Романенко Михаил, ФТ-403

0 задание:

Для анализа был выбран текст романа американского писателя Стивена Кинга - «11/22/63». Оригинал текста: <https://archive.org/stream/11-22-63/11-22-63_djvu.txt>

Код: https://github.com/BootyOne/text-decapitalization

1 задание:

В качестве правил капиталиации были взяты следующие:

- первый символ параграфа

- все слова “I”

- первая буква после некоторых знаков (’.’, ’ !’, ’ ?’, ‘”’)

Применил правила капитализации, и все заглавные буквы, которые не подходят под правила, отметил единицами в битовом массиве. Энтропия составила 0.83. То есть распределение символов этого текста очень скошено.  
Поэтому алгоритм Хаффмана (вероятность != степени двойки) и арифметический метод (некоторые отрезки будут слишком малы) не очень подходят.

В итоге было выбрано дельта-кодирование. Среди рассмотренных нами методов сжатия целых чисел оно достаточно эффективно по памяти на отрезке до 256. Используя этот метод, получилось сжать 1476960 бит до 318427 бит, то есть примерно в 5 раз.

2 задание:

Для подстрок построил контекстные модели

При подсчете размера моделей контекста каждая модель включает в себя размер строки, а также размер записей для каждого символа, включая код символа (1 байт) и количество его встреч с левым текстом (4 байта). Эти размеры суммируются для каждой модели, чтобы получить общий размер моделей контекста.

Для заданного текста результат получился равен 441455 байтам.