Supervision questions: set 1.

Sentiment lexicon

1. (To be done with other members of your supervision group) Each find a short piece of text (100 words or less) expressing an opinion about something, without showing it to the others. This could be a review, but don't use a movie review or similar. Tokenize and then sort the words so they are in alphabical order, removing duplicates. Now swap lists. Mark each word in the list you've been given as positive or negative sentiment and say whether you think the piece the words have come from is overall positive or negative. Compare your answer with the original text. Did you get the overall sentiment right? Were the words used in the way you thought they might be?

   Does your Task 1 system get this right? What about your Task 2 system?

2. What sort of words change the polarity of the sentiment words? *not* is an obvious example: can you think of 10 others? Are there any examples in the text you looked at in 1? Which words in a sentence can have their sentiment flipped if there's a *not* in the sentence?

3. Try looking at some social media posts and work out whether you could find words which indicated different types of sentiment: e.g., could you use a lexicon to classify posts according to how emotionally involved someone was feeling?

4. In a test set with 412 examples, 328 are correctly classified. What is the accuracy?

5. Why is accuracy not necessarily a good measure of success if the classes have very different probabilities?

Naive Bayes

1. Suppose that you are using Naive Bayes on a task where you have 100 documents in a training set, which is equally divided between class A and class B. There are three features F1, F2 and F3: each may occur at most once in a document. (Note that the set up here is a little different from the way we used NB in Task 2.) The distribution for the three features among documents is as follows:

	A	B
F1	5	5
F2	0	10
F3	3	27

a. Show the estimated conditional probabilities for each class, given each feature.

b. Assume that you are trying to classify a document which contains only the features F1 and F3: how would you estimate the relative probability of A and B (without add-one smoothing)?

c. What difference would it make if there were 25 documents in class A in the training set and 75 in class B?

d. Which of the features F1, F2 and F3 would be more useful for classification in general? Explain your answer.

e. Given reasonable amounts of training and test data and a feature set with 10 features, how could you establish which features were most useful?

2. (Difficult) The approach we asked you to take for NB incorporated probabilities for all the positions in the document (i.e., all the **tokens**). An alternative approach, often used for document classification, is to count words only once no matter how many times they appear in a document. This model is clearly less informative than the approach we used, but it usually works better. Why?

Statistical properties of language

1. Given that we will always see new words given a sufficiently large corpus, how is it that most people would confidently say that the following are not English words: *pferd*, *abtruce*, *Kx'a*. Are they right?

More to follow