**Київський національний університет імені Тараса Шевченка**

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

**Комп’ютерні системи**

Лабораторна робота № 1

Тема: «Дослідження кількості інформації при різних варіантах кодування»

Роботу виконав

студент 3 курсу

спеціальності “КІ-СА”

Ситниченко Денис Вікторович

Київ 2021

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

**Хід виконання роботи:**

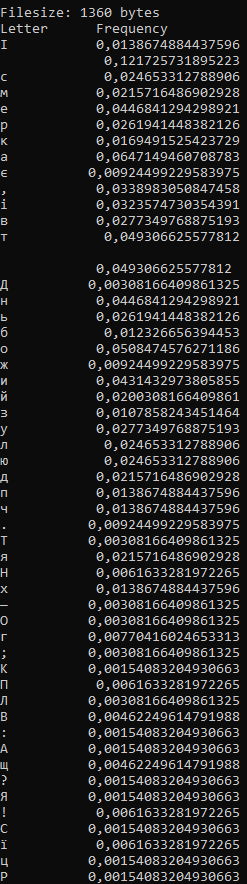
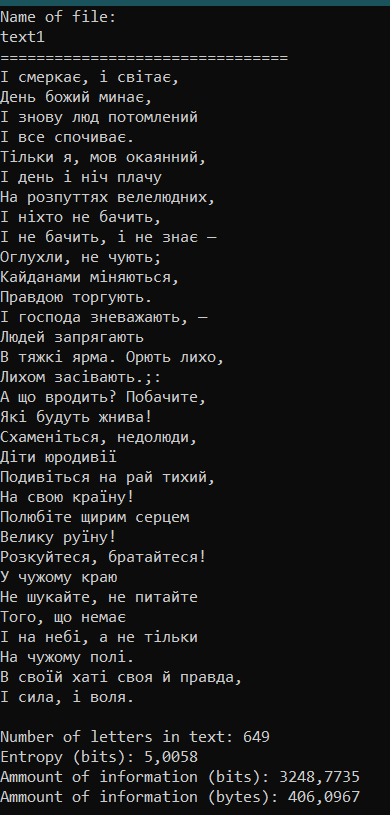
1. Дослідження кількості інформації в тексті
2. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI)

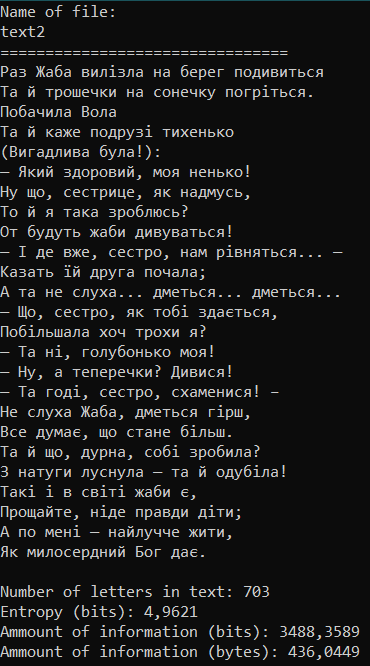
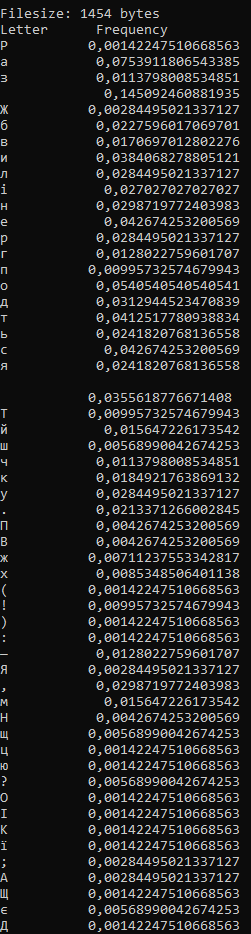
* [І МЕРТВИМ, І ЖИВИМ, І НЕНАРОЖДЕННИМ](https://www.ukrlib.com.ua/books/printit.php?tid=708)
* [Жаба й Віл](https://www.ukrlib.com.ua/books/printit.php?tid=2049)
* [Радіотехніка](https://uk.wikipedia.org/wiki/%D0%A0%D0%B0%D0%B4%D1%96%D0%BE%D1%82%D0%B5%D1%85%D0%BD%D1%96%D0%BA%D0%B0)

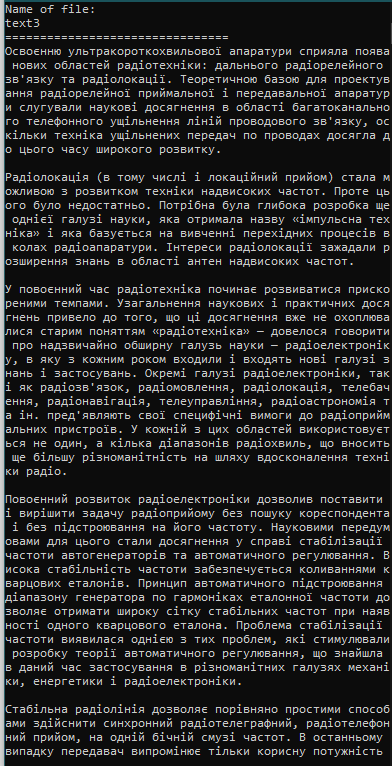
1. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
   1. обраховує частоти (імовірності) появи символів в тексті
   2. обраховує середню ентропію алфавіту для даного тексту
   3. виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
   4. виводить на екран значення частот, ентропії та кількості інформації

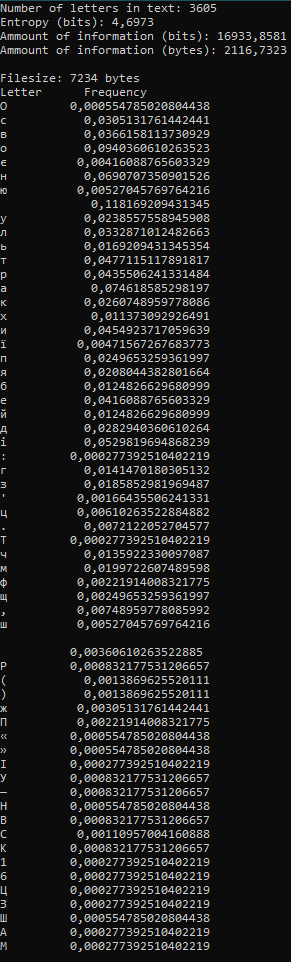
Код розташований за посиланням:

<https://gl.vlabs.knu.ua/frecs/ce/cs/2020-2021/DenysS/lab1>

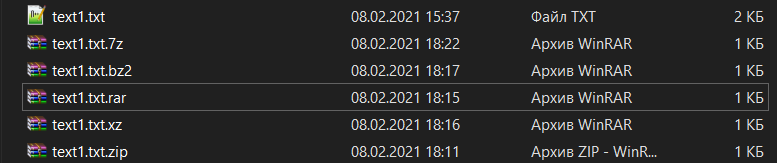
**Текст №1**([І МЕРТВИМ, І ЖИВИМ, І НЕНАРОЖДЕННИМ](https://www.ukrlib.com.ua/books/printit.php?tid=708))

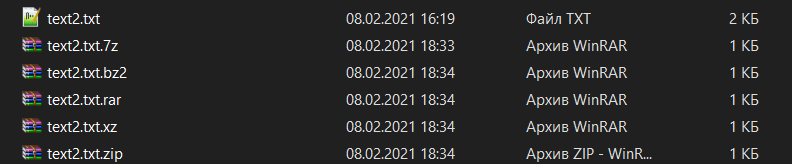
******Текст №2**([Жаба й Віл](https://www.ukrlib.com.ua/books/printit.php?tid=2049))

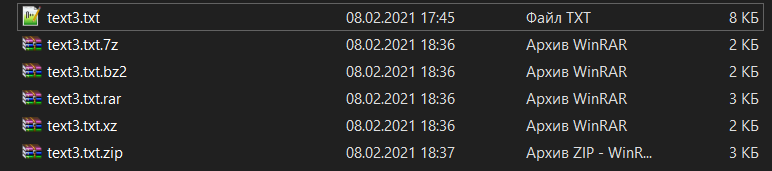
**Текст №3**([Радіотехніка](https://uk.wikipedia.org/wiki/%D0%A0%D0%B0%D0%B4%D1%96%D0%BE%D1%82%D0%B5%D1%85%D0%BD%D1%96%D0%BA%D0%B0))



1. Текстові файли було стиснуто алгоритмами **zip**, 7**z**, **bzip2, xz, rar**

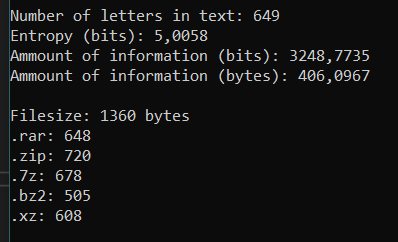




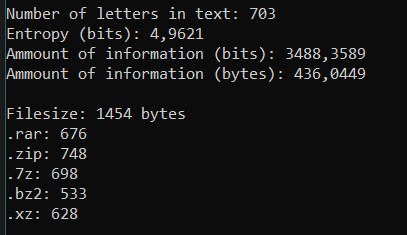


1. Порівняйте результуючі обсяги архівів з обчисленою кількістю інформації та **наведіть у звіті висновки** щодо кореляції цих величин для обраних вами файлів (яка відмінність, що вийшло більше і чому)

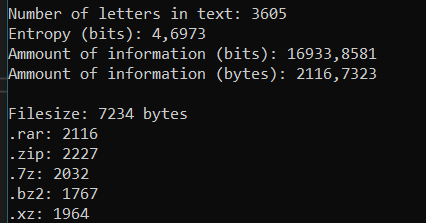
**Текст №1**([І МЕРТВИМ, І ЖИВИМ, І НЕНАРОЖДЕННИМ](https://www.ukrlib.com.ua/books/printit.php?tid=708))



**Текст №2**([Жаба й Віл](https://www.ukrlib.com.ua/books/printit.php?tid=2049))



**Текст №3**([Радіотехніка](https://uk.wikipedia.org/wiki/%D0%A0%D0%B0%D0%B4%D1%96%D0%BE%D1%82%D0%B5%D1%85%D0%BD%D1%96%D0%BA%D0%B0))



Таблиця розмірів файлів та об’ємів інформації в байтах

|  |  |  |  |
| --- | --- | --- | --- |
| **Файл** | text1.txt | text2.txt | text3.txt |
| **Розмір** | 1360 | 1454 | 7234 |
| **Ентропія** | 5,01 bits | 4,96 bits | 4,69 bits |
| **Інформація** | 406,09 | 436,04 | 2116,73 |
| **zip** | 720 | 748 | 2227 |
| 7**z** | 678 | 698 | 2032 |
| **bzip2** | 505 | 533 | 1767 |
| **xz** | 608 | 628 | 1964 |
| **rar** | 648 | 676 | 2116 |

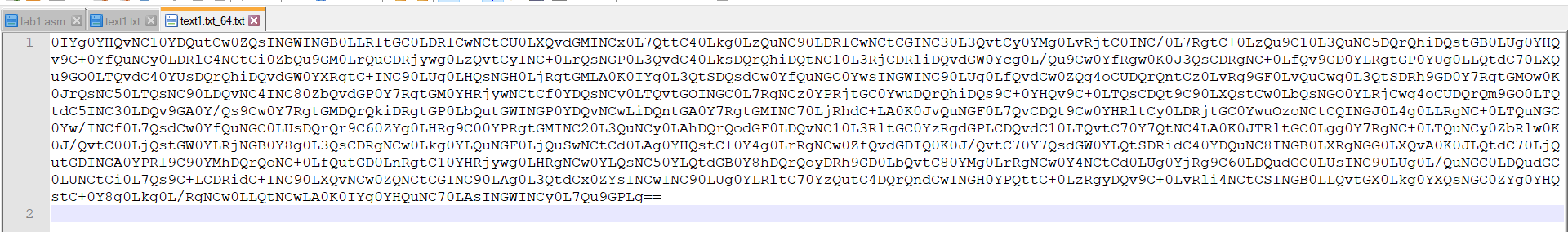
**Висновок**: як видно з таблиці та діаграми реальний розмір архівів виявився досить близьким до апроксимованого значення інформації для текстового файлу великого розміру(text3.txt), і в середньому був навіть дещо менше, що демонструє ефективність сучасних алгоритмів стиснення. Для менших файлів(text1.txt та text2.txt) реальний розмір архівів виявився більшим ніж апроксимоване значення інформації, це пояснюється тим, що розмір текстових файлів менший, тому й службова інформація, що додається при стисненні є більш вагомою в порівнянні з великим текстовим файлом, також так як файл малий - то менше повторюваної інформації, яку можна замінити при стисненні. Серед всіх алгоритмів використаних для порівняння **bzip2** показав себе найкраще в стисненні файлів, але на роботу цього методу стиснення йшло більше часу ніж на інші алгоритми.

1. Дослідження способів кодування інформації на прикладі Base64
2. Ознайомтесь зі стандартом [RFC4648](https://tools.ietf.org/html/rfc4648)
3. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)

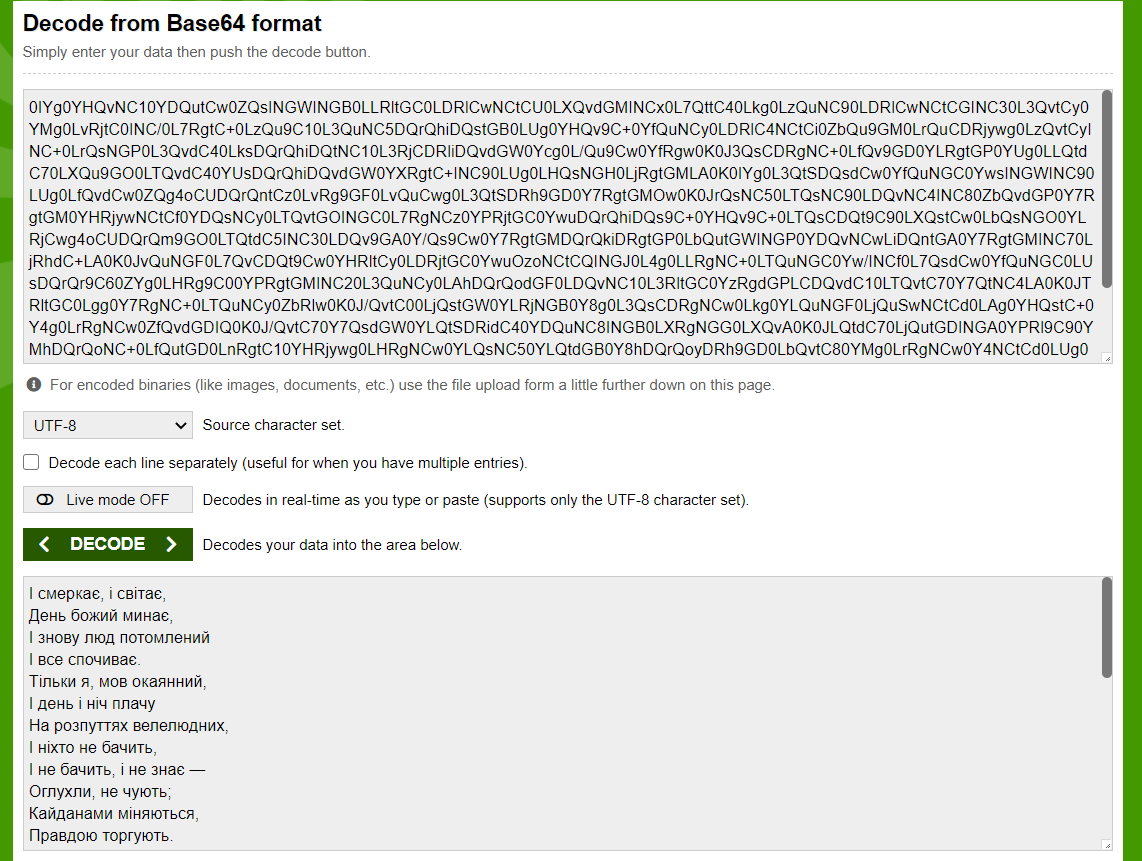
* перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)

Було створено програму, яка кодує довільні файли в base64 шляхом реалізації алгоритму вручну

Вміст файлу text1.txt\_64 після кодування в Base64:

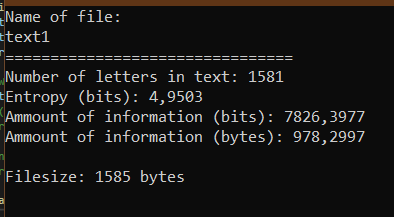


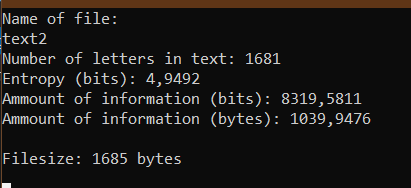
Для перевірки правильності кодування, текст було декодовано за допомогою [base64decode](https://www.base64decode.org/):

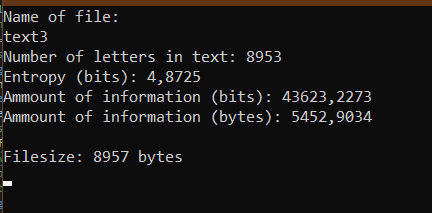


Як можна побачити, все було правильно декодовано.

1. Закодуйте в Base64 обрані вами текстові файли
2. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
3. Порівняйте отримане значення з кількістю інформації вихідного файлу
4. Зробіть висновки з отриманого результату







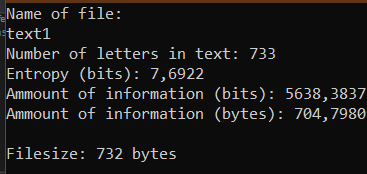
Таблиця об’ємів інформації в байтах оригінального і закодованого файлів

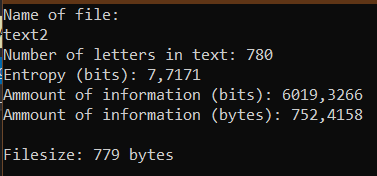
|  |  |  |  |
| --- | --- | --- | --- |
| **Файл** | text1.txt | text2.txt | text3.txt |
| **Інформація** | 406,09 | 436,04 | 2116,73 |
| **Інформація(Base64)** | 978,29 | 1039,94 | 5452,90 |

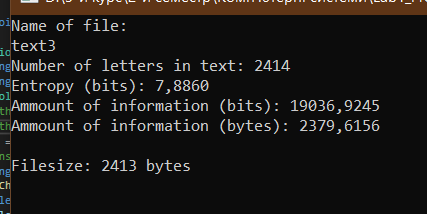
**Висновок:** згідно таблиці та діаграми після кодування Base64 к-сть інформації зросла в середньому в 2.66 рази. Це пояснюється тим, що кожні 3 байти оригінального файлу кодуються 4 –ма символами (збільшення на ¹/₃). Також могло також вплинути на збільшення кінцевої інформації те, що після кодування кожен байт рахується як окремий символ - це і дало таке велике збільшення інформації.

1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
2. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
3. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу
4. Зробіть висновки з отриманого результату

Текстові файли були закодовані в **bzip2**, так як він показав себе найкраще







|  |  |  |  |
| --- | --- | --- | --- |
| **Файл** | text1.txt | text2.txt | text3.txt |
| **Інформація** | 406,09 | 436,04 | 2116,73 |
| **Інформація(Base64 archived)** | 704,79 | 752,41 | 2379,61 |

Як видно з таблиці і графіку overhead у файлів малого розміру складає більшу частку розміру ніж у великих файлів.

**Висновок:** було досліджено що таке ентропія і як її підраховувати, також було розглянуто як працюють алгоритми стиснення і як вони впливають на кінцевий розмір файлу. Згідно отриманих результатів можна зробити висновок, що в архівах інформація зберігається майже без надлишкових даних і в максимально стиснутому виді.