**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

Мургашов Гліб

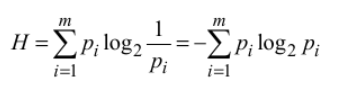
**Київ 2020**

**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).

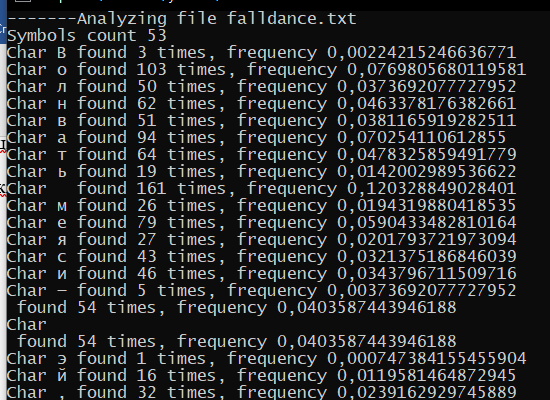
**Кількість інформації в тексті** - середня ентропія вихідного алфавіту помножена на кількість символів тексту. (**HINT**: результат обрахунку для порівняння значення з розміром файлів треба перевести з бітів в байти)

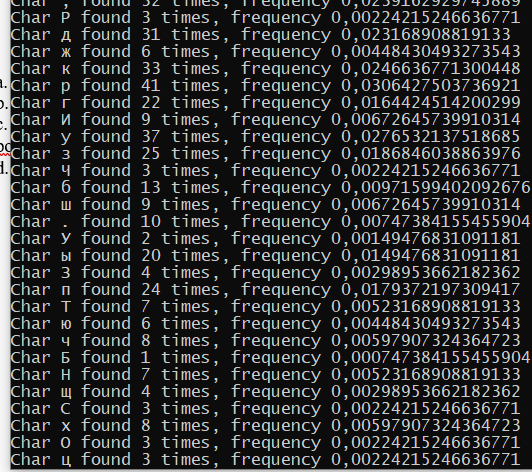
**Хід роботи**

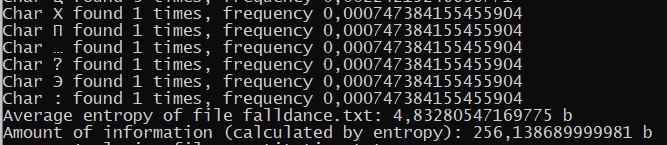
1. Дослідження кількості інформації в тексті

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. **falldance.txt** – Вірш. Автор: Олександр Блок. «Пляски осенние»
   2. **briefHistoryOfTime.txt** – Уривок з книги. Автор: C. Хокинг. Назва: “Краткая история времени”.
   3. **constitution.txt –** уривок із Конституції України
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

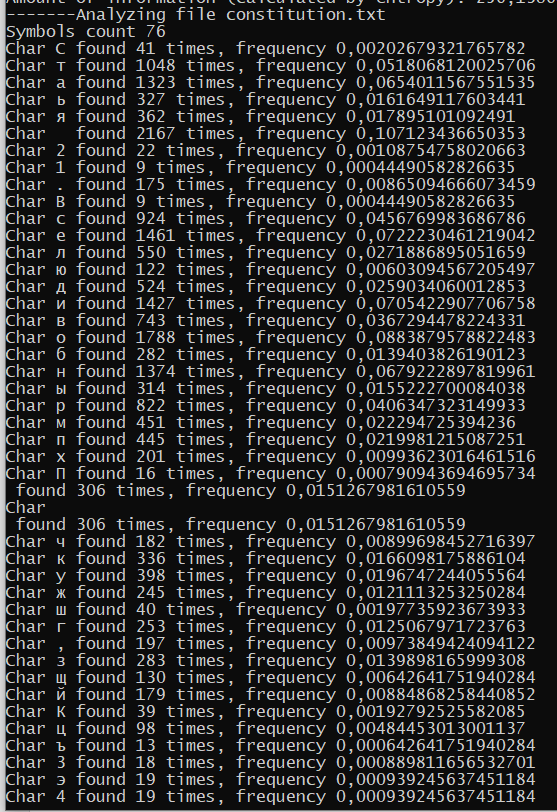
**falldance.txt**

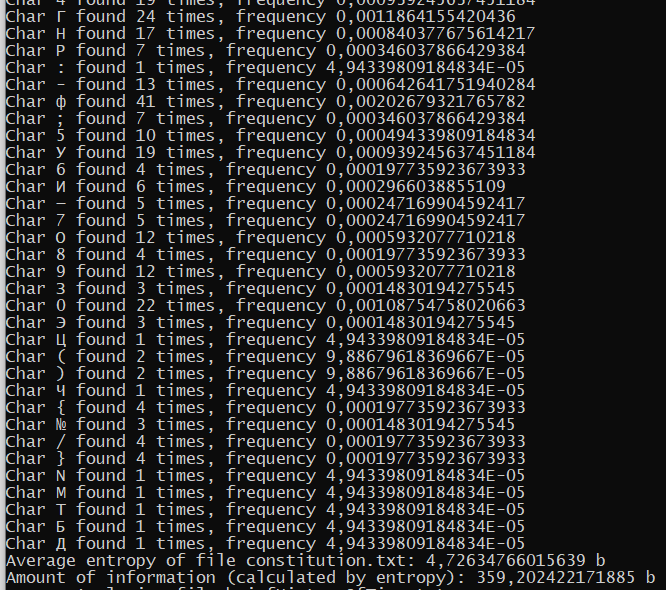




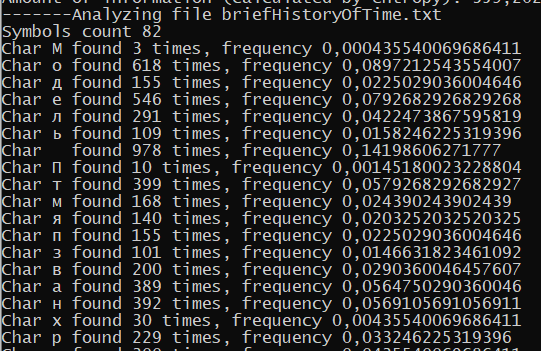


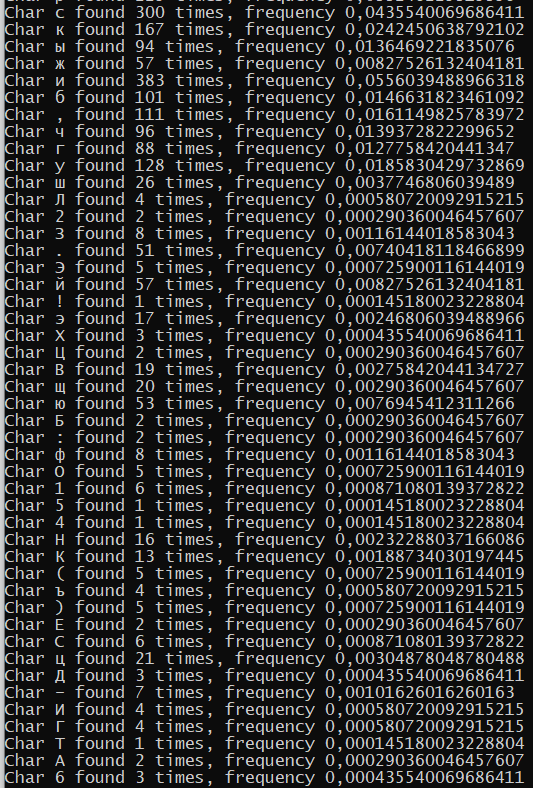
**Constitution.txt**

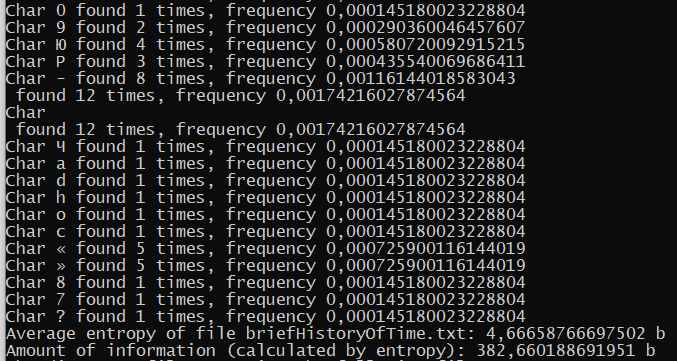


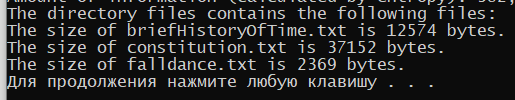


**briefHistoryOfTime:**

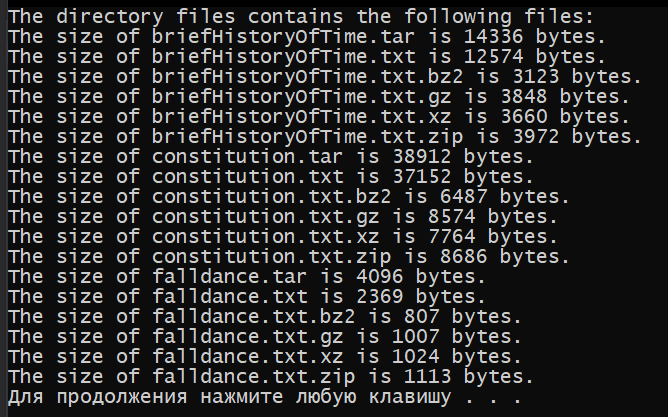
****





****

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)

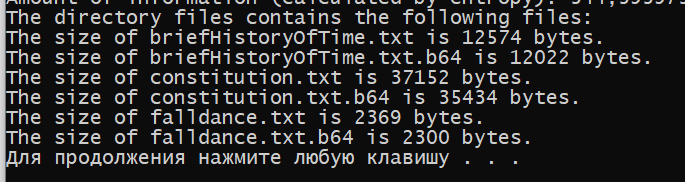


Розміри файлів

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Файл | .txt | ZIP | gzip | LZMA2(.xz) | bzip2 | tar | К-кість інформації (за доп. ентропії) |
| briefHistoryOfTime.txt | 12574 | 3972 | 3848 | 3660 | 3123 | 14336 | 382,7 |
| constitution.txt | 37152 | 8686 | 8574 | 7764 | 6487 | 38912 | 359,2 |
| falldance.txt | 2369 | 1113 | 1007 | 1024 | 807 | 4096 | 4,833 |

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



falldance.txt.b64



constitution.txt.b64



briefHistoryOfTime.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 |
| briefHistoryOfTime.txt | 12574 | 12022 | 4.667 | 5.381 | 382.7 | 344.39 |
| constitution.txt | 37152 | 35434 | 4.726 | 5.385 | 359.2 | 344.67 |
| falldance.txt | 2639 | 2300 | 4.833 | 5.412 | 256.1 | 346.37 |

**Висновок:** в лабораторній роботі було вивчено алгоритми вираховування середної ентропії та кількості інформації, алгоритму кодування бінарного коду в текстовий формат BASE64.

Github: <https://github.com/HleBASS/CS_labs>