Entropy and Information

1 Introduction

The entropy of a random variable X is

$$H(X) = -\sum_{x \in \mathbb{X}} P(X = x) \log P(X = x). \tag{1}$$

The conditional entropy of variable X given variable Y is

$$H(X|Y) = H(X, Y) - H(Y). \tag{2}$$

The mutual information between variables X and Y is

$$I(X; Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y). \tag{3}$$

2 Task

1. Download some texts, DNA sequences, or other discrete symbolic sequences (e.g., music in an appropriate format) in a sufficient amount (say, about 1MB) from the internet.

2. Find the entropy of the empirical distribution of characters/bases.

3. For the empirical distribution of characters, find the mutual information between two characters/bases separated by k characters/bases. Plot the mutual information in a doubly logarithmic scale and try to fit some function to the data points.

4. Describe what you have obtained in a report, attach the used scripts, and send it to me (`ldebowsk@ipipan.waw.pl`).