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 - МА

Співак Богдан Олегович

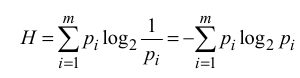
Київ 2021

**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

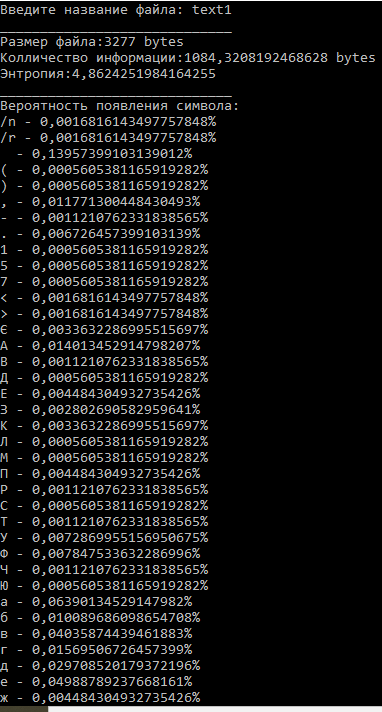
1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування

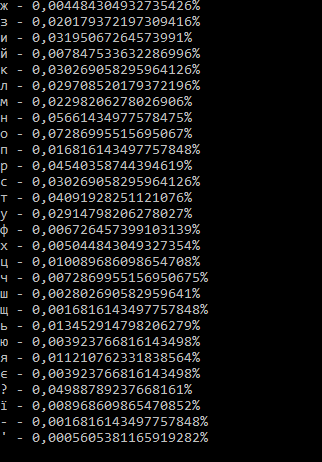
<https://uaf.ua/article/42189> <http://svatovo.ws/stati_uk_glasses.html> <https://onlyart.org.ua/children/kazky/ukrayinska-narodna-kazka-tsap-ta-baran/>

<https://github.com/pilmenchik/CS>

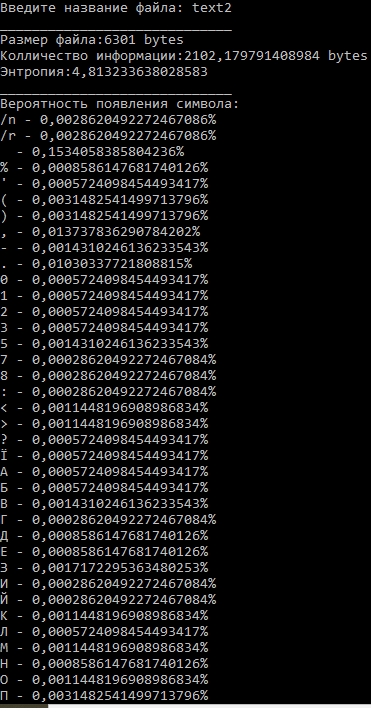
1. Переконайтесь, що тексти, які ви використовуєте є унікальними і не повторюються у ваших колег! Використовуйте наявні електронні засоби зв’язку та документообігу, щоб уникнути дублювання! Вдруге аналіз того самого тексту не зараховується!
2. Створіть програму (будь-якою зручною для вас мовою), яка 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

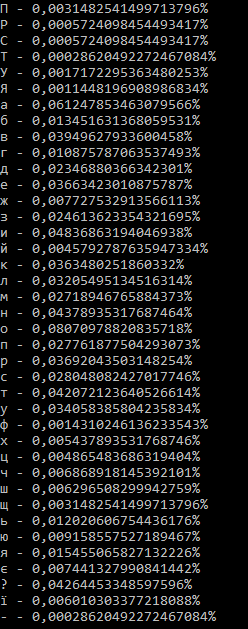
Перший файл:



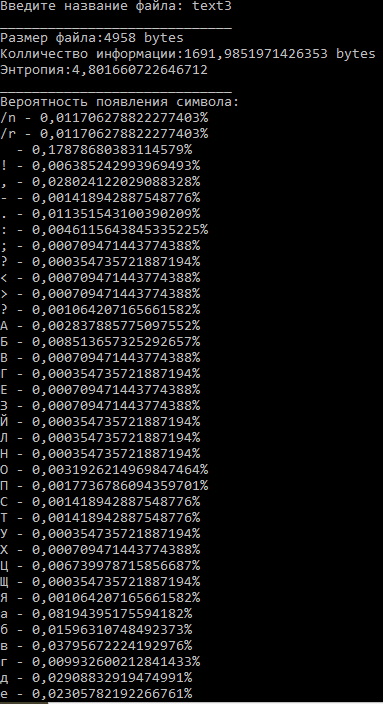


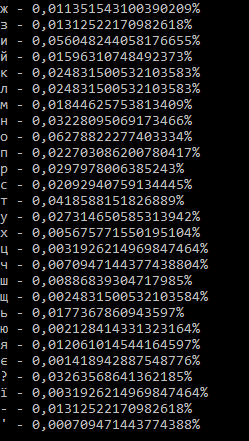
Другий файл:





Третій файл:





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)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Назва | Розмір  файлу | Кількість інформації | 7z | rar | bz2 | xz | zip |
| text1 | 3277 | 1084 | 1315 | 1311 | 1044 | 1248 | 1348 |
| Text2 | 6301 | 2102 | 2342 | 2410 | 1921 | 2272 | 2423 |
| text3 | 4958 | 1692 | 1837 | 1868 | 1456 | 1768 | 1885 |

Найбільш близькі до кількості інформації алгоритми- це bz2, xz, тому ці алгоритми можна вважати кращими за інші наведені у таблиці.

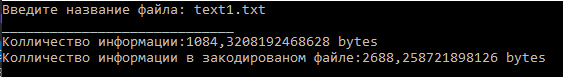
# 2. Дослідження способів кодування інформації на прикладі Base64

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

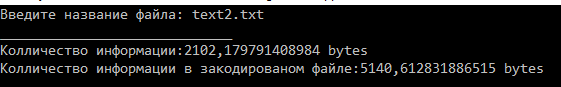
а.Обрахуйте кількість інформації в base64-закодованому варіанті файлу

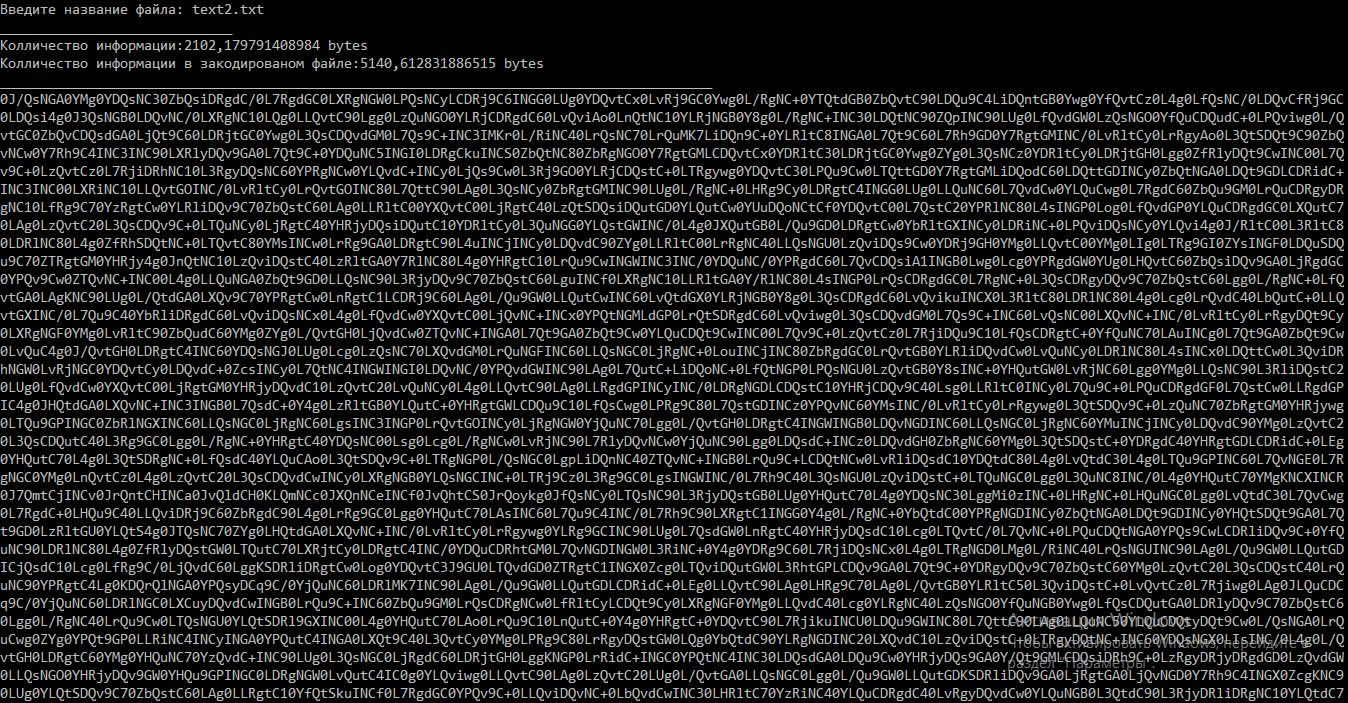
b.Порівняйте отримане значення з кількістю інформації вихідного файлу

c.Зробіть висновки з отриманого результату

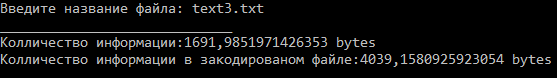












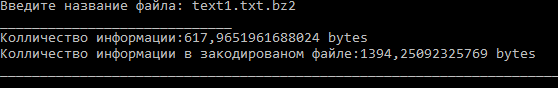


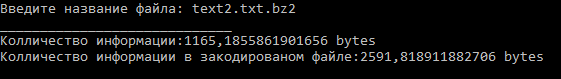
1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли

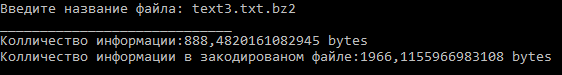
а. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу

b.Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу

c.Зробіть висновки з отриманого результату







|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | text1.txt | text1.txt.bz2 | text2.txt | text2.txt.bz2 | text3.txt | text3.txt.bz2 |
| Кількість інформації | 1084 | 617 | 2102 | 1165 | 1692 | 888 |
| К-ть інформації Base64 | 2688 | 1394 | 5140 | 2591 | 4039 | 1966 |

## **Вихідні коди розроблених програм завантажте в свій репозиторій на GitHub.**

## **В Moodle завантажте звіт, що містить:**

* результати проведеного аналізу кількості інформації обраних текстів (самі тексти в вигляді посилань або в додатках)
* посилання на програму в GitLab gl.vlabs.knu.ua
* приклад роботи створеної програми для підрахунку кількості інформації
* приклад роботи створеної програми для кодування в Base64

***Висновок:***

В данній лабораторній роботі було досліджене поняття середньої ентропії та кількості інформації. Було досліджено декілька алгоритмів стиснення, а саме зміна кількості інформації. Також було досліджено та реалізовано алгоритм кодування Base64 , та його вплив на кількість інформації.