Experiment 1 Entropy of English Text

1. Purpose:

1.1 Understanding the entropy of discrete memoryless and discrete memorized sources.
1.2 Understanding the joint entropy and the conditional entropy, and their differences.
1.3 Understanding the properties of entropy.

2. Principle:

The following formulas may be useful during the experiment.
2.1 Entropy

$$H(X) = -\sum_{i=1}^{q} P(a_i) \log P(a_i)$$

2.2 Average joint entropy

$$\frac{1}{2} H(X,Y) = -\frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{s} P(a_i, b_j) \log P(a_i, b_j)$$

$$\frac{1}{n} H(X_1, \cdots X_n) = -\frac{1}{n} \sum_{\vec{x}} P(\vec{x}) \log P(\vec{x})$$

2.3 Conditional entropy

$$H(X \mid Y) = -\sum_{i=1}^{q} \sum_{j=1}^{s} P(a_i, b_j) \log P(a_i \mid b_j)$$

3. Procedure:

We consider an English text with *26 alphabets and space*. Note that we treat the upper and lower cases of alphabets as the same, and does not consider the punctuations（标点符号）in the text. From the information theory, if the 27 symbols are i.i.d., then the entropy of the English text is given by

$$H_0 = \log(27) = 4.7549 \text{ bit/symbol}.$$

We now consider different entropy of the English text source by the following procedures.

3.1 Preprocess of the English text data

We first need to transform the data provided in the file “English text data” into an $1 \times N$ vector, where N is the total number of symbols. Note that in obtaining the vector for all the symbols, we do not consider the punctuations, and treat the upper and lower cases of the letters as the same.

Hints: One can use the function fscanf() to scan the file, function lower() to change all the upper case letters into lower ones, and function isletter() to obtain the letters from the source data.

3.2 The Entropy $H_1=H(X_1)$

To evaluate $H_1=H(X_1)$, we consider the distribution of the 27 symbols without considering the correlation among consecutive symbols. This can be done by evaluating the probability space of each symbol X_1 based on the English text data, and equation 2.1.

Hints: One can create a $\mathbf{27\times 1}$ vector to evaluate the probability space by looping over the English text data one-by-one.

3.3 The Entropy of $H_2=\frac{1}{2}H(X_1,X_2)$

To evaluate the Entropy $H_2=\frac{1}{2}H(X_1,X_2)$, we now consider both the distribution of the 27 symbols, and the correlation among *two* consecutive symbols. This can be done by evaluating the probability space of two consecutive symbols based on the English text data, and equation 2.2.

Hints: One can create a $\mathbf{27\times 27}$ matrix to evaluate the joint probability space by looping over the English text data *two at a time*, e.g., for text data a1a2a3a4a5a6…, looping as a1a2, a2a3, a3a4, a4a5, a5a6….

3.4 The Entropy of $H(X_2|X_1)$

The Entropy $H(X_2|X_1)$ can be evaluated based on the conditional probability space of two consecutive symbols and equation 2.3.

Hints: One can create a $\mathbf{27\times 27}$ matrix to evaluate the conditional probability space by looping over the English text data *two at a time*. Note that the row represents the next symbol, while the column represents the previous symbol. Remember to check whether $\sum_{i=1}^{27}P(x_i|x_j)=1\forall j$ when finishing the evaluation of conditional probability space.

3.5 The Entropy of $H_3=\frac{1}{3}H(X_1,X_2,X_3)$

Evaluate the Entropy $H_3=\frac{1}{3}H(X_1,X_2,X_3)$.

Hints: One can create a $\mathbf{27\times 27\times 27}$ matrix to evaluate the joint probability space.

3.6 The Entropy of $H_4=\frac{1}{4}H(X_1,X_2,X_3,X_4)$

4. Report Requirements:

4.1 Draw a simple flowchart for the codes of 3.4.

4.2 Show the results of the different entropy in 3.2-3.6.

4.3 Discuss about the differences and inequalities in 3.2-3.6, use the properties of entropy to explain the reason.