Name: Huiyu Song
UNI: hs3160

HW1

Problem 1

1.

$$P(spam) = \frac{3}{5}$$
$$P(ham) = \frac{2}{5}$$

words ‖ *class*	class		words ‖ *class*	class	
words	spam	ham	words	spam	ham
buy	1/12	0	home	1/12	2/7
car	1/12	1/7	bank	2/12	1/7
Nigeria	2/12	1/7	check	1/12	0
profit	2/12	0	wire	1/12	0
money	1/12	1/7	fly	0	1/7

2.

3.

$$\because label = \underset{label}{\operatorname{argmax}} P(label) \prod P(x_i|label)$$

•Nigeria

$$P(spam|Nigeria) \propto P(spam)P(Nigeria|spam) = \frac{3}{5} \cdot \frac{2}{12} = 0.1$$
$$P(ham|Nigeria) \propto P(ham)P(Nigeria|ham) = \frac{2}{5} \cdot \frac{1}{7} = 0.0571$$
$$P(spam|Nigeria) > P(ham|Nigeria)$$

Therefore, it belongs to spam.

•Nigeria home

$$P(spam|Nigeria\ home) \propto P(spam)P(Nigeria|spam)P(home|spam) = \frac{3}{5} \cdot \frac{2}{12} \cdot \frac{1}{12} = 0.0083$$
$$P(ham|Nigeria\ home) \propto P(ham)P(Nigeria|ham)P(home|ham) = \frac{2}{5} \cdot \frac{1}{7} \cdot \frac{2}{7} = 0.01632$$

Therefore, it belongs to ham.

•home bank money

$P(spam|home\ bank\ money) \propto P(spam)P(home|spam)P(bank|spam)P(money|spam) = 0.000694$

$P(ham|home\ bank\ money) \propto P(ham)P(home|ham)P(bank|ham)P(money|ham) = 0.00233$

Therefore, it belongs to ham.

Solution to problem 2

1. **Base Case**: If there is only one word in all sentence, because vocabulary size is V, therefore $\sum_{i=1}^{V} P(w_i|START) = 1$

   Assumption:
   Assume the number of all sentences with n words is N.
   Assume for all sentences with k words, this equation holds:

$$\sum_{w_1,w_2,\cdots,w_k} P(w_1, w_2, \cdots, w_k) = \sum_{w_1,w_2,\cdots,w_k} P(w_1|START)P(w_2|w_1)\cdots P(w_k|w_{k-1}) = 1 \quad (1)$$

   Induction Steps:
   For all sentence with (k+1) words, they all equals to adding one word to one of the sentences with k words. Therefore the probability of all sentences with (k+1) words is:

$$\begin{aligned}
\sum_{w_1,w_2,\cdots,w_{k+1}} P(w_1, \cdots, w_{k+1}) &= \sum_{w_1,w_2,\cdots,w_{k+1}} P(w_1|START)P(w_2|w_1)\cdots P(w_{k+1}|w_k) && (2)\\
&= \sum_{w_1,w_2,\cdots,(w_k,w_{k+1})} P(w_1|START)P(w_2|w_1)\cdots P(w_{k+1}|w_k) && (3)\\
&= \sum_{w_1,w_2,\cdots,w_k} \sum_{(w_k,w_{k+1})} P(w_1|START)\cdots P(w_k|w_{k-1})P(w_k|w_{k+1}) && (4)\\
&= \sum_{w_1,w_2,\cdots,w_k} P(w_1|START)\cdots P(w_k|w_{k-1}) \sum_{(w_k,w_{k+1})} P(w_k|w_{k+1}) && (5)\\
&= 1 \cdot \sum_{(w_k,w_{k+1})} P(w_k|w_{k+1}) && (6)\\
&= 1 \cdot \sum_{(w_k,w_{k+1})} \frac{Count(w_k, w_{k+1})}{Count(w_k)} && (7)\\
&= 1 && (8)
\end{aligned}$$

   Therefore, the assumption is correct.