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

Хоменко Іван Валентинович

Київ 2020

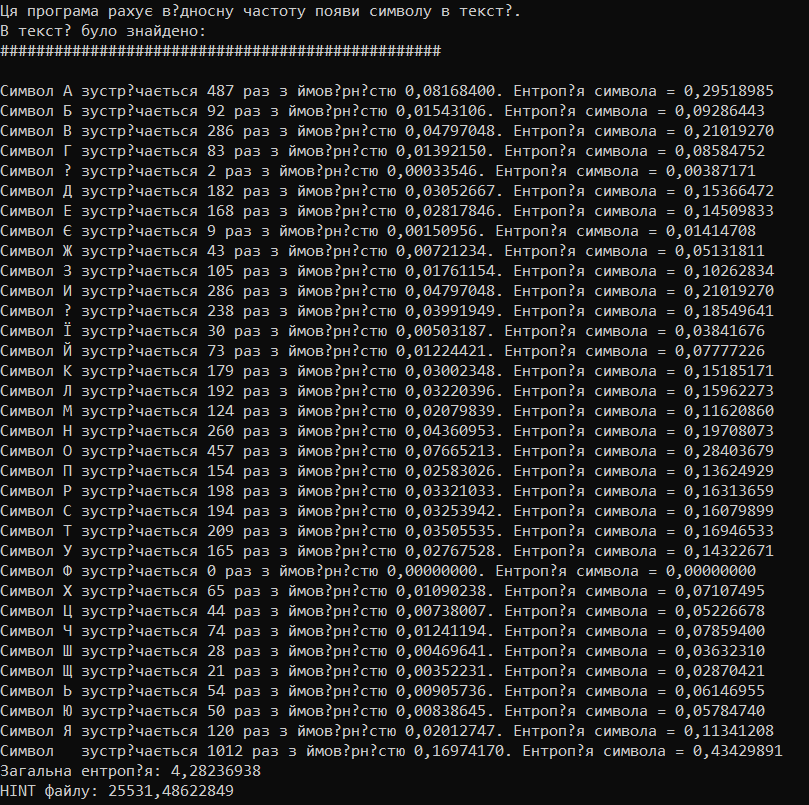
*Мета: Дослідити імовірнісні параметри української мови для оцінки кількості інформації текстів. Дослідити вплив різних методів кодування інформації на її кількість.*

Хід виконання роботи:

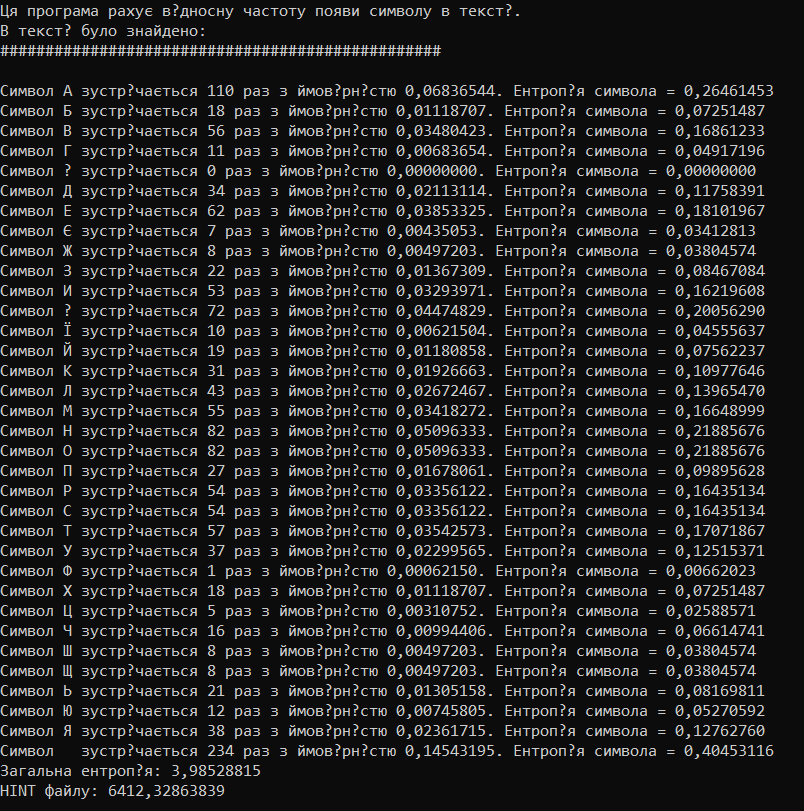
Дослідження кількості інформації в тексті

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)
   * + text\_1.txt –уривок роману Панаса Мирного «Хіба ревуть воли як ясла повні…»
     + text\_2.txt –вірш Івана Франка «Тричі мені являлася любов».
     + text\_3.txt – уривок з біографії Панаса Мирного».
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).

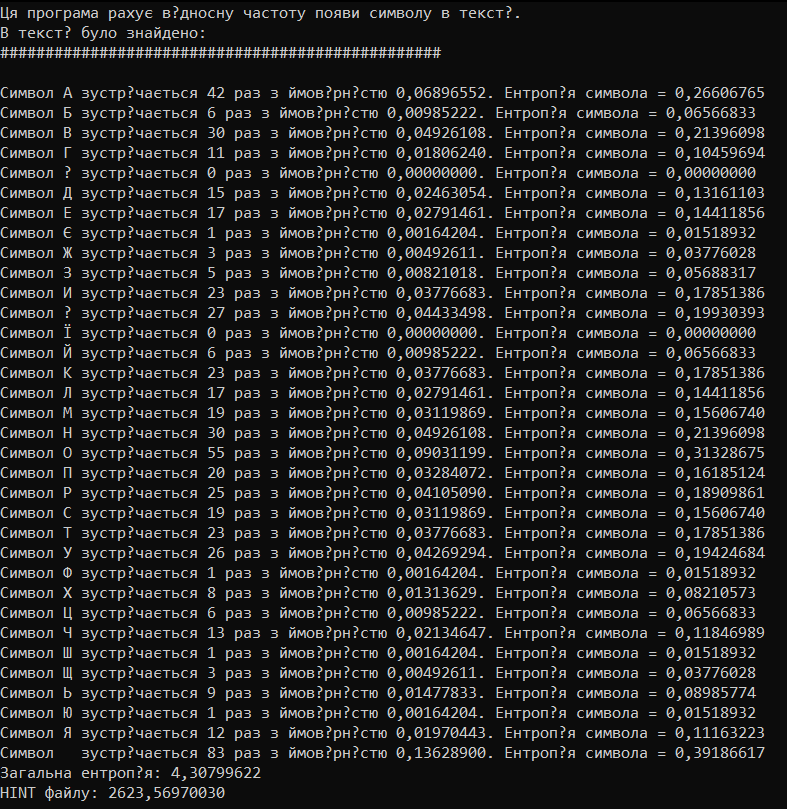
Перевіримо 1 файл



Перевіримо 2 файл



Перевіримо 3 файл



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Назва файлу | BZip2 | 7Z | ZIP | XZ | GZIP | Розмір файлу без стиснення |
| text1.txt | 2974 | 3700 | 3821 | 3624 | 3845 | 3191.375 |
| text2.txt | 979 | 1289 | 1308 | 1212 | 1210 | 801.5 |
| text3.txt | 438 | 641 | 638 | 564 | 514 | 327.875 |

Результати стиснень

# 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)
3. 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

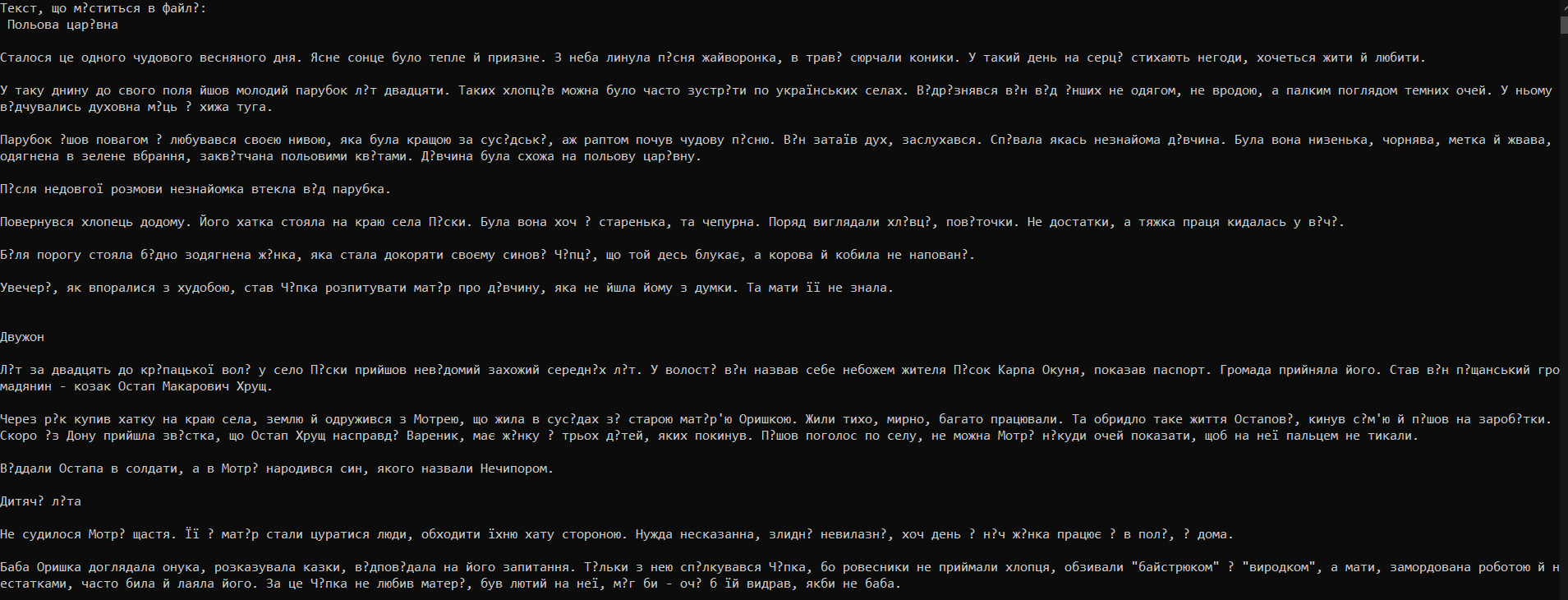
1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли

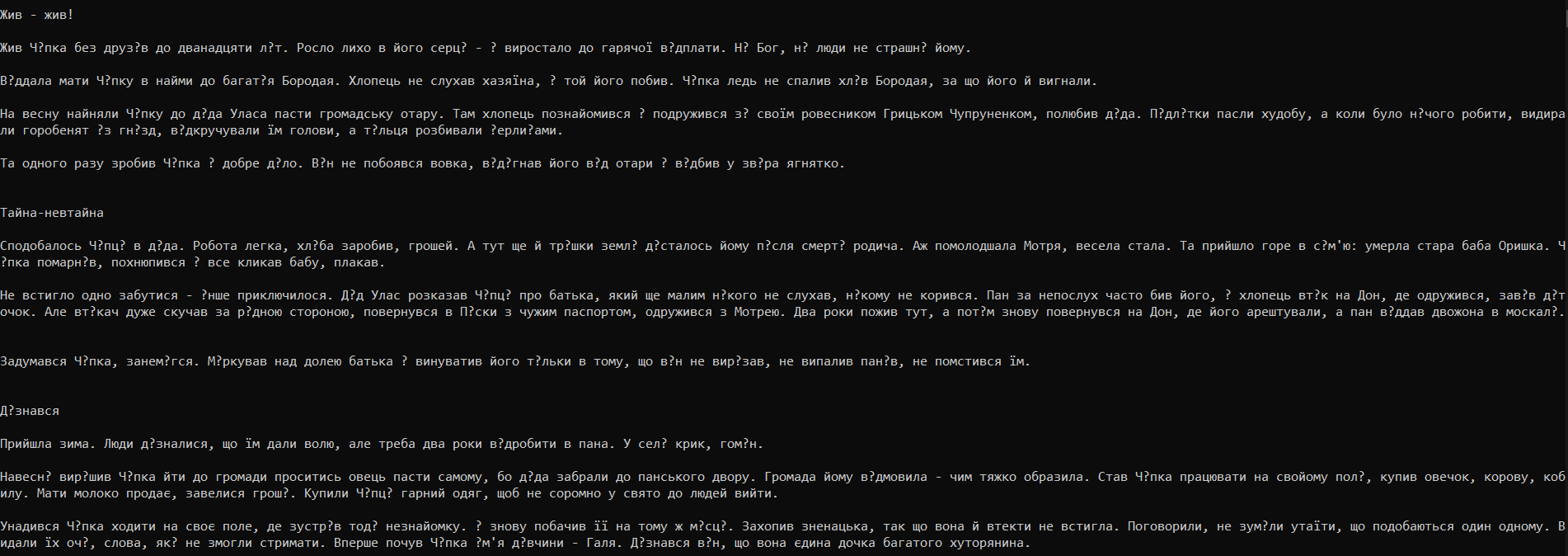
Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу

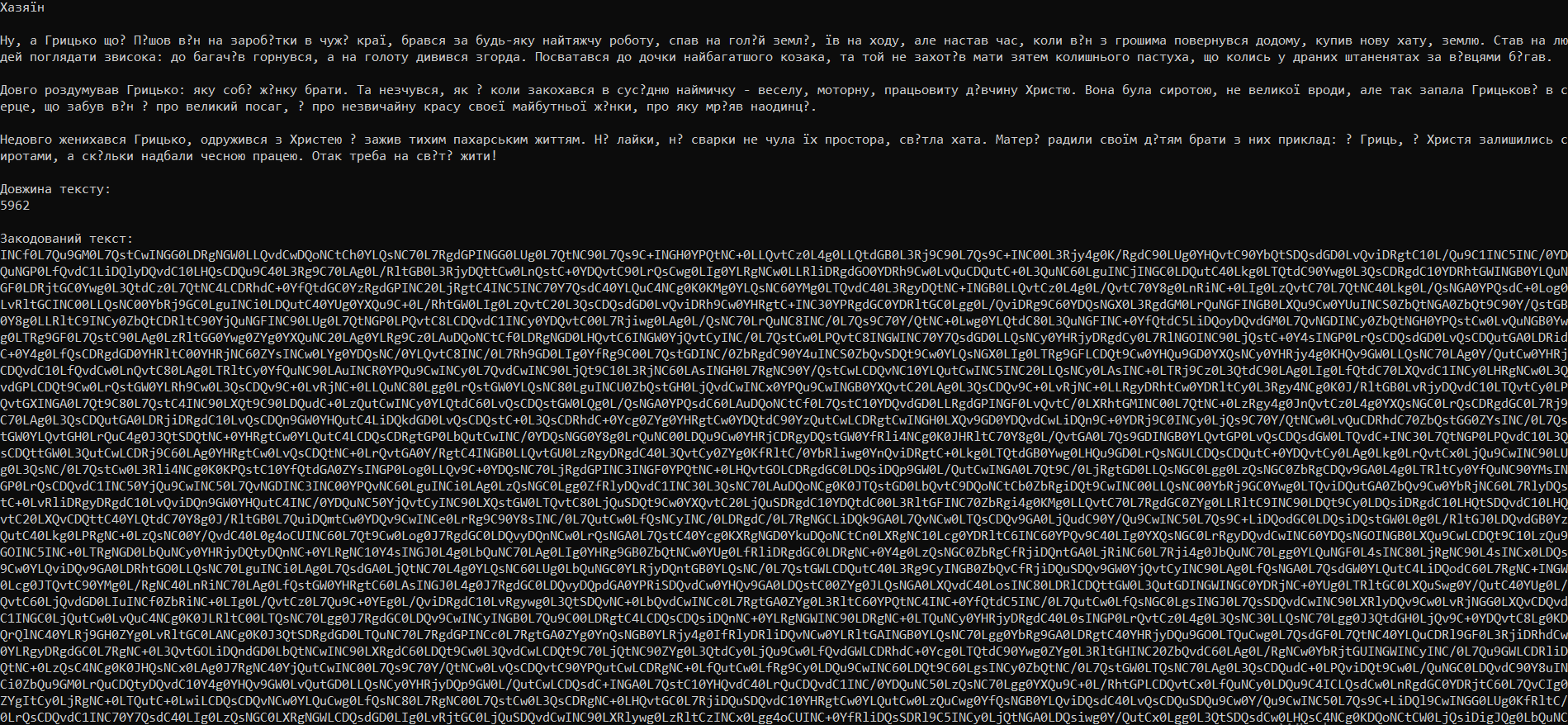
Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу

Зробіть висновки з отриманого результату

Кодуваня 1 файлу

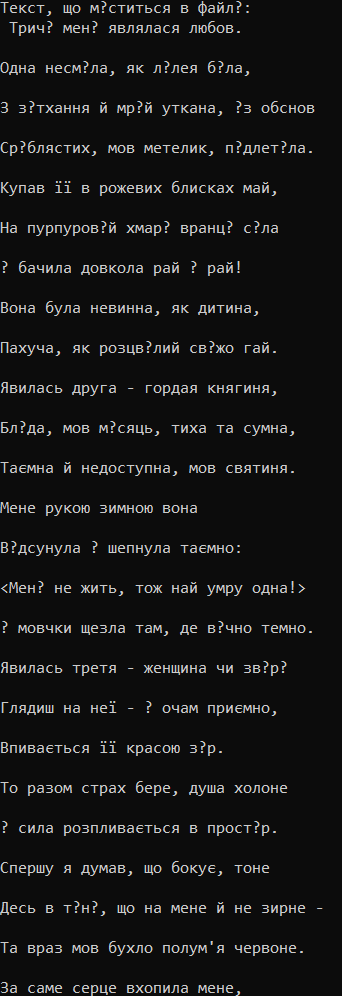
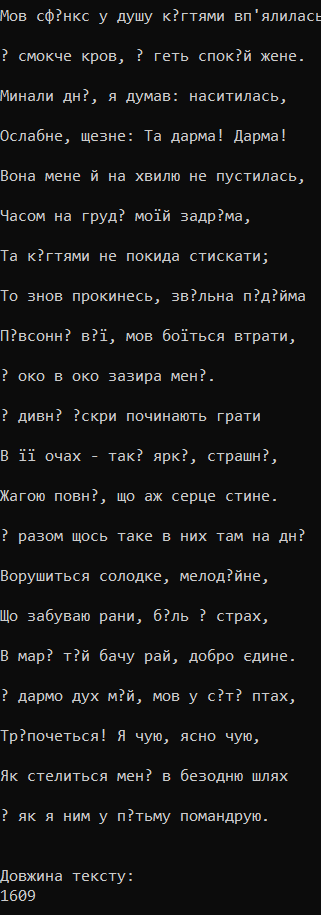


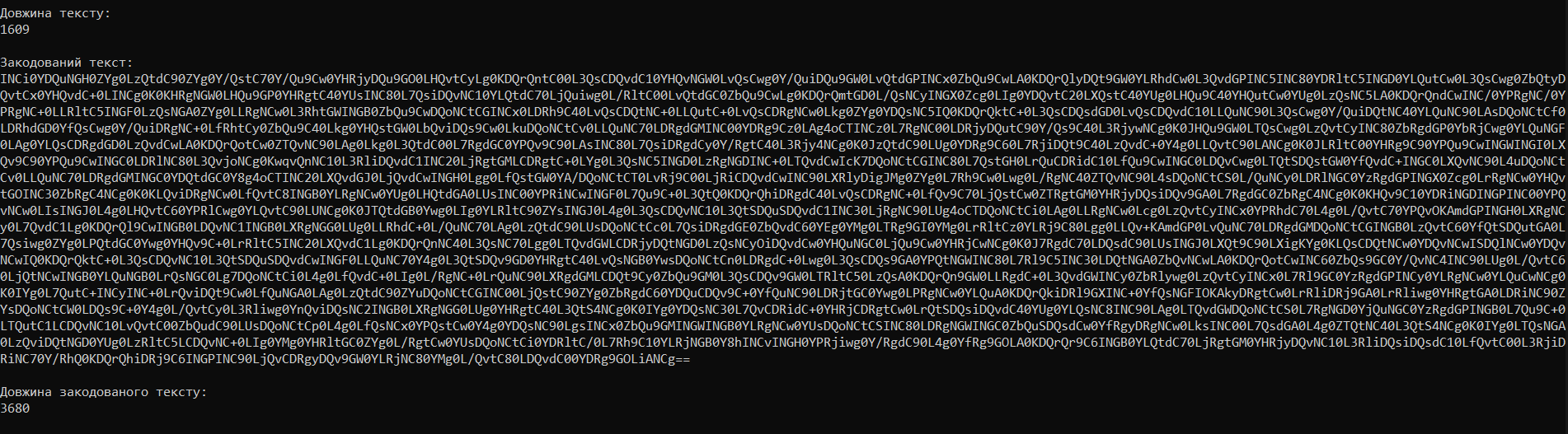




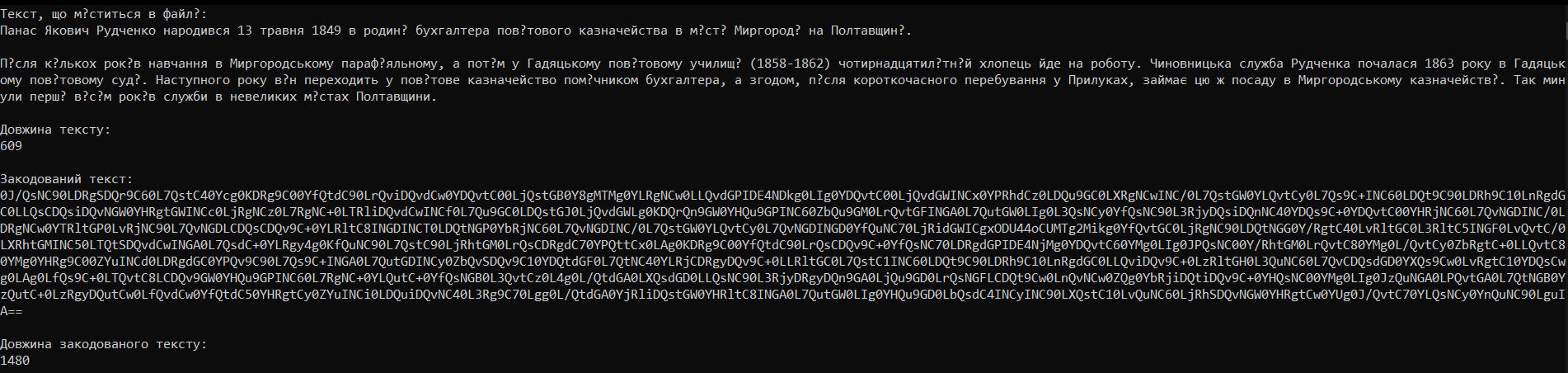


Кодування 2 файлу



Кодування 3 файлу



Розмір та ентропія першого файлу



Розмір та ентропія Другого файлу



Розмір та ентропія третього файлу



|  |  |  |  |
| --- | --- | --- | --- |
| Назва файлу | BZip2 | Розмір оригіналу | Розмір файлу закодованого |
| text1\_1.txt | 3765 | 3191.375 | 7277 |
| text2\_1.txt | 1295 | 801.5 | 1871,625 |
| text3\_1.txt | 659 | 327.875 | 762,5 |

|  |  |  |  |
| --- | --- | --- | --- |
| Назва файлу | BZip2 кількість інформації | Розмір оригіналу | Розмір файлу без стиснення |
| text1\_1.txt | 907,75 | 3191.375 | 7277 |
| text2\_1.txt | 307,8 | 801.5 | 1871,625 |
| text3\_1.txt | 152,125 | 327.875 | 762,5 |

Висновок: в ході лабораторної роботи було розроблено програма, яка аналізує текст та обчислює імовірність появи букв в тексті, ентропію, загальний обсяг інформації. Було здійснено стиснення текстів та помічено, що обсяг інформації є манший ніж стиснутий файл. Було створено програму, яка закодовує текст в Base64, при повторному кодуванні обсяг інформації збільшується.

Посилання на GitHub: