Київський національний університет імені Тараса Шевченка

ФРЕКС

ЗВІТ

до лабораторної роботи №1

«Дослідження кількості інформації при різних варіантах кодування»

Виконав:

Студент 3-го курсу КІ-СА

Момот Дмитро Володимирович

Київ, 2020

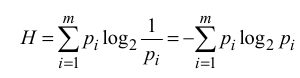
**Мета роботи:**

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)

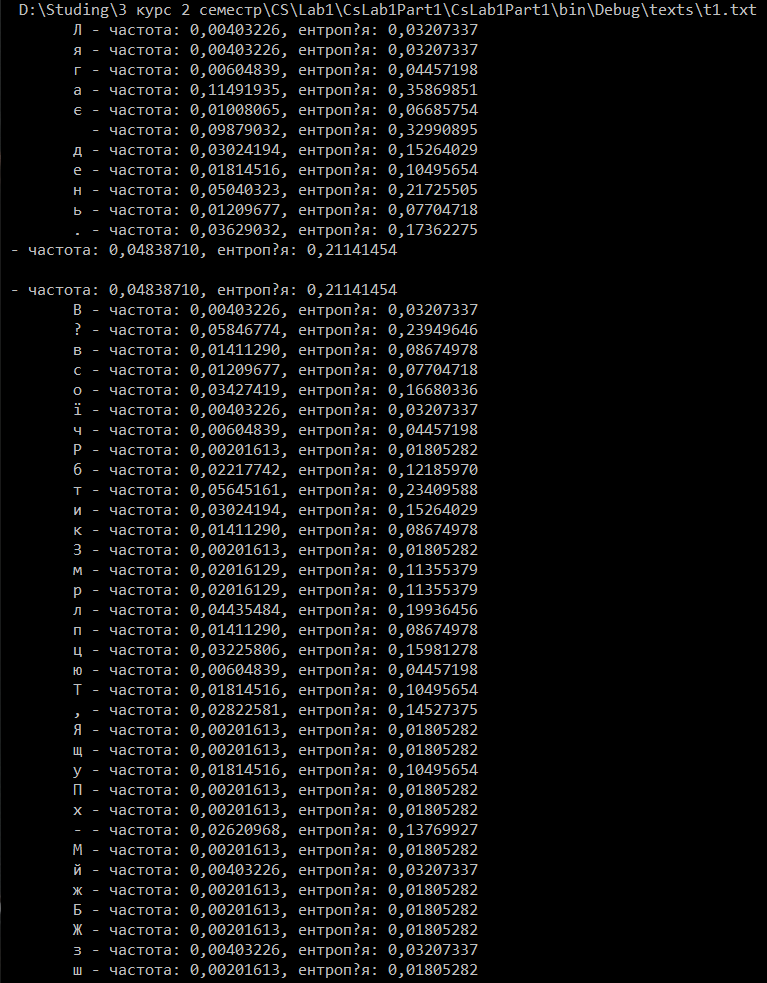
**Хід виконання роботи:**

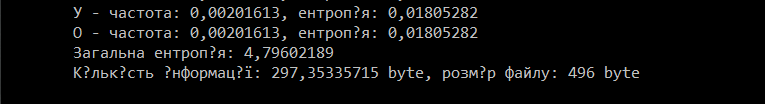
**Дослідження кількості інформації в тексті**

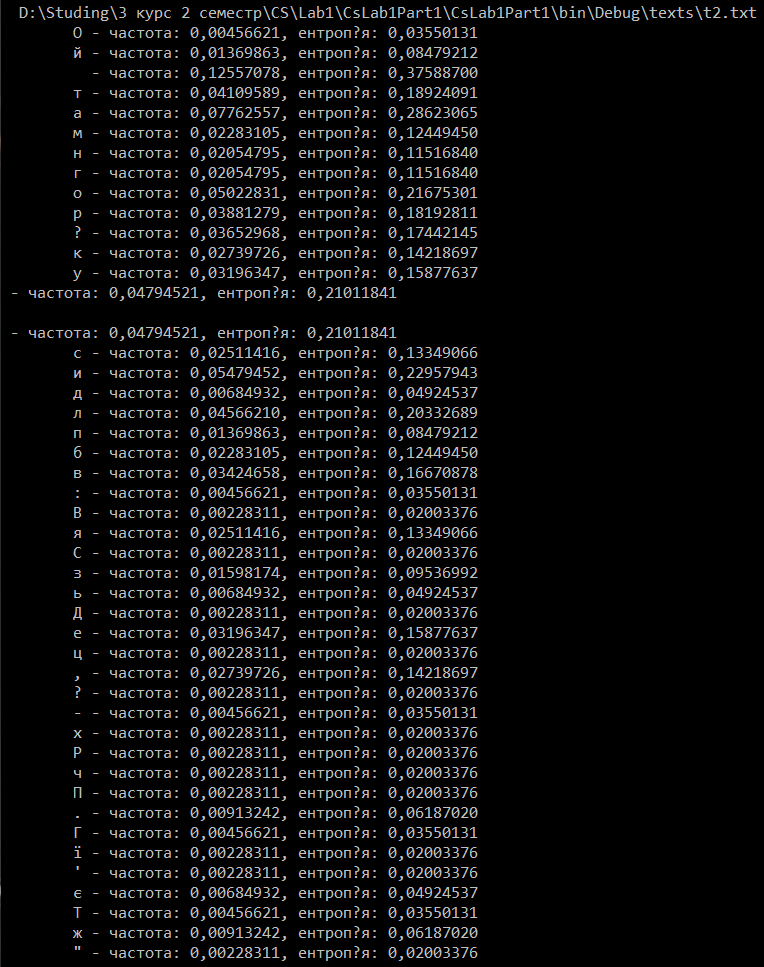
1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI)

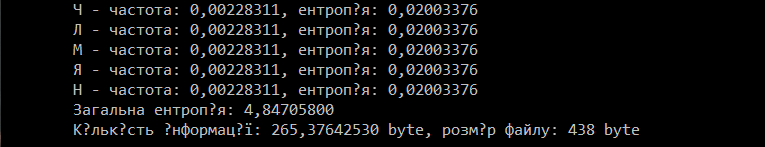
* Текст пісні Воплi Вiдоплясова – Танці (t1.txt)
* Текст пісні Українська народна пісня - Ой там на горі (t2.txt)
* Українські народні казки - Шовкова держава (t3.txt)

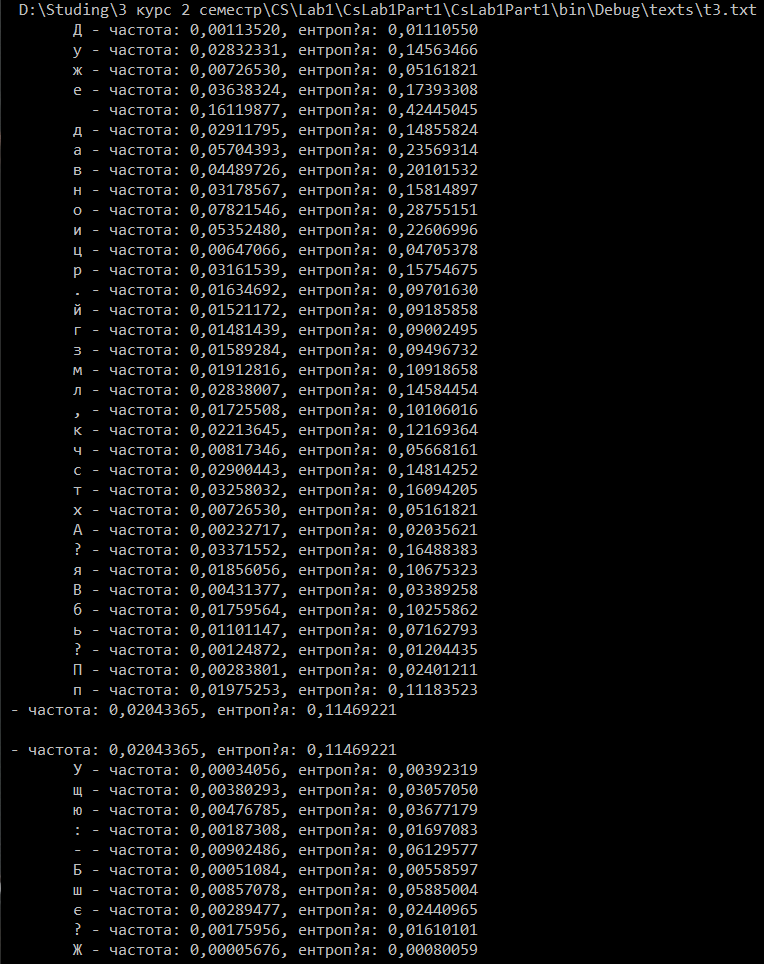
1. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
   1. обраховує частоти (імовірності) появи символів в тексті
   2. обраховує середню ентропію алфавіту для даного тексту
   3. виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
   4. виводить на екран значення частот, ентропії та кількості інформації

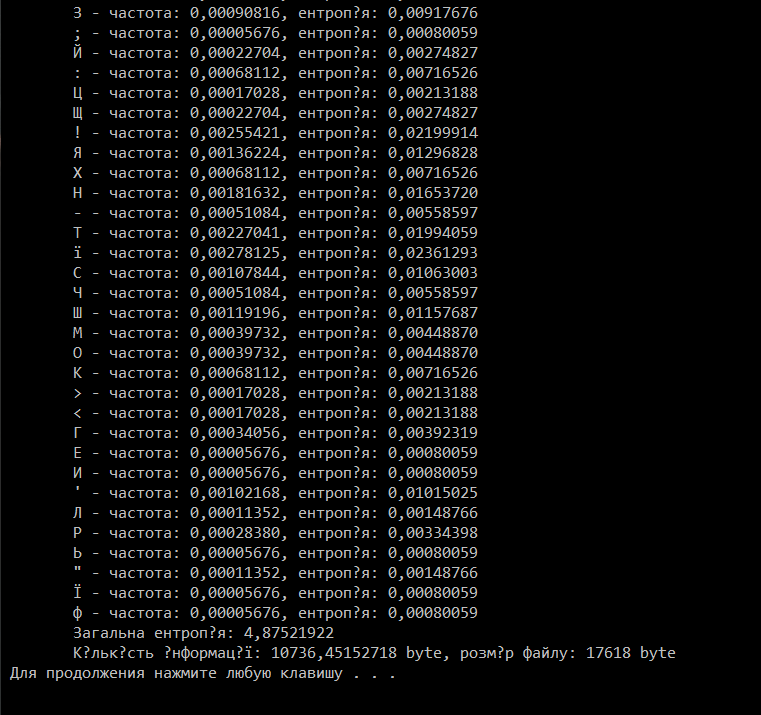


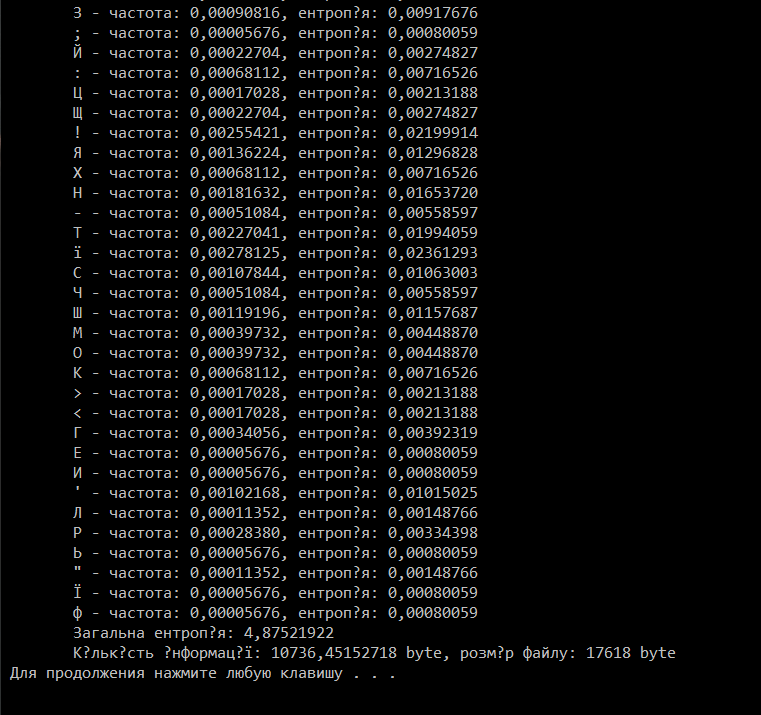




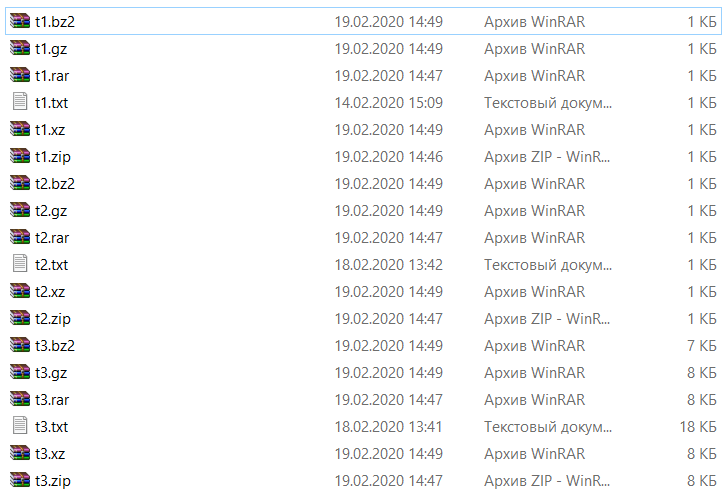








1. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення).

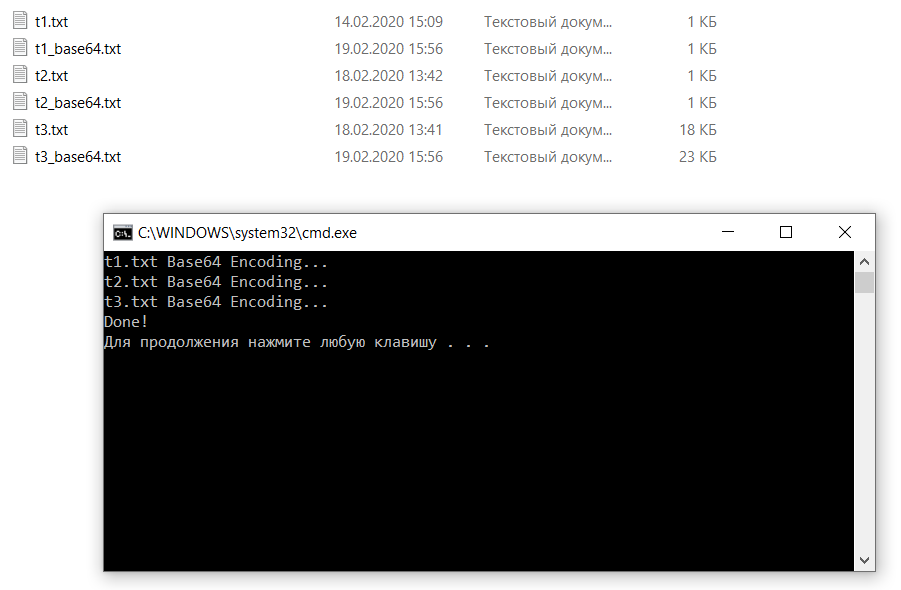


1. Порівняйте результуючі обсяги архівів з обчисленою кількістю інформації та **наведіть у звіті висновки** щодо кореляції цих величин для обраних вами файлів (яка відмінність, що вийшло більше і чому)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | bzip2 | gzip | rar | xz | zip | кількість інформації |
| t1.txt | 324 байт | 309 байт | 376 байт | 372 байт | 394 байт | 297,35 байт |
| t2.txt | 324 байт | 307 байт | 376 байт | 376 байт | 392 байт | 265,38 байт |
| t3.txt | 6 929 байт | 7 820 байт | 7 877 байт | 7 664 байт | 7 923 байт | 10736,45 байт |

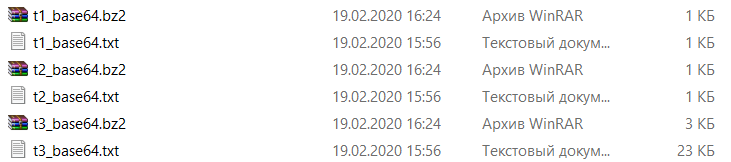
**Дослідження способів кодування інформації на прикладі Base64**

1. Ознайомтесь зі стандартом RFC4648
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)
3. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)
4. Закодуйте в Base64 обрані вами текстові файли
5. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
6. Порівняйте отримане значення з кількістю інформації вихідного файлу
7. Зробіть висновки з отриманого результату



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | t1.txt | t2.txt | t3.txt | t1\_base64.txt | t2\_base64.txt | t3\_base64.txt |
| Кількість інформації | 297,35 байт | 265,38 байт | 10736,45 байт | 288,6 байт | 246,6 байт | 9398,2 байт |

1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
2. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
3. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу
4. Зробіть висновки з отриманого результату



|  |  |
| --- | --- |
| Назва файлу | Кількість інформації |
| t1.txt | 297,35 байт |
| t1\_base64.txt | 288,6 байт |
| t1\_base64.bz2 | 150,45 байт |
| t2.txt | 265,38 байт |
| t2\_base64.txt | 246,6 байт |
| t2\_base64.bz2 | 142,14 байт |
| t3.txt | 10736,45 байт |
| t3\_base64.txt | 9398,2 байт |
| t3\_base64.bz2 | 2247 байт |

**Висновок:**

На даній лабораторній роботі було досліджено імовірнісні параметри української мови для оцінки кількості інформації текстів та вплив різних методів кодування інформації на її кількість. Також було досліджено, що кількість інформації менша ніж розмір стиснутого файлу та те, що кращим алгоритмом стиснення з використаних виявився BZip2.

GitHub: <https://github.com/dima64668/CS>