IST 441 - Assignment 5 - Due 4/13

In this assignment you will be classifying documents using Naïve Bayes classifier. You should read chapter 13 in Introduction to Information Retrieval, and get familiar with examples 13.1 and 13.2

Problem 1

Given the following table that contains training documents that you will use in building a Naïve Bayes classifier, each document is classified into being either about Apple or not.

docID	Words in document	C = Apple?
1	Steve Jobs CEO Apple Apple	Yes
2	Google Apple Android iOS tablet	Yes
3	Microsoft tablet billion	No
4	Apple cider	No
5	Ipad tablet iphone smartphone	Yes
6	Apple computer operating system	Yes

1. Estimate the multinomial parameters of the model
2. Estimate the Bernoulli parameters of the model
3. Given the following documents, estimate whether they should be classified to be about Apple or not, and justify your answer using both the multinomial and the Bernoulli models

docID	Words in document	C = Apple?
7	google jobs	?
8	Microsoft operating system	?
9	Apple Apple Apple tablet microsoft	?

Problem 2

You are asked to design a classifier that identifies spam tweets on twitter. You may choose a naïve Bayes classifier or any other classifier of your choice. If you choose a Naïve Bayes classifier, please mention whether you would choose the Bernoulli or the multinomial model, and why? If you choose a different classifier, please motivate your choice.
What features would you use in building the classifier?