CSC 4309 : NATURAL LANGUAGE PROCESSING (SEM 2, 18/10)
SAMPLE PRACTICE QUESTION (TEST 1)

1. Given the following training data and testing data:

[15 marks]

Training data	Testing data
They visit their parents in Dubai.	My brother does not travel to Dubai.
She does not like to visit her parents.	
My sisters travel to Europe.	

a. Using `Maximum Likelihood Estimation (MLE)` and `Laplace` smoothing, calculate the `likelihood` of the test data using a `trigram` model. (8 marks)

b. Using the same model in (1), calculate the `log likelihood` of the test data. (4 marks)

c. Compute the `per word entropy` and `perplexity` for the test data. (4 marks)

2. Create a python **regular expression** that is able to find all words that contain **'ter'** but not at the beginning or the end. (e.g.: 'in**ter**mediate', 'i**ter**ate', 'en**ter**tain', 'lan**ter**n'). You are allowed to test your solution with a Python interpreter and editor (2 marks)

Script:

3. According to Shannon's information theory, ***entropy*** is the "average number of bits needed to encode information". Choose the possible binary bit representation for the following phrase.

as high as a kite

a. 0001110
b. 1000110
c. 10001010
d. 10010110