Information Theory Assignment

s1140740

1 Source Coding

1.1 Character Statistics

The entropy $H(X_n)$ is 4.168.

1.2 Bigram Statistics

a. The joint entropy $H(X_n, X_{n+1})$ is 7.572.
b. Because they are not i.i.d.
c. The conditional entropy $H(X_{n+1} \mid X_n) = H(X_n, X_{n+1}) - H(X_n)$ is 3.404.

1.3 Compression with known distributions

The maximum number of bits an arithmetic coder would use to represent `thesis.txt` assuming i.i.d. model is 1,433,836. The length is computed as:

$$length = \left\lceil -\sum_{n=1}^{N} \log_2 P(x_n) \right\rceil + 2$$

where x_n is the n-th character in the file and N is the length of the file.

The maximum number of bits an arithmetic coder would use to represent `thesis.txt` using the joint probability from question 2 is 1,171,194. The length is computed as:

$$length = \left\lceil -\log_2 P(x_1) - \sum_{n=1}^{N} \log_2 P(x_{n+1}|x_n) \right\rceil + 2$$