\*\* EMAIL CLASSIFIER USING LINEAR SUPPORT VECTOR MACHINE \*\*

1)-- Email preprocessing --

To classify an email first the email is needed to be converted into a input feature vector.

An email header consists of body and sender reciever details, all we need is the body of mail.

The body of an email is a paragraph consisting of sentences which can not directly be used to classify.

We find some key words in a mail body that may hint towards a spam or not spam mail.

THE IMPORTANT PREPROCESSING INVOLVES:

a) Lower-Casing: every character is converted to lower case.

b) Stripping HTML: Generally every mail consists of some links to some web pages. We remove the html tags as they add no value

to SVM classifier.

c) Normalizing URLs: All URLs are replaced with text 'httpaddr'.

d) Normalizing email addresses: all email addressses are replaced with 'emailaddr'.

e) Normalizing Numbers: all the number are replaced with 'number'.

f) Normalizing currency chars: $->'dollar', ₹->'rupee'.

g) Word Stemming: words are converted to stemmed form. extra chars of a word are cuttoff.

Eg: (Offer, Offers, Offering, Offerings)->'offer'

h) Removal of non-words: Non-words like punctuation, multiple blank space, tab, new line etc are replaced with single space.

VOCABULARY LIST:

Now in our processed mail we choose the words to take as input feature.To prevent our model from over-fitting we pick only the most occuring words.

\*vocab.txt\* file contains