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 III курсу
КІ-МА
Грищук Олександр

Київ 2020

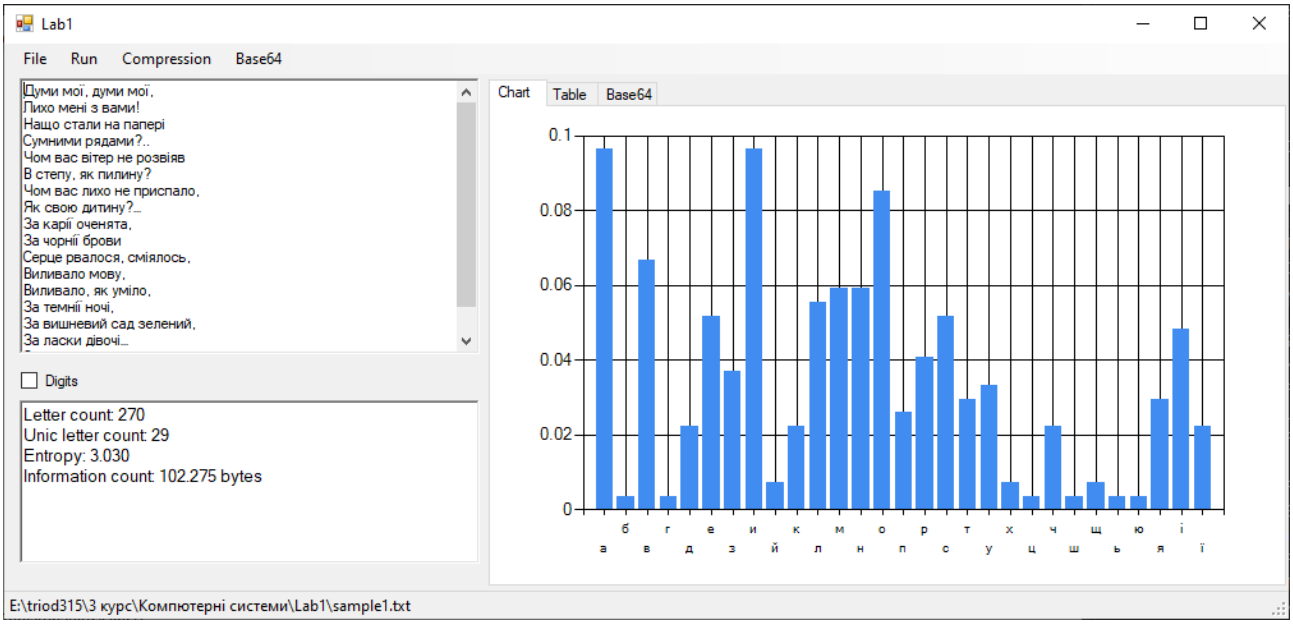
Хід виконання роботи:

1) Дослідження кількості інформації в тексті
   1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування

      Sample1.txt – вірш Т. Г. Шевченка «Думи мої думи»
      https://github.com/triod315/CS/blob/master/Lab1/sample1.txt
      Sample2.txt – фрагмент статті про PHP з lurkore.to
      https://github.com/triod315/CS/blob/master/Lab1/sample2.txt
      Sample3.txt – RFC 2795 (IMPS)
      https://github.com/triod315/CS/blob/master/Lab1/sample2.txt

   2. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
      a. обраховує частоти (імовірності) появи символів в тексті
      b. обраховує середню ентропію алфавіту для даного тексту
      c. виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
      d. 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).

      https://github.com/triod315/CS/tree/master/Lab1

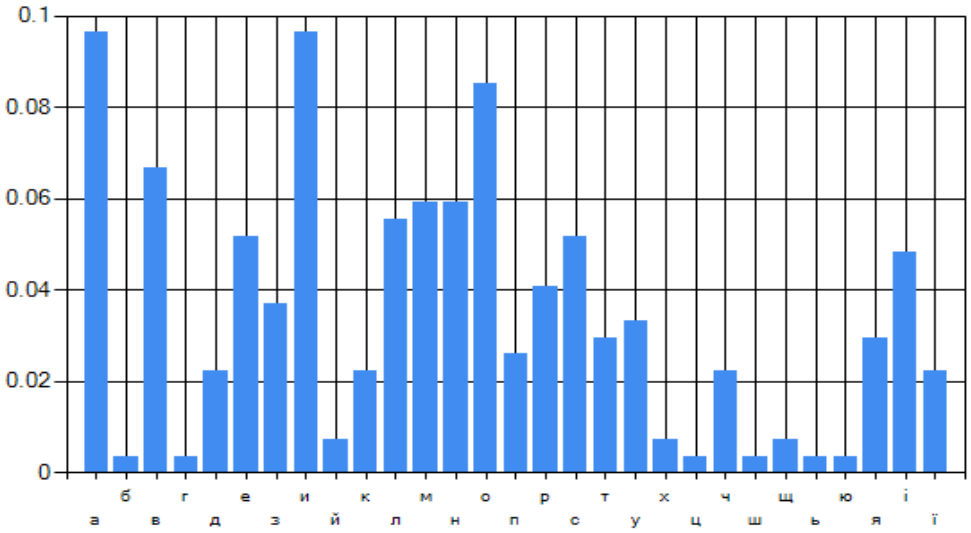


   3. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення).
   4. Порівняйте результуючі обсяги архівів з обчисленою кількістю інформації та наведіть у звіті висновки щодо кореляції цих величин для обраних вами файлів (яка відмінність, що вийшло більше і чому)

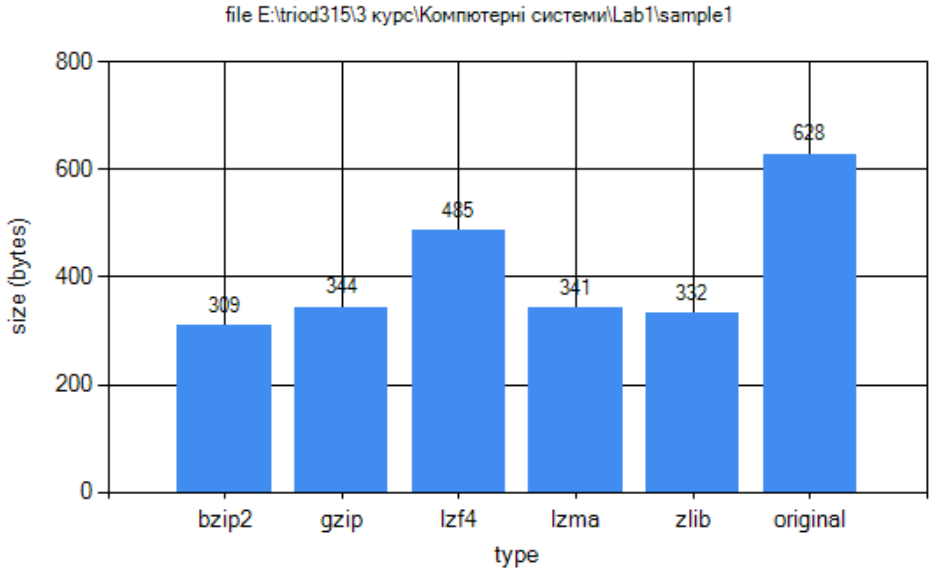
Результати:

1) Sample1

Кількість літер: 270
Кількість унікальних літер: 29
Ентропія: 4.372
Кількість інформації: 147.55 bytes



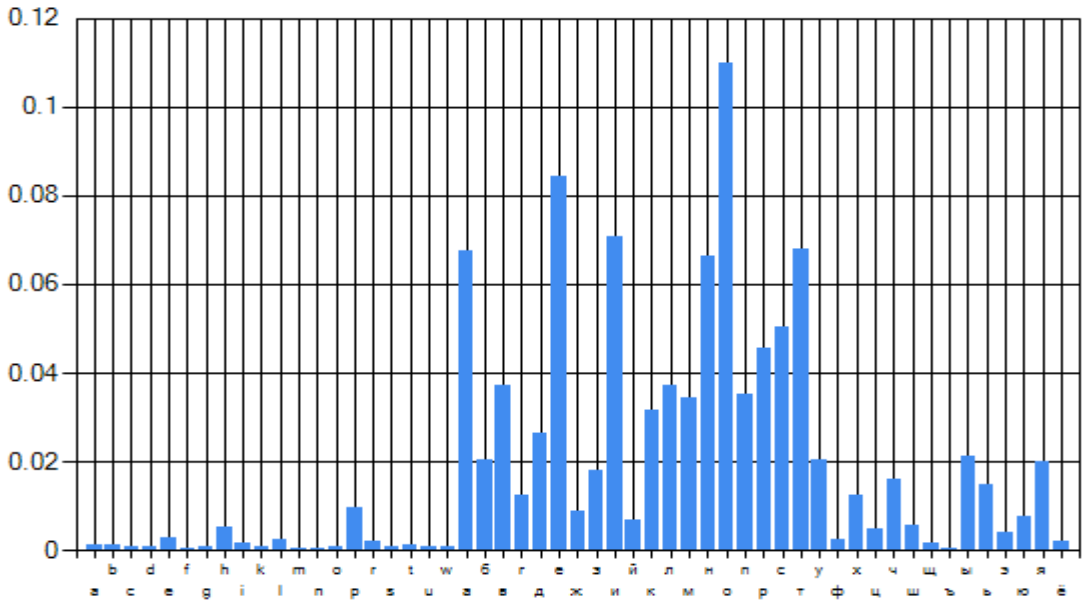
Результат стиснення:



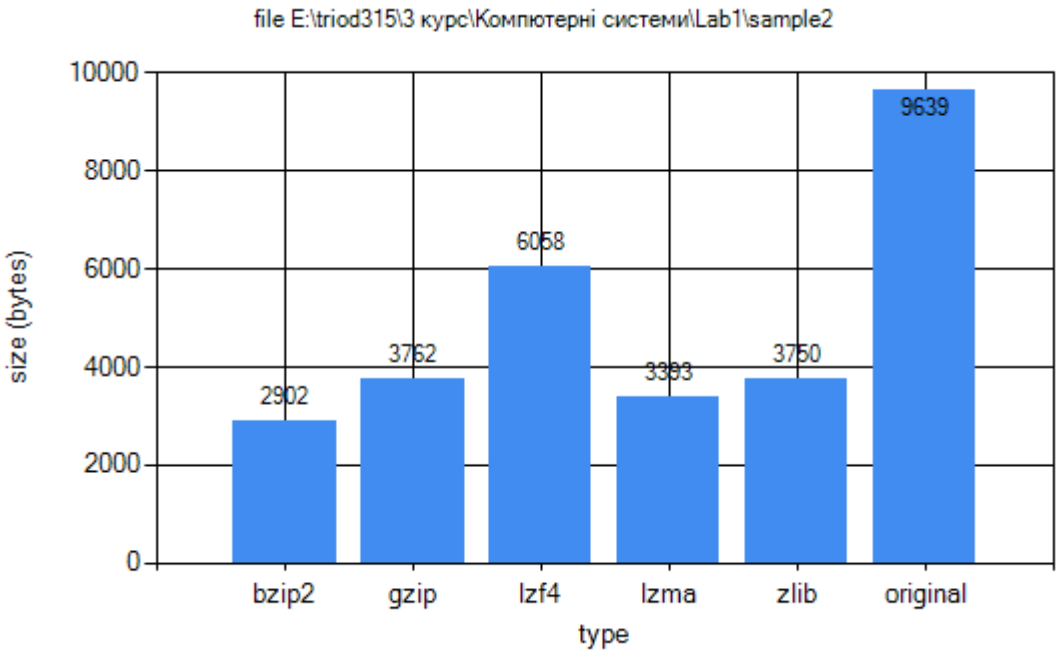
Формат	Bzip2	Gzip	Lzf4	Lzma	Zlib	Txt	Кількість інформації
Обсяг(байт)	309	344	435	341	332	628	147.55

2) Sample2

Кількість літер: 4380
Кількість унікальних літер: 53
Ентропія: 4.653
Кількість інформації: 2547.49 bytes



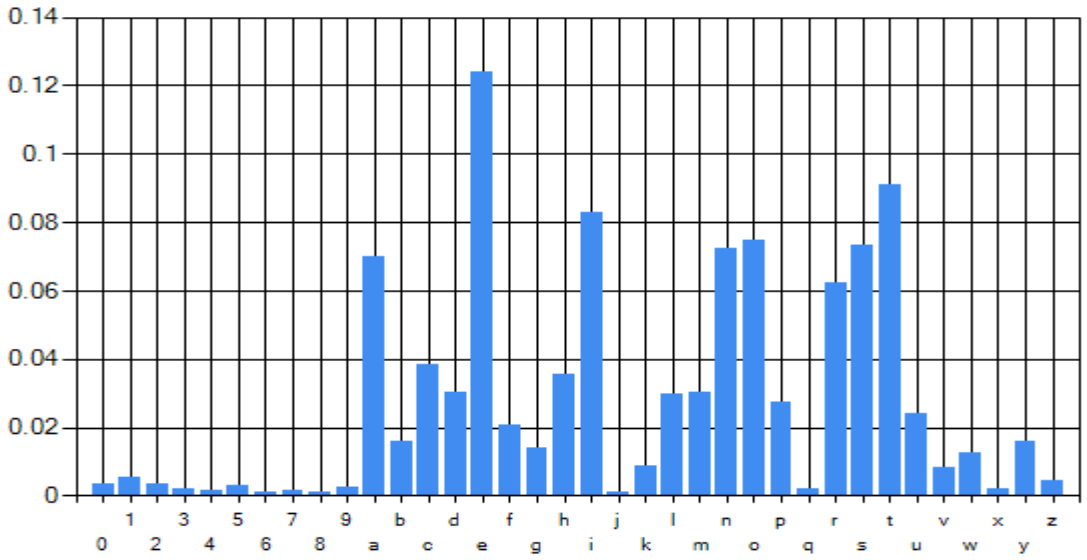
Результат стиснення:



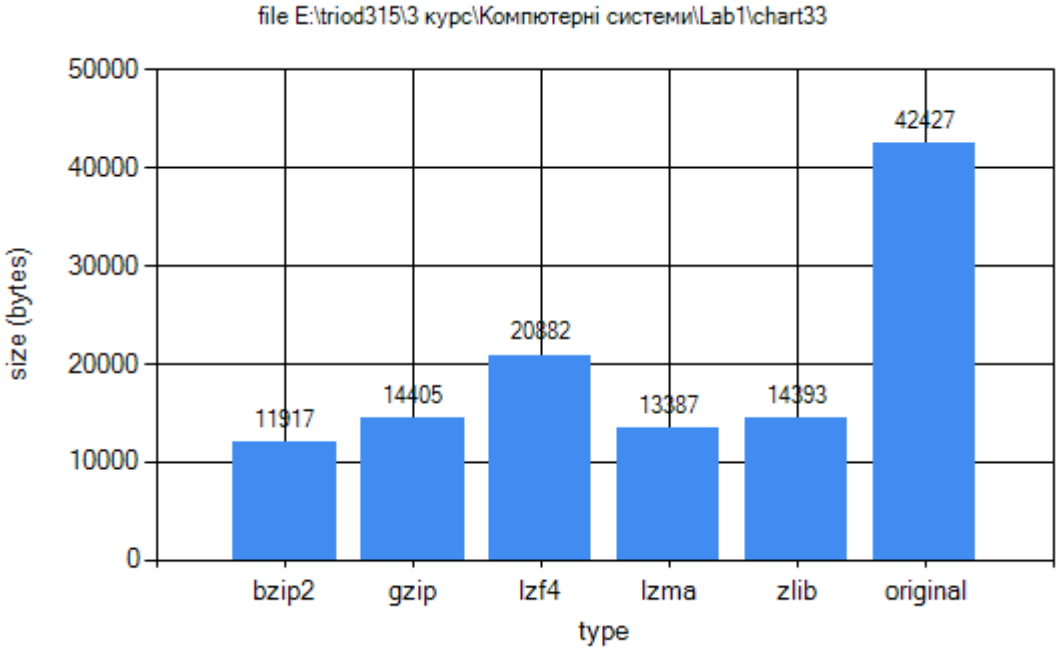
Формат	Bzip2	Gzip	Lzf4	Lzma	Zlib	Txt	Кількість інформації
Обсяг(байт)	2902	3762	6058	3393	3750	9639	2547.49

3) Sample3

Кількість літер: 27010
Кількість унікальних літер: 36
Ентропія: 4.164
Кількість інформації: 13688.416 bytes



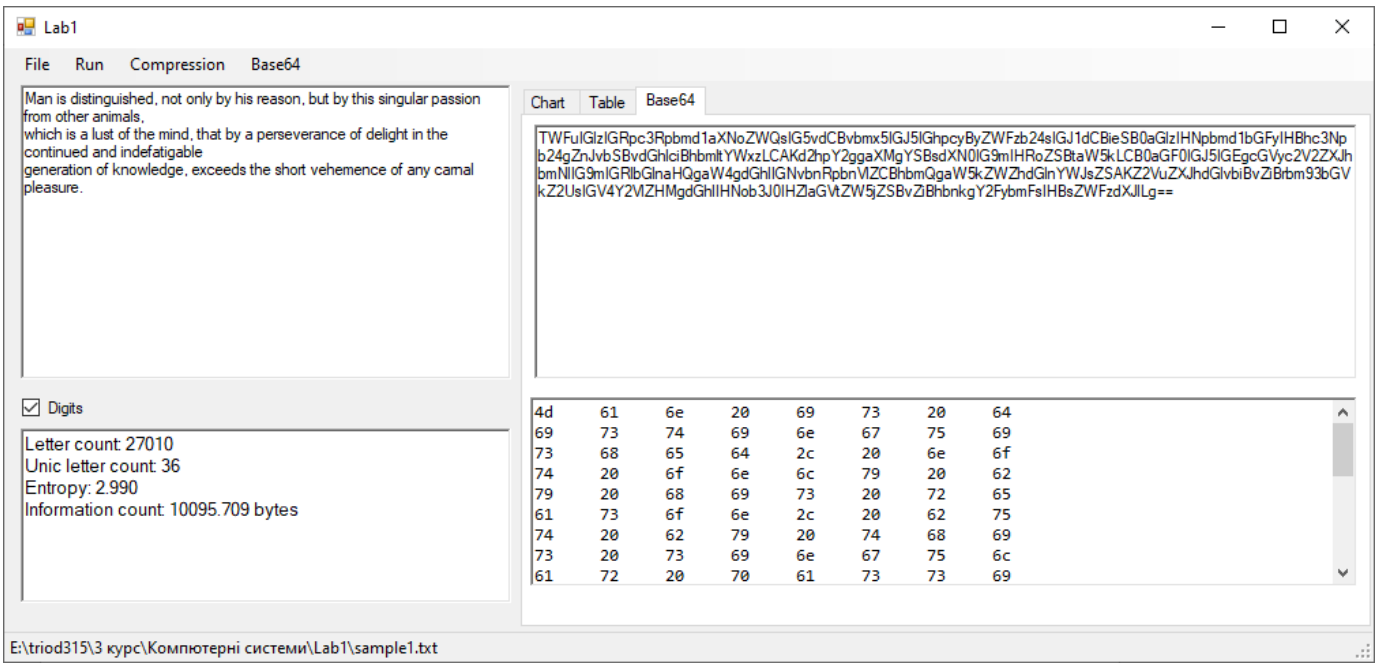
Результат стиснення:



Формат	Bzip2	Gzip	Lzf4	Lzma	Zlib	Txt	Кількість інформації
Обсяг(байт)	11917	14405	20882	13387	14393	42427	13688.4

2) Дослідження способів кодування інформації на прикладі Base64

1. Ознайомтесь зі стандартом RFC4648
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)
    a. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)



3. Закодуйте в Base64 обрані вами текстові файли
    a. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
    b. Порівняйте отримане значення з кількістю інформації вихідного файлу
    c. Зробіть висновки з отриманого результату
4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
    a. 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

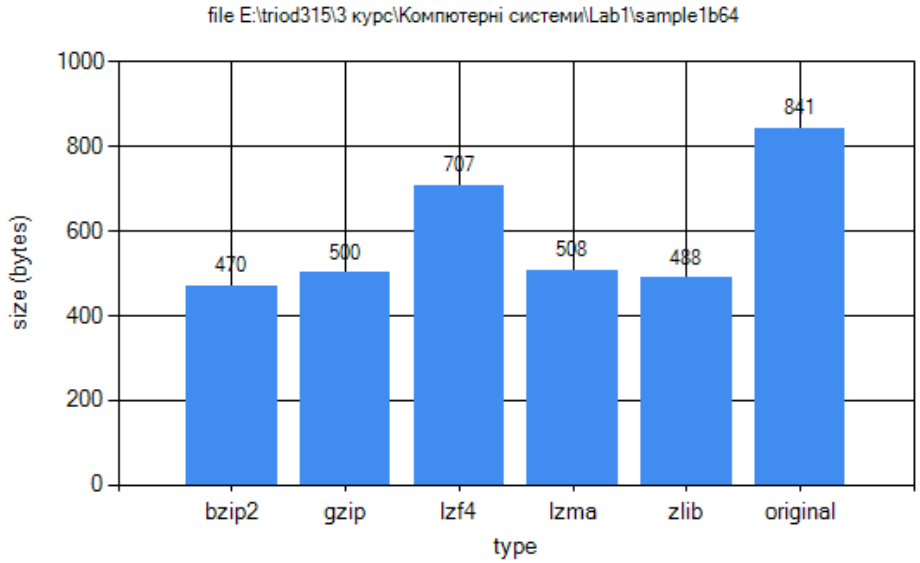
Результати:

1) Sample1

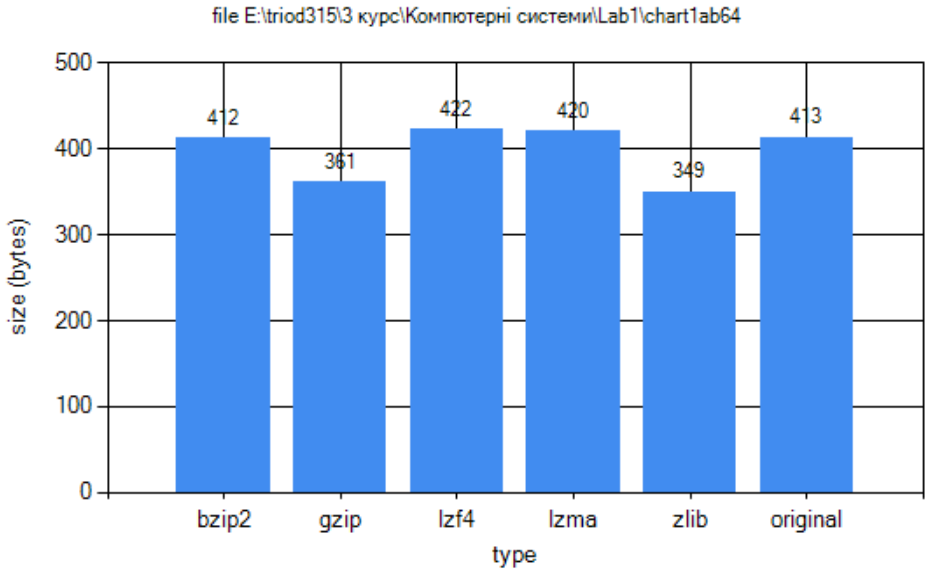
	Normal	Base64
Кількість літер	270	671
Кількість унікальних літер	29	24
Ентропія:	4.372	4.164
Кількість інформації	147.55	349.27

Стиснення файлу

Тип файлу	Кільк. інформації	txt	Base64 txt
Обсяг	147.55	628	841



Стиснутий текст у форматі Base64



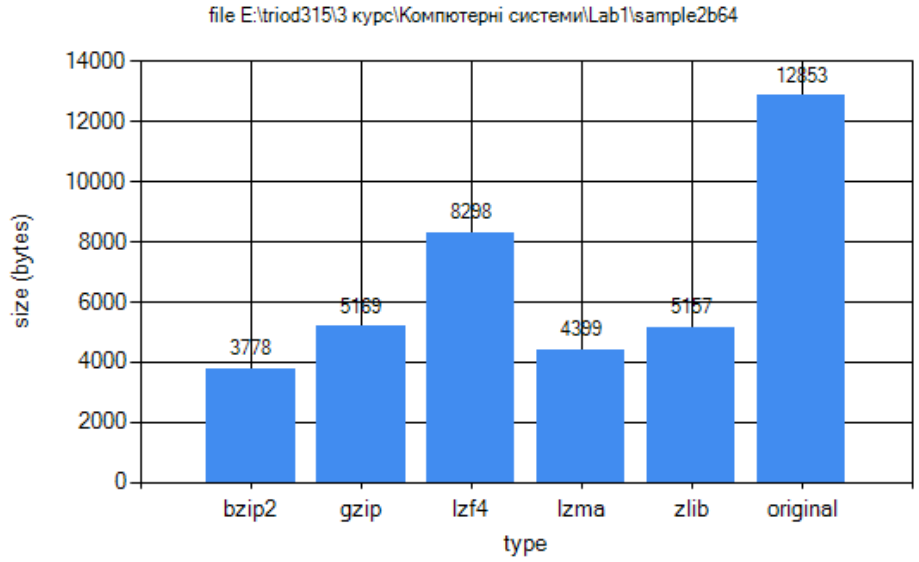
Архів закодований у Base64 і ще раз стиснутий

2) Sample2

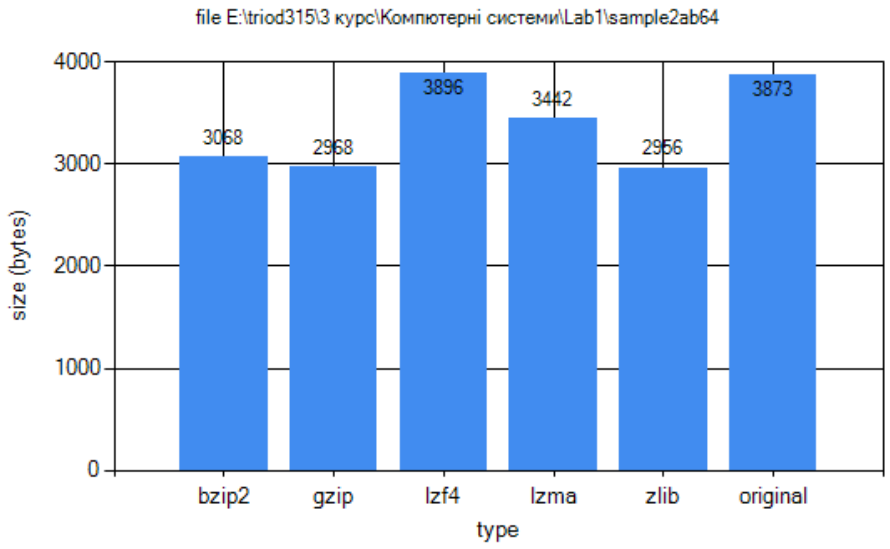
	Normal	Base64
Кількість літер	4380	10061
Кількість унікальних літер	53	26
Ентропія:	4.65	4.104
Кількість інформації	2547.48	5160.83

Стиснення файлу

Тип файлу	Кільк. інформації	txt	Base64 txt
Обсяг	1765.78	9639	12853



Стиснутий текст у форматі Base64



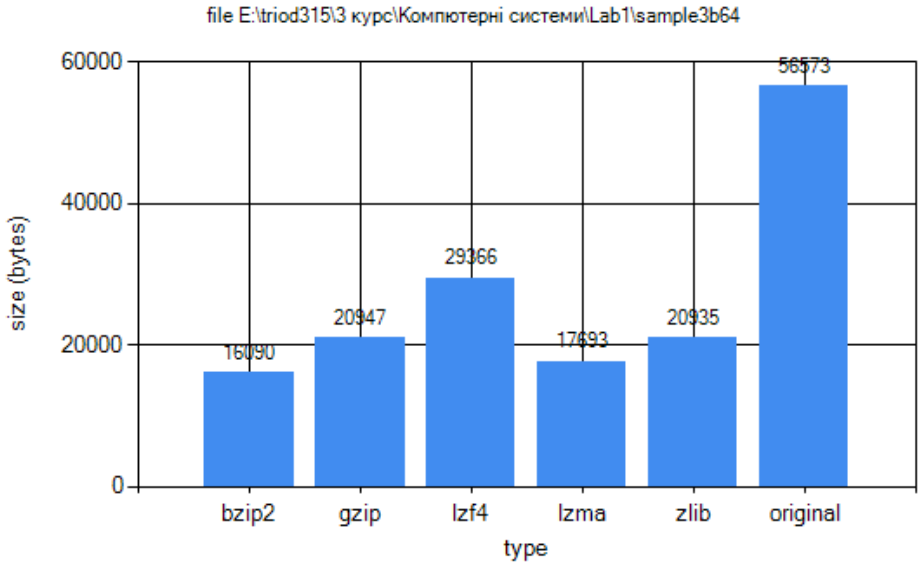
Архів закодований у Base64 і ще раз стиснутий

3) Sample3

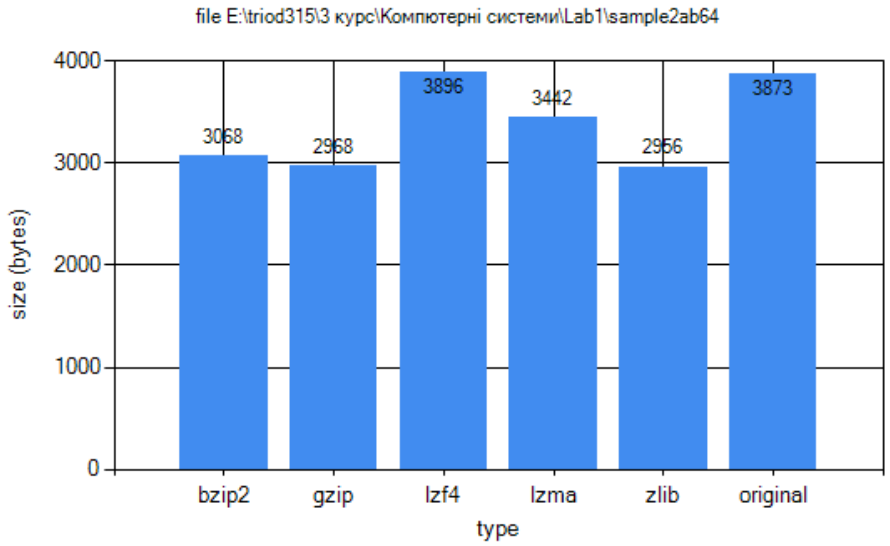
	Normal	Base64
Кількість літер	26296	50979
Кількість унікальних літер	26	26
Ентропія:	4.164	4.474
Кількість інформації	13688.416	28502.4

Стиснення файлу

Тип файлу	Кільк. інформації	txt	Base64 txt
Обсяг	13688.416	42427	19756



Стиснутий текст у форматі Base64



Архів закодований у Base64 і ще раз стиснутий

Висновки: в даній лабораторній було розроблено спеціальне ПЗ для аналізу тексту, яке обчислює кількість інформації та ентропію тексу, проводить стиснення тексту та перетворює у формат Base64. Було проаналізовано три зразки тексту: Sample1.txt – вірш Т. Г. Шевченка «Думи мої думи», Sample2.txt – фрагмент статті про PHP з lurkore.to, Sample3.txt – RFC 2795 (IMPS). Було встановлено що найкращим з перевірених алгоритмом стиснення є BZip2. Також можна помітити що обсяг файлів BZip2 менший за кількість інформації, це пов’язано з неточністю формули для оцінки ентропії для природніх мов, де ймовірність наступних символів може залежати від попередніх.