$$P(C|F_1, \ldots F_n) = \frac{P(C)\,P(F_1, \ldots, F_n|C)}{P(F_1, \ldots F_n)}$$

a) $P(X_{francisco} = true \mid class = SFO) = 2/2 = \underline{1}$

$P(X_{london} = true \mid class = SFO) = 1/2 = \underline{0.5}$

$P(X_{francisco} = true \mid class = JFK) = 1/1 = \underline{1}$

b) $P(X = francisco \mid class = SFO) = 4/14 = \underline{2/7}$

(Assuming no tokenization)

$P(X = london \mid class = SFO) = 1/14 = \underline{1/14}$

$P(X = francisco \mid class = JFK) = \underline{1/8}$

c) i) The Bernoulli model is not very accurate because it ignores frequency information which is important in this domain.

ii) The multinomial model is more accurate because it uses frequency information. However, it ignores position information, so does'nt distinguish between a city name occuring at the beginning/end of the itinerary from one occuring in the middle.

d)

d) We can use as a feature the term that occurs in the last position of each document.
Non standard feature represented with using non-standard words. The non-standard words are classified to 6 categories. using SKIPEZ collection to official, literature, informative, popular, educative and scientific.

$$P(X_{NewYork} = true \mid class = JFK) = 1.0$$
$$P(X_{sanFrancisco} = true \mid Class = SFO) = 1.0$$
$$P(X_{chicago} = true \mid class = ORD) = 1.0$$

2a) It will never choose a category unless all words in a document were seen for that category for the training set. It will rank between classes for which all words we have seen similarity to the smoothed classifier

b) Here it is given that they have doubled the amount of smoothing.

Laplace (add-1) smoothing for Naive Bayes $\Rightarrow$

$$P(W_i|c) = \frac{Count(W_i, c) + 1}{\sum (Count(w, c) + 1)}$$

$$= Count(W_i, c) + 1 \Big/ \sum_{w \in V} Count(W, c) + |V|$$

It will be more likely to choose categories for which some/many of the words in the document were unseen.

III. Given that

System returns 3 relevant documents
2 irrelevant documents
Total 8 Relevant documents in the collection

a) $\text{Precision} = \frac{TP}{TP+FP} = \frac{3}{3+2} = 3/5$

b) $\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+5} = 3/8.$

b) i) An IR system which always returns no results will have high accuracy for most queries, since the corpus usually contains only a few relevant documents. Documents that are truly relevant are the only ones that will be mistakenly classified as non-relevant. and thus the accuracy is close to 1. Recall and precision are two different measures that can jointly capture the trade off b/w returning more relevant results and returning fewer irrelevant results.

ii) There are of course many correct answers. One simple correct answer is

Assume document 1 is the only relevant document.

$$A_q = \{1, 2, 3\}$$
$$B_q = \{3\}$$

Both A_q & B_q made 2 mistakes. So they have same accuracy 80%.

The precision of A_q is 1/3, the precision for B_q is 0. Since B_q did not return any relevant documents, it is of no utility.