

Мини-проект 2: Энтропия

Количественная оценка объема информации, содержащейся в данных - один из основных вопросов теории информации. В этом мини-проекте вы будете использовать теорию информации, чтобы определить, действительно ли изображение стоит больше тысячи слов.

Для этого мы сначала вычислим энтропию (меру информации) всей книги (содержащей более 1000 слов) и сравним ее с энтропией изображения.

- (a) Загрузите текст книги по вашему выбору (книга у каждого своя). Например «Граф Монте-Кристо» Александра Дюма. Загрузите файл и подсчитайте, сколько раз появляется каждый символ x_i в тексте. *Подсказка:* это можно сделать в трех строках кода, используя функции `fileread` и `hist` Matlab.
- (b) Оцените вероятность $p(x_i)$ каждого символа.
- (c) Вычислите энтропию символа как:

$$H(x) = \sum_{x_i} p(x_i) \log_2 \left(\frac{1}{p(x_i)} \right)$$

- (d) Ваш результат из (c) сообщает вам информацию, закодированную в каждом символе книги. Теперь умножьте это количество символов в книге, чтобы получить общую энтропию книги.

Теперь вычислим энтропию изображения.

- (e) Загрузите изображение по вашему выбору (лучше, для наглядности, иллюстрацию из книги). Загрузите изображение и вычислите, сколько раз появляется интенсивность каждого пикселя x_i . *Подсказка:* вы также можете сделать это в трех строках кода.
- (f) Оцените вероятность $p(x_i)$ интенсивности каждого пикселя.
- (g) Вычислите энтропию пикселя по той же формуле.
- (h) Ваш результат из (g) сообщает вам информацию, закодированную в каждом пикселе изображения. Теперь умножьте это на количество пикселей в изображении, чтобы получить общую энтропию изображения.
- (i) Что содержит больше информации, книга или изображение?
- (j) Как вы думаете, это сравнение справедливое? Почему да или почему нет?