Natural Language Processing
Lab 3

This lab sheet is to practice the concepts taught this week so far: Naive Bayes classification and Sentiment Analysis.

1. Assume the following likelihoods for each word being part of a positive or negative movie review, and equal prior probabilities for each class.

	POS	NEG
I	0.09	0.16
usually	0.07	0.06
enjoy	0.29	0.06
foreign	0.04	0.15
film	0.08	0.11

What class will a Naive Bayes classifier assign to the sentence "I usually enjoy foreign film"? Show your workings.

Ans:

$$c_{POS} = P(I|POS)P(always|POS)P(like|POS)P(foreign|POS)P(films|POS)P(POS)$$

$$= 0.09 \times 0.07 \times 0.29 \times 0.04 \times 0.08 \times 0.5$$

$$= 0.0000029232$$

$$c_{NEG} = P(I|NEG)P(always|NEG)P(like|NEG)P(foreign|NEG)P(films|NEG)P(NEG)$$

$$= 0.16 \times 0.06 \times 0.06 \times 0.15 \times 0.11 \times 0.5$$

$$= 0.000004752$$

$$\hat{c} = argmax_c(c_{POS}, c_{NEG}) = c_{NEG}$$

It will assign a negative label to the document.

2. Given the following short, feature-selected, movie reviews, each labelled with a genre, either comedy or action:

1. fun, couple, love, love **comedy**
2. fast, furious, shoot **action**
3. couple, fly, fast, fun, fun **comedy**
4. furious, shoot, shoot, fun **action**
5. fly, fast, shoot, love **action**

and a new document D:

- fast, couple, shoot, fly

compute the most likely class for D. Assume a Naive Bayes classifier and use add-1 smoothing for the likelihoods.

Ans: Class probability:

$logprior[comedy] = \log P(comedy) = \log \frac{2}{5} = -0.3979$
$logprior[action] = \log P(action) = \log \frac{3}{5} = -0.2218$

Vocabulary:

$V = \{fun, couple, love, fast, furious, shoot, fly\}$
$bigdoc[comedy] = \{fun, couple, love, fly, fast\}$
$bigdoc[action] = \{fast, furious, shoot, fun, fly, love\}$

Loglikelihoods:

$loglikelihood[fun, comedy] = \frac{3+1}{9+7} = \frac{4}{16}$
$loglikelihood[fun, action] = \frac{1+1}{11+7} = \frac{3}{18}$
$loglikelihood[couple, comedy] = \frac{2+1}{9+7} = \frac{3}{16}$
$loglikelihood[couple, action] = \frac{0+1}{11+7} = \frac{1}{18}$
$loglikelihood[love, comedy] = \frac{2+1}{9+7} = \frac{3}{16}$
$loglikelihood[love, action] = \frac{1+1}{11+7} = \frac{2}{18}$
$loglikelihood[fast, comedy] = \frac{1+1}{9+7} = \frac{2}{16}$
$loglikelihood[fast, action] = \frac{2+1}{11+7} = \frac{3}{18}$
$loglikelihood[furious, comedy] = \frac{0+1}{9+7} = \frac{1}{16}$
$loglikelihood[furious, action] = \frac{2+1}{11+7} = \frac{3}{18}$
$loglikelihood[shoot, comedy] = \frac{0+1}{9+7} = \frac{1}{16}$
$loglikelihood[shoot, action] = \frac{3+1}{11+7} = \frac{4}{18}$
$loglikelihood[fly, comedy] = \frac{1+1}{9+7} = \frac{2}{16}$

$loglikelihood[fly, action] = \frac{1+1}{11+7} = \frac{2}{18}$

We have the following table:

word	comedy	action
fun	4/16	3/18
couple	3/16	1/18
love	3/16	2/18
fast	2/16	3/18
furious	1/16	3/18
shoot	1/16	4/18
fly	2/16	2/18

Next we can compute classifier output:

$$sum[comedy] = -0.3979 + \log\frac{2}{16} + \log\frac{3}{16} + \log\frac{1}{16} + \log\frac{2}{16} = -4.1351$$
$$sum[action] = -0.2218 + \log\frac{3}{18} + \log\frac{1}{18} + \log\frac{4}{18} + \log\frac{2}{18} = -3.8627$$

Because $sum[action] > sum[comedy]$, we assert that D should be in the action class

3. Why might a Naive Bayes classifier be used as a baseline classifier in sentiment analysis?

Ans: The following are all applicable: Computationally efficient in aspects such as model size and training time, good performance, simplifying assumptions, effective on small datasets, interpretable, simple handling of class imbalance.

4. Why is the 'Naive' assumption made in Naive Bayes, and how does it affect sentiment classification?

Ans: The 'Naive' assumption in Naive Bayes refers to the assumption that all features (words, in the case of text data) are independent of each other. This simplifies the computation, making the model computationally efficient. In sentiment classification, this means that the algorithm treats each word as contributing independently to the sentiment, ignoring any possible correlation between words. While this is a strong and often unrealistic assumption, Naive Bayes can still perform surprisingly well in many sentiment classification tasks.