

Мини-проект 2: Количество информации (крайний срок - 15 октября 18.00)

Преподаватель: Р.Х.Латыпов, 26.09.2022

Фундаментальный аспект современных наук, связанных с данными - количественная оценка объема содержащейся в них информации. Это один из основных вопросов теории информации.

В данном мини-проекте используется шенноновское определение количества информации. Вам предлагается провести вычисления и решить, действительно ли изображение стоит больше тысячи слов.

Для этого необходимо вычислить энтропию (меру информации) всей книги (содержащей более 1000 слов) и сравнить ее с энтропией изображения из этой книги.

(a) Загрузите текст книги по вашему выбору.

(b) Вычислите, сколько раз появляется каждый символ x_i .
Оцените вероятность $p(x_i)$ каждого символа

(c) Вычислите энтропию символа по формуле

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

(d) Ваш результат из (c) сообщает вам ожидаемую информацию, закодированную в каждом символе текста книги. Теперь умножьте полученную энтропию на количество знаков в книге, чтобы получить общую энтропию книги.

После этого вычислить энтропию изображения.

(e) Загрузите изображение из книги по вашему выбору. Если рисунок цветной, перевести в черно-белое изображение. Вычислите, сколько раз появляется каждая интенсивность пикселя x_i (диапазон от 0 до 255).

(f) Оцените вероятность $p(x_i)$ каждой интенсивности пикселя.

(g) Вычислите энтропию пикселя.

(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 рисунка.

(i) Что содержит больше информации, книга или изображение?

(j) Как вы думаете, это сравнение справедливо? Почему да или почему нет?