Analytical Component

Problem 1

1. For finding prior probabilities , we have 3 spam emails and 2 ham emails.
   So P(spam) = 3/5
   P(ham) = 2/5
2. For spam emails
   buy – 1
   car -1
   Nigeria – 2
   profit -2
   money-1
   home-1
   bank-2
   check-1
   wire-1

   Total words = 12

   P(buy | spam) = P(car | spam) = P(money | spam) = P(home | spam) = P(check | spam) = P(wire | spam) = 1/12

   P(Nigeria | spam) = P(profit | spam) = P(bank | spam) = 2/12 = 1/6

   For ham emails
   money -1
   bank-1
   home-2
   car-1
   Nigeria-1
   fly-1

   Total words = 7

   P(money | ham) = P(bank | ham) = P(car | ham) = P(Nigeria | ham) = P(fly | ham) = 1/7
   P(home | ham) = 2/7

3. For Nigeria
   P(spam | Nigeria) = (P(Nigeria | Spam) . P (spam))/ P(Nigeria)

   Since the denominator will remain constant in both cases of ham and spam, we are ignoring the denominator during our following calculations.

P(spam | Nigeria) = P(Nigeria | Spam) . P (spam)
= 2/12 * 3/5
= 0.1

P(ham | Nigeria) = P(Nigeria | ham) . P (ham)
= 1/7 * 2/5
= 0.0571

Since P(spam | Nigeria) is greater, predicted class = spam

For Nigeria home
P(spam | Nigeria home) = P(Nigeria home | spam) P(spam)
= P(Nigeria | spam) . P(home | spam) P(spam) [assuming conditional independence of words]
= 2/12 * 1/7 * 3/5
= 0.01428

P(ham | Nigeria home) = P(Nigeria home | ham) P(ham)
= P(Nigeria | ham) . P(home | ham) P(ham)
= 1/7 * 2/7 * 2/5
= 0.01632
So, predicted class = ham

For home bank money
P(spam | home bank money)
= P(home | spam) . P(bank | spam) . P(money | spam) . P(spam)
=1/12 *2/12 *1/12 * 3/5
= 0.000694

P(ham | home bank money)
= P(home | ham) . P(bank | ham) . P(money | ham) . P(ham)
=1/7 * 1/7 * 2/7* 2/5
= 0.00233
So, predicted class = ham

Problem 2

Let P(n) denote sum of probabilities of all sentences of length n.

For P(1)

Let vocab = {a,b,c}, then sentences are -

START a (probability is 1/3)

START b (probability is 1/3)

START c (probability is 1/3)

So total sum is 1. So P(1) = 1

Lets take generalized case of all sentences of length n-1.

Our vocab is {a,b,c}

Lets say we have k possible sentences of length n-1.

Probability of each sentence (using bigram model) = 1/k (as we can see from n=1, and n=2)

Now if we move to all sentence of length n.

Our total no. of sentences = 3k (since our vocab size is 3)

In each of sentences , the probability product upto n-1th word = 1/k

And the probability of last word $P(w_n \mid w_{n-1}) = 1/3$

So, probability of the entire sentence = 1/k * 1/3 = 1/3k

So sum of probability of all sentences of length n = P(n) = 1/3k * 3k = 1

Hence proved, $\sum P(w_1, w_2, \ldots w_n) = 1$