Assignment: Build a SPAM/HAM Classifier using Naïve Bayes Classification

Problem Statement

We have a message m = (w*1*, w*2*, ... , w*n*), where (w*1*, w*2*, ... , w*n*) is a set of unique words contained in the message. We need to find

$$P(spam|w1 \cap w2\cap \ldots \cap wn) = \frac{P(w1 \cap w2\cap \ldots \cap wn|spam).P(spam)}{P(w1 \cap w2\cap \ldots \cap wn)}$$

If we assume that occurrence of a word is independent of all other words, we can simplify the above expression to

$$\frac{P(w1|spam).P(w2|spam)\ldots P(wn|spam).P(spam)}{P(w1).P(w2)..P(wn)}$$

In order to classify we have to determine which is greater

$$P(spam|w1 \cap w2\cap \ldots \cap wn)\ versus\ P(\sim spam|w1 \cap w2\cap \ldots \cap wn)$$

Sample of the dataset to be used:

Class (Spam/ Ham)	Mail Content
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune

spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030

Steps to be included:

1. Preprocess the data.
   a) Remove repeated entries if exists
   b) Remove punctuation
   c) Remove not relevant special characters
   d) Obtain a list of clean text words
2. Split train and test dataset
3. Create the model (convert strings to integer counts, obtain frequency of the words and create the model using Naïve Bayes formula).
4. Test the model.
5. Create a confusion matrix on your prediction for test dataset.
6. Make analysis report for your model and results.