**Лабораторная работа №1**

**Реализация алгоритма для сжатия данных**

**Исполнитель:** Клюжев Игорь Адрианович, студент, группа M4106

**Проверяющий:** Беляев Евгений Александрович

**Исходный код:** [github.com:Igor-bite/compression-algo-2024](https://github.com/Igor-bite/compression-algo-2024)

**Описание алгоритма**

В рамках данной лабораторной работы был взят алгоритм PPMA с его обычной реализацией. Длина контекста может задаваться внутри кода, но она должна быть задана одинаковой для кодера и декодера. Запись в файл длины контекста кодера не производится.

В качестве escape символа использую число 257. Контексты с их вероятностями храню рекурсивно: то есть контекст буквы может хранить в себе для каждой буквы ещё контекст и так далее. Длина цепочки как раз будет равняться длине сохраняемого контекста. В данный момент в коде выставлена длина контекста равная 3, так как это наилучшая длина контекста для выданного датасета (было проверено перебором).

Для запуска кодера необходимо перейти в директорию с кодом и запустить скрипт на питоне командой:

python compress.py input\_file output\_file

заменив параметры имён входного и выходного файлов на необходимые.

Таким же образом можно запустить и декодер:

python decompress.py input\_file output\_file

При передаче программе не 2 параметра – будет ошибка и программа остановится.

Также при сжатии файла после окончания сжатия будет выведена информация размера входного файла, посчитанных энтропий для входного файла, а также затрат бит на символ.

**Посчитанные параметры для датасета**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **File** | **H(X)** | **H(X|X)** | **H(X|XX)** | **Size** | **Compressed size** | **Compress rate** | **Bits by symbol** |
| bib | 5,20 | 3,36 | 2,31 | 111261 | 32768 | 71% | 2,36 |
| book1 | 4,53 | 3,58 | 2,81 | 768771 | 249856 | 67% | 2,60 |
| book2 | 4,79 | 3,75 | 2,74 | 610856 | 188416 | 69% | 2,47 |
| geo | 5,65 | 4,26 | 3,46 | 102400 | 90112 | 12% | 7,04 |
| news | 5,19 | 4,09 | 2,92 | 377109 | 143360 | 62% | 3,04 |
| obj1 | 5,95 | 3,46 | 1,40 | 21504 | 12288 | 43% | 4,57 |
| obj2 | 6,26 | 3,87 | 2,27 | 246814 | 102400 | 59% | 3,32 |
| paper1 | 4,98 | 3,65 | 2,33 | 53161 | 16384 | 69% | 2,47 |
| paper2 | 4,60 | 3,52 | 2,51 | 82199 | 24576 | 70% | 2,39 |
| paper3 | 4,67 | 3,55 | 2,56 | 46526 | 16384 | 65% | 2,82 |
| paper4 | 4,70 | 3,48 | 2,21 | 13286 | 4096 | 69% | 2,47 |
| paper5 | 4,94 | 3,53 | 2,04 | 11954 | 4096 | 66% | 2,74 |
| paper6 | 5,01 | 3,61 | 2,25 | 38105 | 12288 | 68% | 2,58 |
| pic | 1,21 | 0,82 | 0,71 | 513216 | 61440 | 88% | 0,96 |
| progc | 5,20 | 3,60 | 2,13 | 39611 | 12288 | 69% | 2,48 |
| progl | 4,77 | 3,21 | 2,04 | 71646 | 16384 | 77% | 1,83 |
| progp | 4,87 | 3,19 | 1,76 | 49379 | 12288 | 75% | 1,99 |
| trans | 5,53 | 3,35 | 1,93 | 93695 | 20480 | 78% | 1,75 |
|  |  |  |  |  |  |  |  |
|  |  |  |  | **Overall size** | **1019904** |  |  |