***THE ALGORITHM :*  
  
1. Preprocess or normalize each email**  
 While many emails contain similar types of entities ,like numbers, URLs, or email addresses), the specific entities will be different in almost every email. Thus, one method often used in processing emails is to “normalize” these values, so that all URLs , all numbers are treated likewise.   
 1.1 Convert entire email to lowercase  
 1.2 Strip all HTML tags  
 1.3 Normalize URLs  
 1.4 Normalize email addresses  
 1.5 Normalize numbers  
 1.6 Remove non-alphanumeric characters

**2. Create word index for each email**

With the help of vocabulary list, we map each word in the preprocessed emails into a list of word indices which contains the index of the word present in the vocabulary list.  
  
**3. Construct feature vector**

Now we can implement the process of 'feature extraction', that converts each email into a vector in R n , where n=number of words in vocabulary list. So, the feature x i ∈ {0, 1} for an email corresponds to the fact that whether the i-th word in the dictionary occurs in the email or not. Thus, x i = 1 if the i-th word is in the email and x i = 0 if the i-th word is not present in the email.  
  
**4. Train SVM for spam classification algorithm** We have divided our dataset into training set and test set. The training set has 4000 examples of emails while the test set contains 1000 examples of spam and non spam emails .Each email was pre-processed or normalized using the 'processEmail' and and converted into a vector x (i) ∈ R 1899 using 'emailFeatures' function.  
We run our SMO algorithm on our training set. Then we make observations on our test set. We calculate mean (accuracy) for both.