoded as 1s and negative sentiments are encoded as 0s
   2. Create a word cloud using positive reviews alone and identify important terms that are frequently appearing
   3. Create a word cloud using negative reviews alone and identify important terms that are frequently appearing. Identify some products which are very frequently appearing in negative reviews
2. Without cleaning the dataset, identify sentiment for each review using VADER package in nltk. This is unsupervised prediction, hence you will not be using the labelled column
3. Apply Random Forest or Adaboost model to treat this as an supervised model by using the labelled column
4. Using KMeans algorithm cluster the reviews in to two groups. Create wordclouds separately for each cluster. Using the word cloud, see if you can identify which cluster could be for positive and negative reviews
5. Compare the accuracy of supervised and unsupervised using VADER, unsupervised using KMeans, and comment on which one gives the best result