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?

   The words Ben gave me were:

word	polarity	word	polarity	word	polarity
!		's		(	
)		,		.	
a		advanced	positive	algorithms	positive
also		am		an	
and		are		as	
at		beginner-friendly	positive	being	
bought	negative	companion	positive	concepts	positive
consider	positive	course	positive	covered	positive
definitely	positive	despite	negative	ece358	
especially	positive	everything	positive	for	
go		great	positive	hardcover	positive
heavy	negative	helpful	positive	i	
if		includes	positive	introductory	positive
is		it		large	negative
like	positive	loving	positive	maybe	negative
myself	negative	not	negative	of	
overall	negative	pseudocode	positive	quite	negative
recommend	positive	said		snippets	positive
some	negative	still	negative	textbook	
that		the		this	
uoft		very		visualization	positive
would		you			

   I will not be associating polarities to tokens such as "!", "'s" or "(" which are not words but punctuation – nor will I be assigning polarities to words such as "a", "and" or "an" which are grammatical necessities and so can convey little-to-no opinion whatsoever.

   There are 22 positive words but only 11 negative words. I will therefore conclude that this text is most likely a positive review.
   A more human reason for this would be that most of the positive words are stronger than the negative words – most of the negative words are words which I interpret to be negative but could also be used in positive contexts – ie "heavy": is the reviewer complaining the book is physically heavy, is a heavy read or saying there is a heavy emphasis on algorithmic proof?

   The full text was:

   "I bought this as a course companion for ECE358 at UofT (an algorithms course) and am loving it. The pseudocode snippets for everything is very helpful and it also includes visualization of algorithms. Despite being an introductory textbook, some of the concepts covered are quite advanced and maybe not as beginner-friendly. That being said, it's still a great algorithms textbook overall and I would definitely recommend it! Also consider that it is a very large and heavy textbook, especially if you go for hardcover like myself."

This was a 4-star review to *Introduction to Algorithms*. A positive review. This has the overall sentiment of what I expected. However – not everything was used as I anticipated.
For example I saw "beginner-friendly" and introductory and assumed they would be used in conjunction rather than disjunction: ie "this textbook is beginner-friendly and introductory".

Both of my tick1 predictors (the sentiment analyser with a cutoff at zero and the one with a nonzero cutoff) predicted that the sentiment was positive. The lexicon agreed with me that there were significantly more positive words than negative. This makes sense since the lexicon was not trained on film reviews and so is generalisable to other sources.

However, both of my tick2 predictors (unsmoothed and smoothed) predicted that the sentiment was negative. This was surprising but on more thought also makes sense – they were trained on film reviews which have a totally different set of interpretations to an advanced algorithms textbook: for example in film review some words (such as "introductory") have different polarities and many of the positive words simply will never turn up – "advanced", "hardcover", "beginner-friendly".

This shows two things:

- Predictors are only usable on the type of data that they were trained on
- Predictors that are better in one situation are not always better than other predictors in every situation. Shows by the tick1 lexicon predictor (which achieved no higher than a 60% accuracy on the film reviews) getting the correct result while the Naïve Bayes predictor which achieved 82% got the incorrect result.

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?

   "But" and "however" negate the sentiment of the previous sentence.

   In other cases words can change the polarity – for example "would", "wanted", "could", "maybe" sometimes change the polarity of subsequent words ie "I wanted a textbook filled with interesting algorithms and examples", "maybe this is more useful for beginners".

   Most positive words have their sentiment flipped when negated (with *not*) but many negative words do not have their sentiment flipped. For example "It's a bad book" is obviously negative, but "It's not a bad book" is still a negative statement since the implication is that it is not a good book either.

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?

   You could – but it wouldn't be as effective since "emotional involvement" is not binary and so the data would be more varied – and especially given the broad range of topics which people discuss on social media the lexicon required would be vast. Training data about emotional envolvement in politics or sport could not be applied to emotional envolvement in breakups for example. So in summary: yes you could but it would take a lot of effort and at the end of it still wouldn't be particularly good unless you greatly narrowed down the criteria and only analysed one very particular type of social media post.

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

   Accuracy is the total correct predictions divided by the total number of predictions.

$$A = \frac{328}{412}$$
$$A = 0.796 \text{ to 3.S.F} \tag{1}$$

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

   If one class has a significantly higher probability of occuring than another then a predictor may have a negligible (or zero) probability of choosing the smaller class – and so be near-useless in most applications (†) – however this predictor may still have a very high accuracy.

   Often better metrics to use are recall and precision (ideally a combination of both)– recall is the proportion of false negatives and precision is the proportion of false positives.

   A simple but powerful metric to use is the F-measure and is a weighted geometric average of recall and precision.

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{2}$$

   A higher β weights recall higher and a lower beta weights precision higher. β should be chosen dependent on the relative cost of false-positives and false-negatives.

   Take a cancer screening test. Assume that there is a very low prevelance of this cancer and only 1 in 1000 people have it. If the test doesn't work and always gives negative results then it has a 99.9% accuracy. Using accuracy as a metric we would conclude that this test is very good – despite the fact that it never says anyone has cancer and is worse than useless.

   In this specific case recall would be more important – since the cost of a false negative is far higher then the cost of a false positive. So an appropriate metric would be F_β with a very high β.

   (†) The main exceptions to "low probabilities of one class" being bad are where false-positives must never happen or where there is a secondary predictor that is far more expensive to run so you wish to run a cheaper predictor to deal with the "obvious" cases with and make the whole system faster.

2 Naïve Bayes

1. (a) 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

(b)

$$\begin{aligned}
\hat{\mathbb{P}}(A|F1) &= \frac{A \wedge F1}{\sum_{c \in \{A,B\}} c \wedge F1} \\
&= \frac{5}{10} \\
&= 0.5 \\
\hat{\mathbb{P}}(A|F2) &= \frac{A \wedge F2}{\sum_{c \in \{A,B\}} c \wedge F2} \\
&= \frac{0}{10} \\
&= 0 \\
\hat{\mathbb{P}}(A|F3) &= \frac{A \wedge F2}{\sum_{c \in \{A,B\}} c \wedge F3} \\
&= \frac{3}{30} \\
&= 0.1
\end{aligned} \tag{3}$$

$$\begin{aligned}
\hat{\mathbb{P}}(B|F1) &= \frac{B \wedge F1}{\sum_{c \in \{A,B\}} c \wedge F1} \\
&= \frac{5}{10} \\
&= 0.5 \\
\hat{\mathbb{P}}(B|F2) &= \frac{B \wedge F2}{\sum_{c \in \{A,B\}} c \wedge F2} \\
&= \frac{10}{10} \\
&= 1.0 \\
\hat{\mathbb{P}}(B|F3) &= \frac{B \wedge F2}{\sum_{c \in \{A,B\}} c \wedge F3} \\
&= \frac{27}{30} \\
&= 0.9
\end{aligned} \tag{4}$$

(c) 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)?

$$\begin{aligned}
\hat{\mathbb{P}}(A|(F1 \wedge F3)) &\approx \frac{\hat{\mathbb{P}}(A) \times \hat{\mathbb{P}}(F1|A) \times \hat{\mathbb{P}}(F3|A)}{\hat{\mathbb{P}}(F1 \wedge F3)} \\
&= \frac{\frac{1}{2} \times \frac{5}{50} \times \frac{3}{50}}{\frac{1}{10} \times \frac{3}{10}} \\
&= \frac{1}{10} \\
&= 0.1 \\
\hat{\mathbb{P}}(B|(F1 \wedge F3)) &\approx \frac{\hat{\mathbb{P}}(B) \times \hat{\mathbb{P}}F1|B \times \hat{\mathbb{P}}(F3|B)}{\hat{\mathbb{P}}(F1 \wedge F3)} \\
&= \frac{\frac{1}{2} \times \frac{5}{50} \times \frac{27}{50}}{\frac{1}{10} \times \frac{3}{10}} \\
&= \frac{9}{10} \\
&= 0.9
\end{aligned} \tag{5}$$

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

This would change $\hat{\mathbb{P}}(A)$ from 0.5 to 0.25 and $\hat{\mathbb{P}}(B)$ from 0.5 to 0.75.
However it would also double the probabilities $\hat{\mathbb{P}}(F1|A)$, $\hat{\mathbb{P}}(F2|A)$ and $\hat{\mathbb{P}}(F3|A)$ (since the observed number is the same while the number of observations halved). $\hat{\mathbb{P}}(F1|B)$, $\hat{\mathbb{P}}(F2|B)$ and $\hat{\mathbb{P}}(F3|B)$ would decrease by $\frac{1}{3}$ since there were now more observations of B.

With these new probabilities:

$$\begin{aligned}
\hat{\mathbb{P}}(A|(F1 \wedge F3)) &\approx \frac{\hat{\mathbb{P}}(A) \times \hat{\mathbb{P}}(F1|A) \times \hat{\mathbb{P}}(F3|A)}{\hat{\mathbb{P}}(F1 \wedge F3)} \\
&= \frac{\frac{1}{4} \times \frac{5}{25} \times \frac{3}{25}}{\frac{1}{10} \times \frac{3}{10}} \\
&= \frac{1}{5} \\
&= 0.2 \\
\hat{\mathbb{P}}(B|(F1 \wedge F3)) &\approx \frac{\hat{\mathbb{P}}(B) \times \hat{\mathbb{P}}F1|B \times \hat{\mathbb{P}}(F3|B)}{\hat{\mathbb{P}}(F1 \wedge F3)} \\
&= \frac{\frac{3}{4} \times \frac{5}{75} \times \frac{27}{75}}{\frac{1}{10} \times \frac{3}{10}} \\
&= \frac{9}{10} \\
&= 0.9
\end{aligned} \tag{6}$$

So the new $\hat{\mathbb{P}}(A|(F1 \wedge F2))$ is 0.2 and the new $\hat{\mathbb{P}}(B|(F1 \wedge F2))$ is 0.8.

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

F2 and F3 are the features which would be most useful for classification.

F1 is evenly split between both A and B so (assuming the original model in which the classes A and B have the same size), the probabilities $\hat{\mathbb{P}}(A|F1)$ and $\hat{\mathbb{P}}(B|F1)$ are the same. So we cannot tell anything about which class the document belongs to based on F1.

In the observed data, F2 never occurred with A and is uncommon with B. Naïvely, one may think that F2 would be the best indicator of whether a class belonged to A or B. However, due to the small size of the training set we cannot use this feature like this. The probability of observing F1 given A may not be zero. The sample is so small that we can rule very little out and eliminating A completely irrelevant of what the other evidence says is naïve. This is why smoothing is useful – we deal with the case where we do not observe any feature while keeping the probability low.

A truly naïve algorithm (like the unsmoothed one we were forced to use for the first part of tick 2) would keep F2 as a feature on B but not as a feature on A and so multiply B by $\hat{\mathbb{P}}(B|F2)$ and not multiply A therefore decreasing the relative probability of B despite observing a feature uniquely seen in B.

In general F3 is the best indicator – we have a high number of positives and so a good confidence that $\hat{\mathbb{P}}(B|F3) >> \hat{\mathbb{P}}(A|F3)$ and F3 is also common in B so can be used in many instances.

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

To remove ambiguity with the question I will clarify that I am assessing the usefulness of given features for a given (Naïve Bayes) model rather than their overall predictive power.

Unlike many other models; Naïve Bayes has no intrinsic score for the usefulness of a feature. $\hat{\mathbb{P}}(C|F_i)$ is not a good measure since if F_i is a very rare feature than it can have a high $\hat{\mathbb{P}}(C|F_i)$ but very low overall importance. A similar argument applies to using $\hat{\mathbb{P}}(F_i)$ as a measure of importance – F_i could be a very frequent feature with no correllation at all to any class – ie F_i is whether the mm digit of a persons height is odd and the classes are whether the film review they wrote was positive.

For Naïve Bayes the best way of working out feature importance is experimentally. A good way to work this out is called Permutation Importance. You first train and test the data with all features as normal. Record the accuracy (or F-measure or recall or precision – whichever metric you are using) of the full model. Then for each feature: shuffle that one feature randomly and redo the training and testing. Record the resulting drop in metric score. A high decrease means that the feature is more important while a low decrease means that the feature is less important (and an increase means that we have bad data or a bad predictor [like the unsmoothed predictor in tick2]).

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?

   One of the Naïve Bayes assumptions is that the probabilities of different words occuring is independent. This assumption is broken between the same word. For example if one author uses the word "favourite" then they are more likely to reuse it. This is a clear breach of the Naïve Bayes assumption.

   As with many things in machine learning if we do not handle edge cases then we will end up with worse results. A simple way to handle this edge case that breaks our assumption is to simply ignore multiple occurrences of the same word and consider the text as a set-of-words rather than a bag-of-words (only consider a word once irrelevant of how many times it occurs in the text). So although we consider less data and information; the model of the data that we use fits our assumptions and the Naïve Bayes model of text better.

Since our predictive algorithm is now better suited to the data, we can often end up with better results than if we considered the data as it really is even if we have less actual data to process.

3 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?

   Pferd is Horse in German. Abtruce is not a word used for anything. Kx'a is the name of a language group in Africa.

   While none of these words are commonly used – some of them do have meaning – and it is undoubtable that they have all been used at some point. Regulation of *what* words constitute a language is very informal and aims to model accurately what the population actually uses. For example there are words which are simply never used anymore that are sometimes removed from the lexicon even though they are written down and were used for hundreds of years – and there are other words which are always being added. For example "food-baby" and "NFT" were recently added to the dictionary. So one argument is that words are only part of English if they are actively being used in some English-speaking community. This is the approach that most people would take. Under this definition only Kx'a would be an English word – it is the recognised name of a language group and used in linguistics.

   However one could take a more liberal approach. For example in the BBC comedy Blackadder a dictionarian was attempting to write a dictionary of every word. To which Blackadder offered his contrafibularities upon it's completion. A liberal view would say that since this word had been used to convey meaning it was now an English word. Under this interpretation *pferd* and *Kx'a* would be English words since both have undoubtably been used to convey meaning (the former likely among german speakers speaking english) and the latter in linguistics. This is in practice a reasonable approach – any attempt to map the english language will not be completely successful since new words are emerging all the time. Under this interpretation it is almost undoubtable that there are people who have used all three words *pferd*, *abtruce* and *Kx'a* to convey meaning and so all three should be considered as English words.

   In my opinion the latter is a more accurate representation of the reality of English – the language is always evolving and so does not keep a static or even consistent (ie literally – it is its own antonym) lexicon. However like with everything we must make practical modelling assumptions to have a feasible chance of using anything consistently or well and so the first interpretation is more practical.

   My interpretation is that *pferd* is primarily a German word – and not English, *abtruce* is not a word in any mainstream context so should in general not be considerd a word, *Kx'a* is a word.