## 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

Поліщук Богдан Миколайович

## 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****:*

https://lh6.googleusercontent.com/Jw8pwa_P8wurUquJoka9fS3K7o6GJJ0BKedhb0yOvsHjHs1DwmMZ9o6SzT7g_OgUBpwHxe2FLJCHs3BYFJKW8Fz2FCCHXxlGsq4wldEdp5YamqOsDKm3XnWEANk1ZFWy8DVYPMXJ

*де m - кількість символів алфавіту, p - імовірність появи символу Ентропія вимірюється в* ***БІТАХ*** *(як представлення кількості можливих варіантів).*

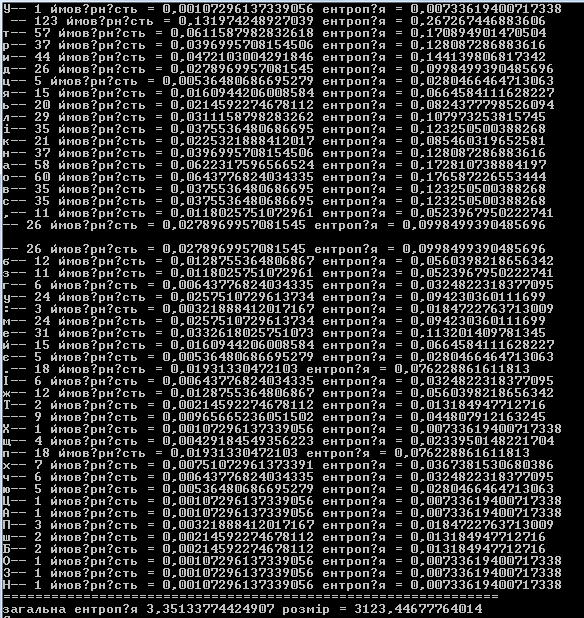
***Кількість інформації в тексті*** *- середня ентропія вихідного алфавіту помножена на кількість символів тексту. (****HINT****: результат обрахунку для порівняння значення з розміром файлів треба перевести з бітів в байти)*

## Хід виконання роботи:

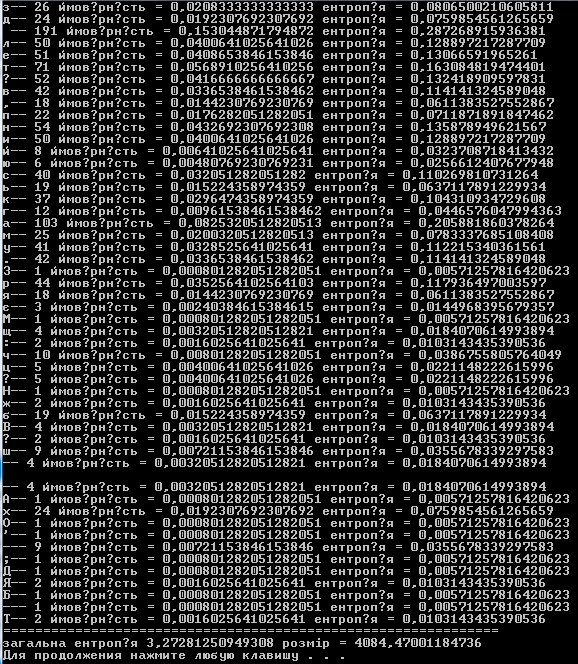
Дослідження кількості інформації в тексті

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)
   * text1.txt – Василь Стус – У тридцять лiт ти тiльки народився.
   * text2.txt – уривок із Intermezzo.
   * text3.txt – Англійська казка.
2. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
   * **обраховує частоти (імовірності) появи символів в тексті**
   * **обраховує середню ентропію алфавіту для даного тексту**
   * **виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів**
   * **виводить на екран значення частот, ентропії та кількості інформації 4. 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).**

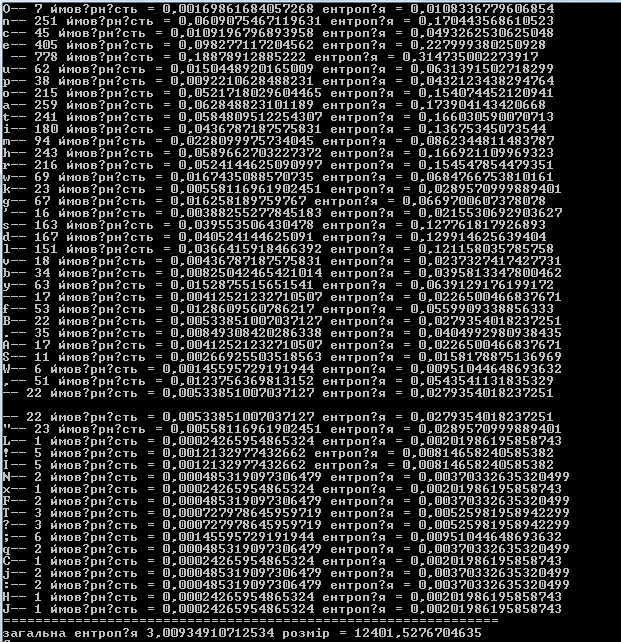
Аналіз van***.txt***



Аналіз ***int.txt***



Аналіз tale***.txt***



Ентропія:

van.txt – 3,35133774424907

int.txt – 3,27281250949308

tale.txt – 3,00934910712534

Розмір файлу(в байтах):

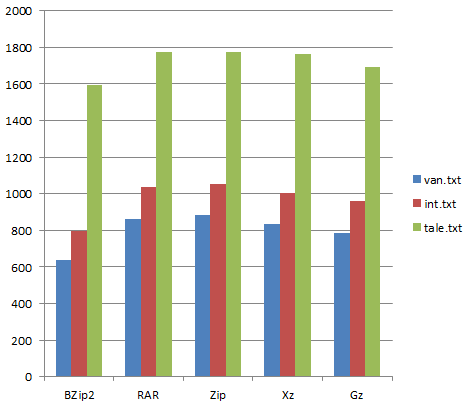
van.txt – 3123біт/8 = 390 байт

int.txt – 4084біт/8 = 510 байт

tale.txt – 12401біт/8 = 1550 байт

Результати стиснення

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Назва файлу | BZip2 | RAR | Zip | Xz | Gz | Кількість інформації |
| van.txt | 638 | 860 | 881 | 836 | 787 | 390 |
| int.txt | 795 | 1037 | 1053 | 1004 | 959 | 510 |
| tale.txt | 1597 | 1777 | 1777 | 1764 | 1692 | 1550 |



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).

Перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)

1. Закодуйте в Base64 обрані вами текстові файли

Обрахуйте кількість інформації в base64-закодованому варіанті файлу Порівняйте отримане значення з кількістю інформації вихідного файлу Зробіть висновки з отриманого результату

1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли

Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу

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

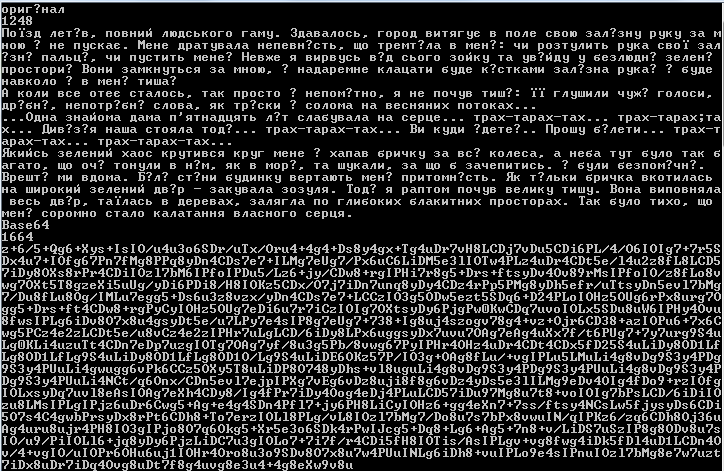
Зробіть висновки з отриманого результату

## Хід виконання роботи:

Base64 ***van.txt***



Base64 ***int.txt***

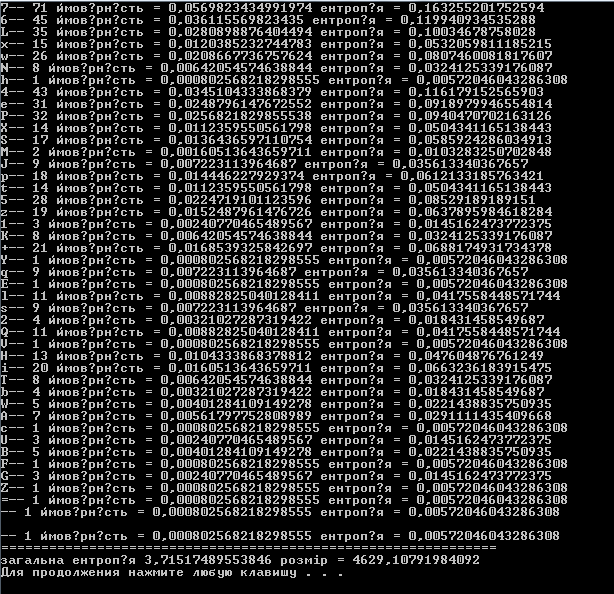


Base64 ***tale.txt***

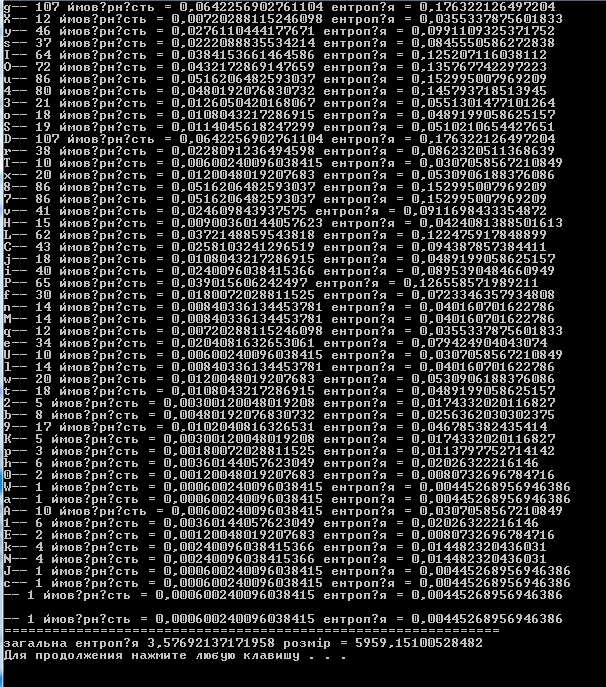




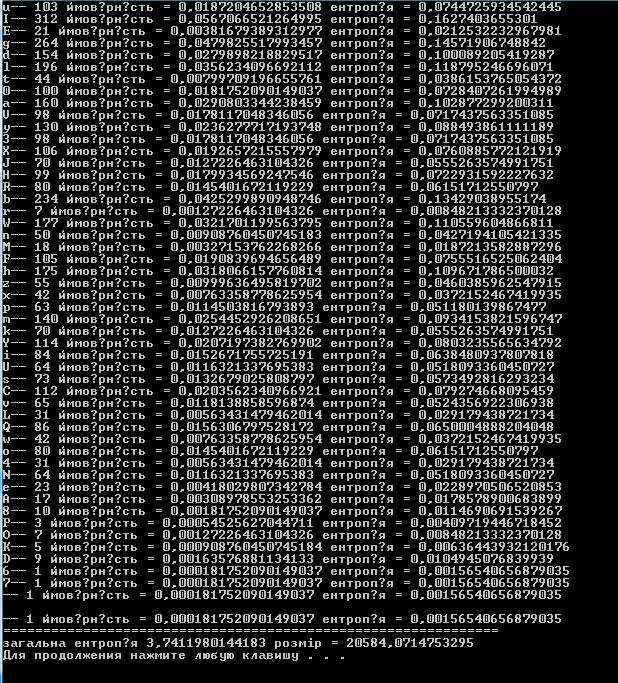
**Ентропія vanencr.txt**



**Ентропія intencr.txt**



**Ентропія taleencr.txt**



# **Розмір файлу :**

vanencr.txt = 4629bit/8 = 578 байт

intencr.txt = 5959 bit/8 = 745 байт

taleencr.txt = 20584bit/8 = 2573 байт

**Ентропія:**

van.txt – 3,35133774424907

int.txt – 3,27281250949308

tale.txt – 3,00934910712534

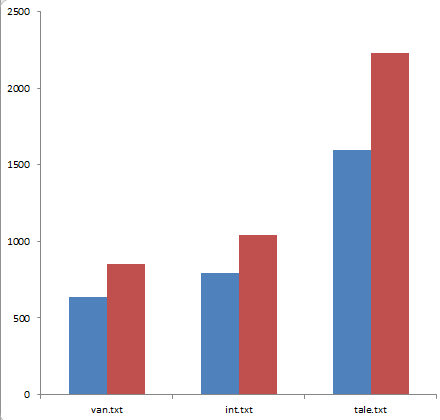
**Розмір файлу:**

van.txt – 3123біт/8 = 390 байт

int.txt – 4084біт/8 = 510 байт

tale.txt – 12401біт/8 = 1550 байт

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Назва файлу | Ентропія | Довжина | Кількість інформації (біт) | |  |
| text1encrypted.txt | 3,7151 | 1244 | 4629 | |
| text2encrypted.txt | 3,5769 | 1664 | 5959 | |
| text3encrypted.txt | 3,7411 | 5500 | 20584 | |
| Назва файлу | BZip2(BASE64) | Кількість інформації | |
| text1encrypted.txt | 853 | 578 | |
| text2encrypted.txt | 1039 | 745 | |
| text3encrypted.txt | 2232 | 2573 | |



BZIP2 - ▉

BZIP2(Base64) - ▉

**Висновок:** В даній лабораторній роботі я працював з аналізом тексту. Навчився визначати ентропію і працювати з алгоритмом Base64. Також було досліджено , що кількість інформації займає менше ніж сам стиснутий який зберігається на диску і було виявлено що найкращий алгоритм зжимання це BZip2.

Github: <https://github.com/Zhukbo>