Лабораторна робота №1

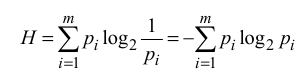
Дослідження кількості інформації при різних варіантах кодування

**Мета:** Дослідити імовірнісні параметри української мови для оцінки кількості інформації текстів. Дослідити вплив різних методів кодування інформації на її кількість.

# Теоретичні відомості

***Відносна частота появи символу*** - імовірність появи певного символу в певному місці тексту - відношення числа появи символу в тексті до загальної кількості символів.

***Середня ентропія нерівноймовірного алфавіту***:

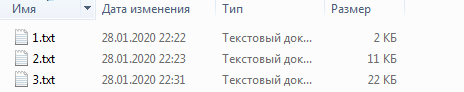


де *m* - кількість символів алфавіту, *p* - імовірність появи символу

Ентропія вимірюється в **БІТАХ** (як представлення кількості можливих варіантів).

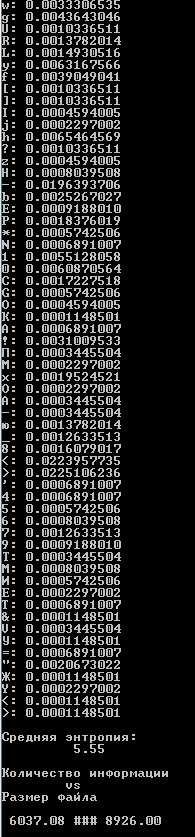
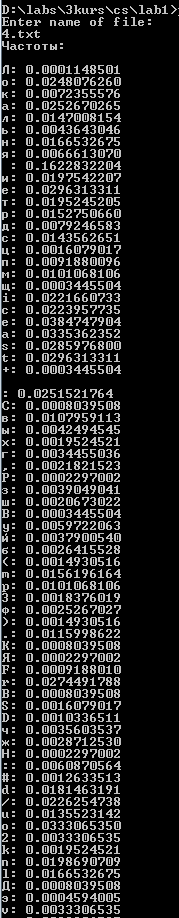
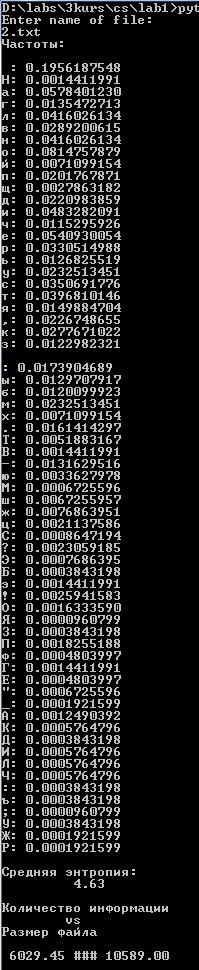
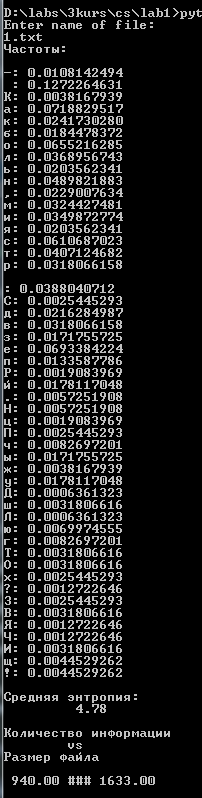
***Кількість інформації в тексті*** - середня ентропія вихідного алфавіту помножена на кількість символів тексту. (***HINT***: результат обрахунку для порівняння значення з розміром файлів треба перевести з бітів в байти)

# 1. Дослідження кількості інформації в тексті

1. 

2. Наврядчи хтось ще візьме «Балладу о прокуренном вагоне», «Улыбку» Р. Бредбері та статтю «Локальная инет радиостанция при помощи icecast +ices» з Яндекс.Дзену :)

3.



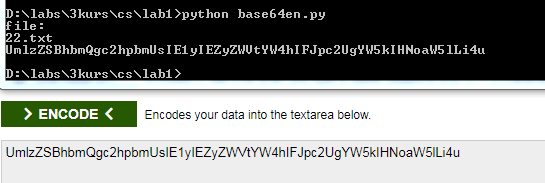
4. Проведіть стисн. кожного вхід. файлу за допомогою 5 різних алгоритмів стиснення

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ориг. файл | Размер | кол-во информации. | rar | zip | gz | bz2 | xz |  |  |  |  |  |
| 1.txt | 1633 | 940.00 | 822 | 862 | 730 | 784 | 836 |  |  |  |  |  |
| 2.txt | 10589 | 6029.45 | 5535 | 5440 | 5224 | 4889 | 5356 |  |  |  |  |  |
| 3.txt | 8926 | 6037.08 | 3892 | 3942 | 3774 | 3619 | 3684 |  |  |  |  |  |

5. У всіх трьох випадках кількість інформації була меншою розміру файлу, бо сам файл окрім даних містить інші службові дані та атрибути. Також вона завжди була більшою за розміри архівів. Це і логічно, бо використовуються досконалі алгоритми стискання інформації. Також було помічено, що файл, де ентропія більша, вдається стиснути на більший відсоток відносно кількості інформації, відсоток ж відносно розміру файлу на диво сталий і складає близько 40-50%.

# 2. Дослідження способів кодування інформації на прикладі Base64

1. Ознайомтесь зі стандартом [RFC4648](https://tools.ietf.org/html/rfc4648)
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)
   1. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)



1. Закодуйте в Base64 обрані вами текстові файли
   1. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
   2. Порівняйте отримане значення з кількістю інформації вихідного файлу
   3. Зробіть висновки з отриманого результату
2. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
   1. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
   2. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу[[1]](#footnote-1)
   3. Зробіть висновки з отриманого результату

1. Для кращого сприйняття інформації **обов’язково** подайте отримані значення у вигляді таблиці, що містить всі варіанти значення обрахованої кількості інформації та **відповідні діаграми** на основі табличних даних [↑](#footnote-ref-1)