**Київський національний університет імені Тараса Шевченка**

**факультет радіофізики, електроніки та комп’ютерних систем**

Лабораторна робота № 1

**Тема:** «Дослідження кількості інформації при різних варіантах кодування»

Роботу виконав

студент 3 курсу

КІ - СА

Бондаренко Владислав

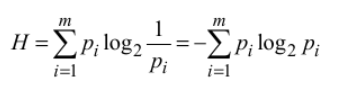
Київ 2020

**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).

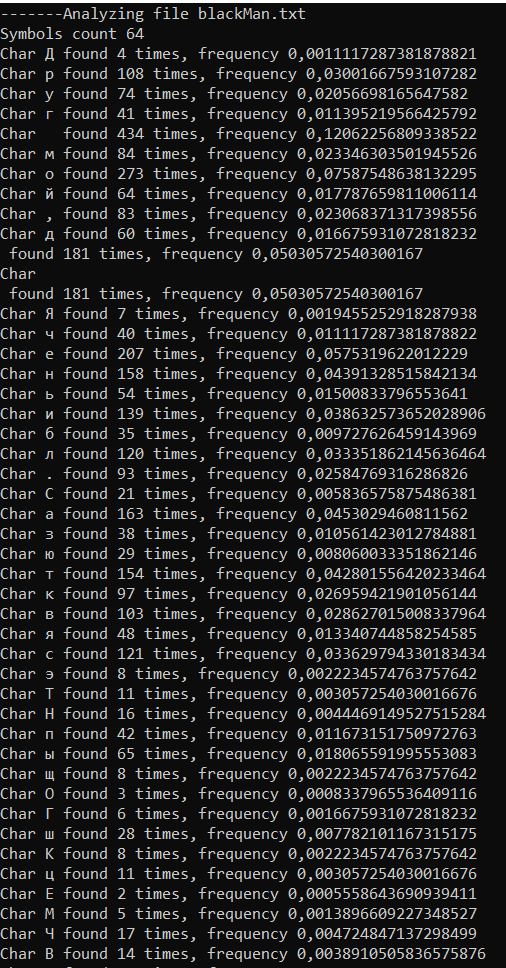
**Кількість інформації в тексті** - середня ентропія вихідного алфавіту помножена на кількість символів тексту. (**HINT**: результат обрахунку для порівняння значення з розміром файлів треба перевести з бітів в байти)

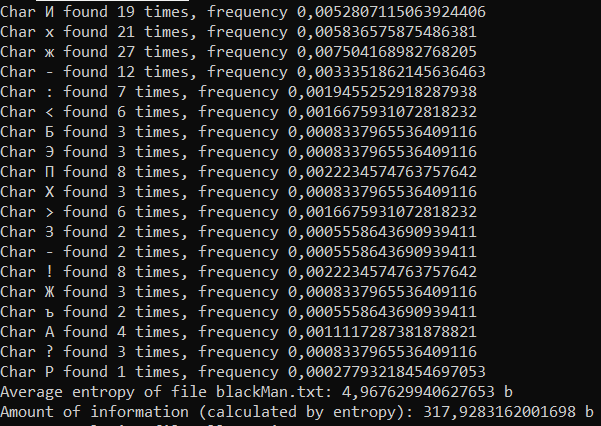
**Хід роботи**

1. Дослідження кількості інформації в тексті

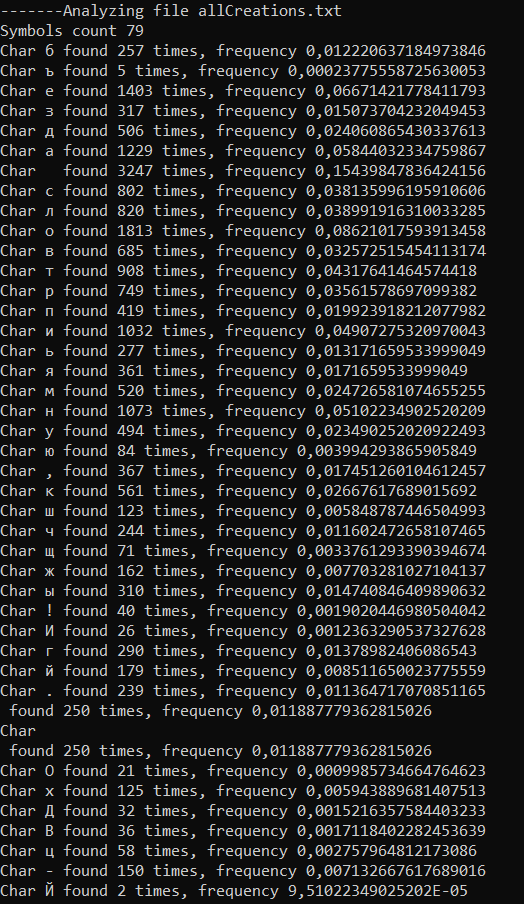
1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI)
   1. **blackMan.txt** – Вірш. Автор: Сергій Есенін. Название: “Черный человек”.
   2. **allCreations.txt** – Уривок з книги. Автор: Д. Херриот. Назва: “О всех созданиях”.
   3. **trash.txt –** Стаття. Автор: Жемжуров Михаил Леонидович. Назва: “Радиотоксичность облученного ядерного топлива реактора ВВЭР-1200 в зависимости от выгорания и времени выдержки”.
2. Переконайтесь, що тексти, які ви використовуєте є унікальними і не повторюються у ваших колег! Використовуйте наявні електронні засоби зв’язку та документообігу, щоб уникнути дублювання! Вдруге аналіз того самого тексту не зараховується!
3. 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

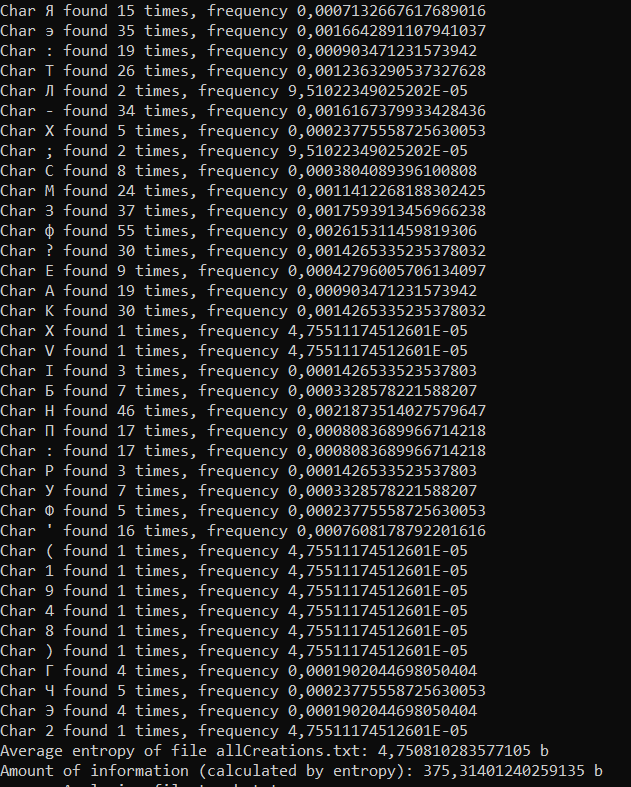
**blackMan.txt**



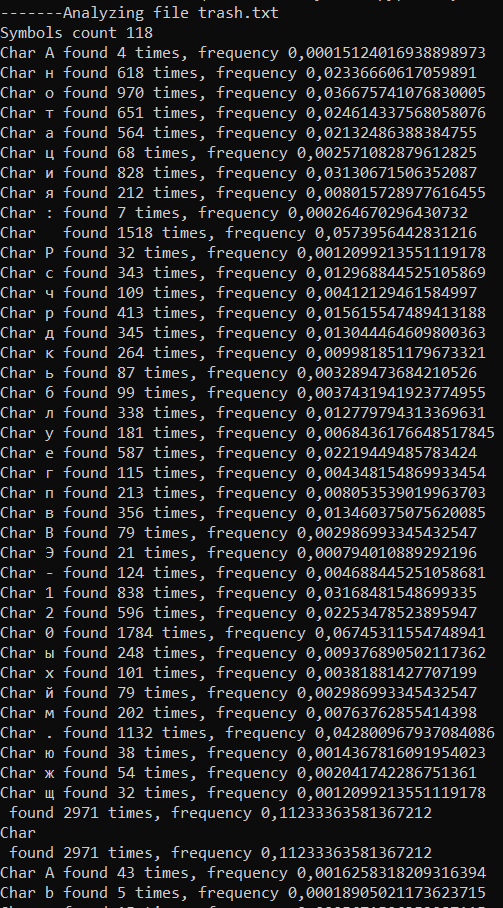


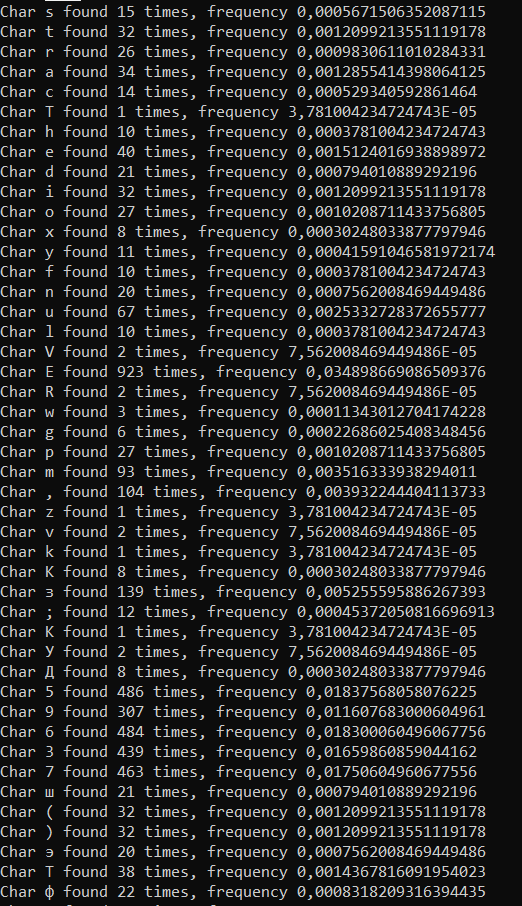
**allCreations.txt**

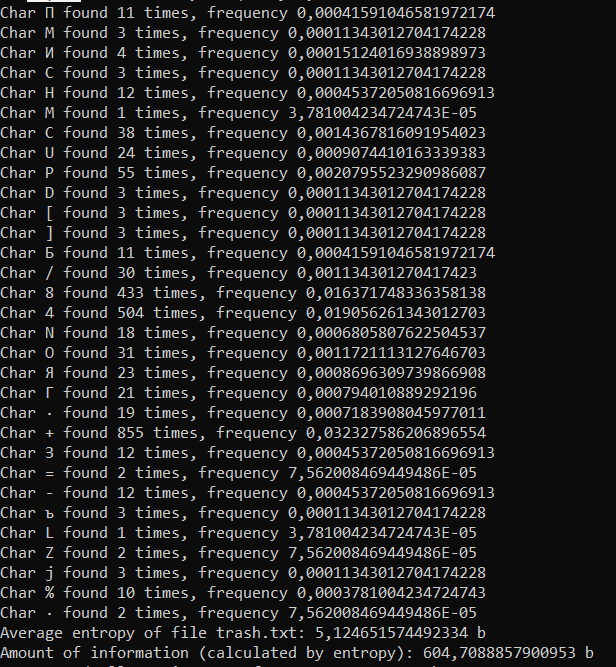
****

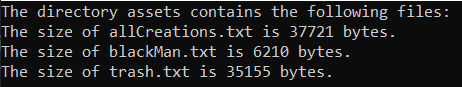


**trash.txt**

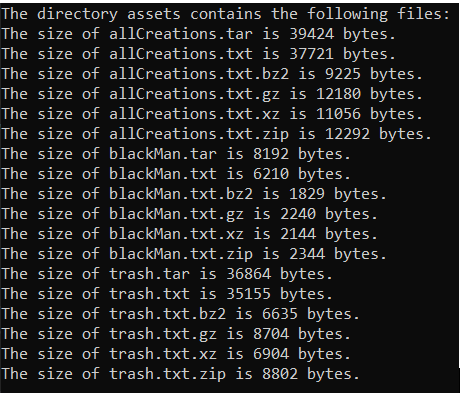






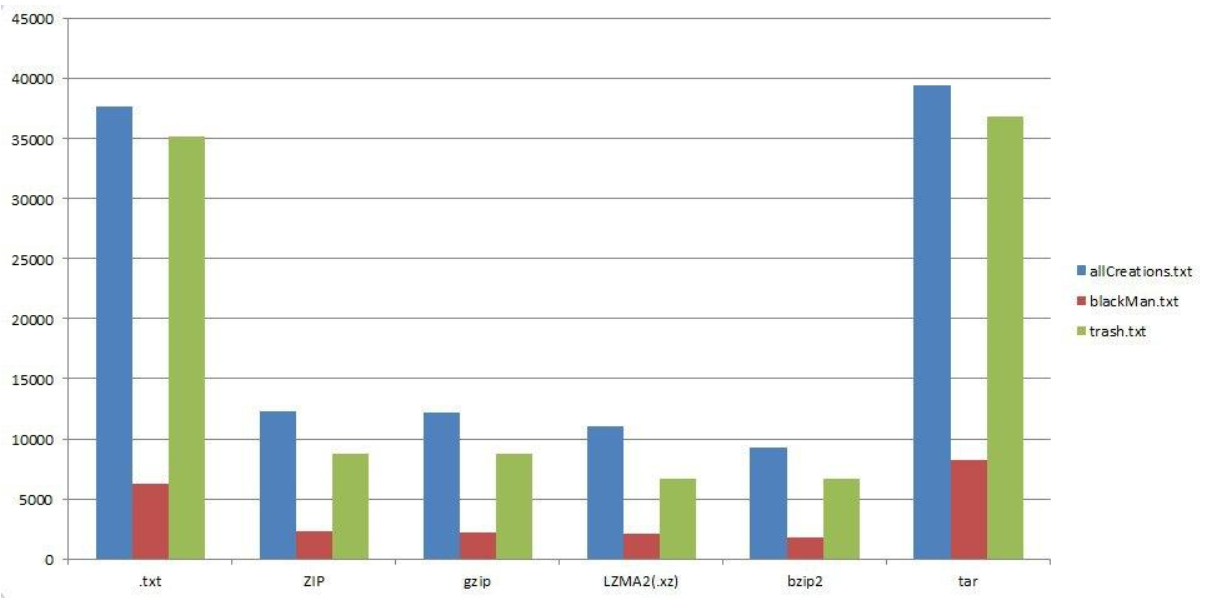


1. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення).
2. Порівняйте результуючі обсяги архівів з обчисленою кількістю інформації та **наведіть у звіті висновки** щодо кореляції цих величин для обраних вами файлів (яка відмінність, що вийшло більше і чому)



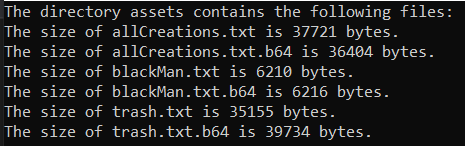
Розміри файлів

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Файл | .txt | ZIP | gzip | LZMA2(.xz) | bzip2 | tar | К-кість інформації (за доп. ентропії) |
| allCreations.txt | 37721 | 12292 | 12180 | 11056 | 9225 | 39424 | 375.3 |
| blackMan.txt | 6210 | 2344 | 2240 | 2144 | 1829 | 8192 | 317.9 |
| trash.txt | 35155 | 8802 | 8704 | 6635 | 6635 | 36864 | 604.7 |



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)
3. 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
4. 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
   3. Зробіть висновки з отриманого результату



blackMan.txt.b64



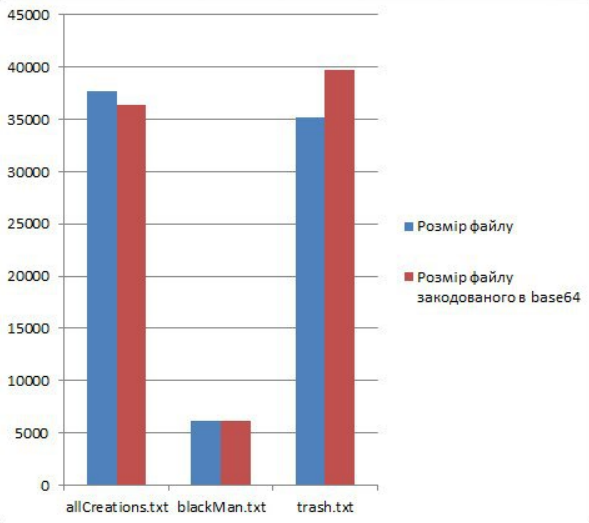
allCreations.txt.b64



trash.txt.b64



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Файл | Розмір файлу | Розмір файлу закодованого в base64 | Середня ентропія файлу | Середня ентропія файлу закодованого в base64 | К-кість інформації файлу | К-кість інформації файлу закодованого в base64 |
| allCreations.txt | 37721 | 36404 | 4.75 | 5.43 | 375.3 | 364 |
| blackMan.txt | 6210 | 6216 | 4.97 | 5.5 | 318 | 369 |
| trash.txt | 35155 | 39734 | 5.12 | 5.37 | 604.7 | 360 |



**Висновок:** В даній лабораторній роботі було навчено обрахуванню ентропії, написанню алгоритму для обрахування ентропії, написання алгоритму для кодування бінарного коду в текстовий формат base64, який широко використовується у Всесвітній паутині, наприклад, для вставки зображень і інших двійкових ресурсів в HTML і CSS.

Github: <https://github.com/Lrazerz/CompSystemsLabs>