Київський національний університет імені Тараса Шевченка
Факультет радіофізики, електроніки та комп’ютерних систем
Навчальна дисципліна «Комп’ютерні системи»

Звіт з лабораторної роботи №1

на тему «**Дослідження кількості інформації при різних варіантах кодування**»

Роботу виконав
Студент 3 курсу
КІ, група СА
Кравченко В’ячеслав
Васильович

Київ 2019

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

Хід роботи

Дослідження кількості інформації в тексті

1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (файли також є у репозиторії):
   - Об’єкт SCP-173
   - Гальмування двигуном
   - Перекладений текст пісні Rick Astley - Never Gonna Give You Up (натисніть, щоб піймaтися на рікролл українською)
2. 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!
   - Заради цього створено мною цей документ
3. Код створеної програми міститься у репозиторії, посилання буде вкінці. Результат роботи програми:

```
Файл для аналізу: scp_1.txt
Загальна кількість символів файлу: 1216

Відносна частота появи літери "а" у тексті = 0,0559210526315789 ; Літера присутня у тексті: 68 разів.
Відносна частота появи літери "б" у тексті = 0,0271381578947368 ; Літера присутня у тексті: 33 разів.
Відносна частота появи літери "в" у тексті = 0,0493421052631579 ; Літера присутня у тексті: 60 разів.
Відносна частота появи літери "г" у тексті = 0,00986842105263158 ; Літера присутня у тексті: 12 разів.
Відносна частота появи літери "ґ" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "д" у тексті = 0,0279605263157895 ; Літера присутня у тексті: 34 разів.
Відносна частота появи літери "е" у тексті = 0,0625 ; Літера присутня у тексті: 76 разів.
Відносна частота появи літери "є" у тексті = 0,0148026315789474 ; Літера присутня у тексті: 18 разів.
Відносна частота появи літери "ж" у тексті = 0,00657894736842105 ; Літера присутня у тексті: 8 разів.
Відносна частота появи літери "з" у тексті = 0,0222039473684211 ; Літера присутня у тексті: 27 разів.
Відносна частота появи літери "и" у тексті = 0,0592105263157895 ; Літера присутня у тексті: 72 разів.
Відносна частота появи літери "і" у тексті = 0,0493421052631579 ; Літера присутня у тексті: 60 разів.
Відносна частота появи літери "ї" у тексті = 0,00246710526315789 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "й" у тексті = 0,0230263157894737 ; Літера присутня у тексті: 28 разів.
Відносна частота появи літери "к" у тексті = 0,0361842105263158 ; Літера присутня у тексті: 44 разів.
Відносна частота появи літери "л" у тексті = 0,0205592105263158 ; Літера присутня у тексті: 25 разів.
Відносна частота появи літери "м" у тексті = 0,0287828947368421 ; Літера присутня у тексті: 35 разів.
Відносна частота появи літери "н" у тексті = 0,0847039473684211 ; Літера присутня у тексті: 103 разів.
Відносна частота появи літери "о" у тексті = 0,100328947368421 ; Літера присутня у тексті: 122 разів.
Відносна частота появи літери "п" у тексті = 0,0328947368421053 ; Літера присутня у тексті: 40 разів.
Відносна частота появи літери "р" у тексті = 0,0534539473684211 ; Літера присутня у тексті: 65 разів.
Відносна частота появи літери "с" у тексті = 0,0304276315789474 ; Літера присутня у тексті: 37 разів.
Відносна частота появи літери "т" у тексті = 0,0567434210526316 ; Літера присутня у тексті: 69 разів.
Відносна частота появи літери "у" у тексті = 0,0328947368421053 ; Літера присутня у тексті: 40 разів.
Відносна частота появи літери "ф" у тексті = 0,00164473684210526 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "х" у тексті = 0,00904605263157895 ; Літера присутня у тексті: 11 разів.
Відносна частота появи літери "ц" у тексті = 0,00328947368421053 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "ч" у тексті = 0,00986842105263158 ; Літера присутня у тексті: 12 разів.
Відносна частота появи літери "ш" у тексті = 0,00411184210526316 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "щ" у тексті = 0,000822368421052632 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "ь" у тексті = 0,00904605263157895 ; Літера присутня у тексті: 11 разів.
Відносна частота появи літери "ю" у тексті = 0,00411184210526316 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "я" у тексті = 0,0205592105263158 ; Літера присутня у тексті: 25 разів.

Середня ентропія нерівноймовірного алфавіту у заданому тексті: 4,33820760849823
Кількість інформації у тексті: 659,407556491731
```

```
Файл для аналізу: brake_2.txt
Загальна кількість символів файлу: 7338

Відносна частота появи літери "а" у тексті = 0,0806759334968656 ; Літера присутня у тексті: 592 разів.
Відносна частота появи літери "б" у тексті = 0,0160806759334969 ; Літера присутня у тексті: 118 разів.
Відносна частота появи літери "в" у тексті = 0,0509675660943036 ; Літера присутня у тексті: 374 разів.
Відносна частота появи літери "г" у тексті = 0,0177159989097847 ; Літера присутня у тексті: 130 разів.
Відносна частота появи літери "ґ" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "д" у тексті = 0,0363859362224039 ; Літера присутня у тексті: 267 разів.
Відносна частота появи літери "е" у тексті = 0,0496047969473971 ; Літера присутня у тексті: 364 разів.
Відносна частота появи літери "є" у тексті = 0,00790406105205778 ; Літера присутня у тексті: 58 разів.
Відносна частота появи літери "ж" у тексті = 0,00926683019896429 ; Літера присутня у тексті: 68 разів.
Відносна частота появи літери "з" у тексті = 0,0235759062414827 ; Літера присутня у тексті: 173 разів.
Відносна частота появи літери "и" у тексті = 0,0630962115017716 ; Літера присутня у тексті: 463 разів.
Відносна частота появи літери "і" у тексті = 0,0491959662033252 ; Літера присутня у тексті: 361 разів.
Відносна частота появи літери "ї" у тексті = 0,00517852275824475 ; Літера присутня у тексті: 38 разів.
Відносна частота появи літери "й" у тексті = 0,00817661488143908 ; Літера присутня у тексті: 60 разів.
Відносна частота появи літери "к" у тексті = 0,0332515671845189 ; Літера присутня у тексті: 244 разів.
Відносна частота появи літери "л" у тексті = 0,0331152902698283 ; Літера присутня у тексті: 243 разів.
Відносна частота появи літери "м" у тексті = 0,0342055055873535 ; Літера присутня у тексті: 251 разів.
Відносна частота появи літери "н" у тексті = 0,0756336876533115 ; Літера присутня у тексті: 555 разів.
Відносна частота появи літери "о" у тексті = 0,0902153175252112 ; Літера присутня у тексті: 662 разів.
Відносна частота появи літери "п" у тексті = 0,0287544289997274 ; Літера присутня у тексті: 211 разів.
Відносна частота появи літери "р" у тексті = 0,0363859362224039 ; Літера присутня у тексті: 267 разів.
Відносна частота появи літери "с" у тексті = 0,0362496593077133 ; Літера присутня у тексті: 266 разів.
Відносна частота появи літери "т" у тексті = 0,053147996729354 ; Літера присутня у тексті: 390 разів.
Відносна частота появи літери "у" у тексті = 0,0317525211229218 ; Літера присутня у тексті: 233 разів.
Відносна частота появи літери "ф" у тексті = 0,00095393840283456 ; Літера присутня у тексті: 7 разів.
Відносна частота появи літери "х" у тексті = 0,00804033796674843 ; Літера присутня у тексті: 59 разів.
Відносна частота появи літери "ц" у тексті = 0,0050422458435541 ; Літера присутня у тексті: 37 разів.
Відносна частота появи літери "ч" у тексті = 0,0155355682747343 ; Літера присутня у тексті: 114 разів.
Відносна частота появи літери "ш" у тексті = 0,00558735350231671 ; Літера присутня у тексті: 41 разів.
Відносна частота появи літери "щ" у тексті = 0,00558735350231671 ; Літера присутня у тексті: 41 разів.
Відносна частота появи літери "ь" у тексті = 0,0215317525211229 ; Літера присутня у тексті: 158 разів.
Відносна частота появи літери "ю" у тексті = 0,00885799945489234 ; Літера присутня у тексті: 65 разів.
Відносна частота появи літери "я" у тексті = 0,0249386753883892 ; Літера присутня у тексті: 183 разів.

Середня ентропія нерівноймовірного алфавіту у заданому тексті: 4,46480594664132
Кількість інформації у тексті: 4095,34325455675
```

```
Файл для аналізу: rick_3.txt
Загальна кількість символів файлу: 832

Відносна частота появи літери "а" у тексті = 0,0697115384615385 ; Літера присутня у тексті: 58 разів.
Відносна частота появи літери "б" у тексті = 0,0276442307692308 ; Літера присутня у тексті: 23 разів.
Відносна частота появи літери "в" у тексті = 0,0396634615384615 ; Літера присутня у тексті: 33 разів.
Відносна частота появи літери "г" у тексті = 0,00961538461538462 ; Літера присутня у тексті: 8 разів.
Відносна частота появи літери "ґ" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "д" у тексті = 0,0204326923076923 ; Літера присутня у тексті: 17 разів.
Відносна частота появи літери "е" у тексті = 0,0997596153846154 ; Літера присутня у тексті: 83 разів.
Відносна частота появи літери "є" у тексті = 0,00600961538461538 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "ж" у тексті = 0,0252403846153846 ; Літера присутня у тексті: 21 разів.
Відносна частота появи літери "з" у тексті = 0,0240384615384615 ; Літера присутня у тексті: 20 разів.
Відносна частота появи літери "и" у тексті = 0,0733173076923077 ; Літера присутня у тексті: 61 разів.
Відносна частота появи літери "і" у тексті = 0,03125 ; Літера присутня у тексті: 26 разів.
Відносна частота появи літери "ї" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "й" у тексті = 0,0108173076923077 ; Літера присутня у тексті: 9 разів.
Відносна частота появи літери "к" у тексті = 0,0444711538461538 ; Літера присутня у тексті: 37 разів.
Відносна частота появи літери "л" у тексті = 0,0276442307692308 ; Літера присутня у тексті: 23 разів.
Відносна частота появи літери "м" у тексті = 0,0240384615384615 ; Літера присутня у тексті: 20 разів.
Відносна частота появи літери "н" у тексті = 0,0625 ; Літера присутня у тексті: 52 разів.
Відносна частота появи літери "о" у тексті = 0,0853365384615385 ; Літера присутня у тексті: 71 разів.
Відносна частота появи літери "п" у тексті = 0,0216346153846154 ; Літера присутня у тексті: 18 разів.
Відносна частота появи літери "р" у тексті = 0,0336538461538462 ; Літера присутня у тексті: 28 разів.
Відносна частота появи літери "с" у тексті = 0,0120192307692308 ; Літера присутня у тексті: 10 разів.
Відносна частота появи літери "т" у тексті = 0,0600961538461538 ; Літера присутня у тексті: 50 разів.
Відносна частота появи літери "у" у тексті = 0,0528846153846154 ; Літера присутня у тексті: 44 разів.
Відносна частота появи літери "ф" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "х" у тексті = 0,00480769230769231 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "ц" у тексті = 0,00841346153846154 ; Літера присутня у тексті: 7 разів.
Відносна частота появи літери "ч" у тексті = 0,0132211538461538 ; Літера присутня у тексті: 11 разів.
Відносна частота появи літери "ш" у тексті = 0,015625 ; Літера присутня у тексті: 13 разів.
Відносна частота появи літери "щ" у тексті = 0,0120192307692308 ; Літера присутня у тексті: 10 разів.
Відносна частота появи літери "ь" у тексті = 0,0108173076923077 ; Літера присутня у тексті: 9 разів.
Відносна частота появи літери "ю" у тексті = 0,00841346153846154 ; Літера присутня у тексті: 7 разів.
Відносна частота появи літери "я" у тексті = 0,0132211538461538 ; Літера присутня у тексті: 11 разів.

Середня ентропія нерівноймовірного алфавіту у заданому тексті: 4,33067035550116
Кількість інформації у тексті: 450,389716972121
```

4. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення.

Для цього завдання я використовував WinRAR, WinZip та 7-Zip. Для кожного алгоритму я використовував рівень стиснення Normal, щоб усі були в рівних умовах.

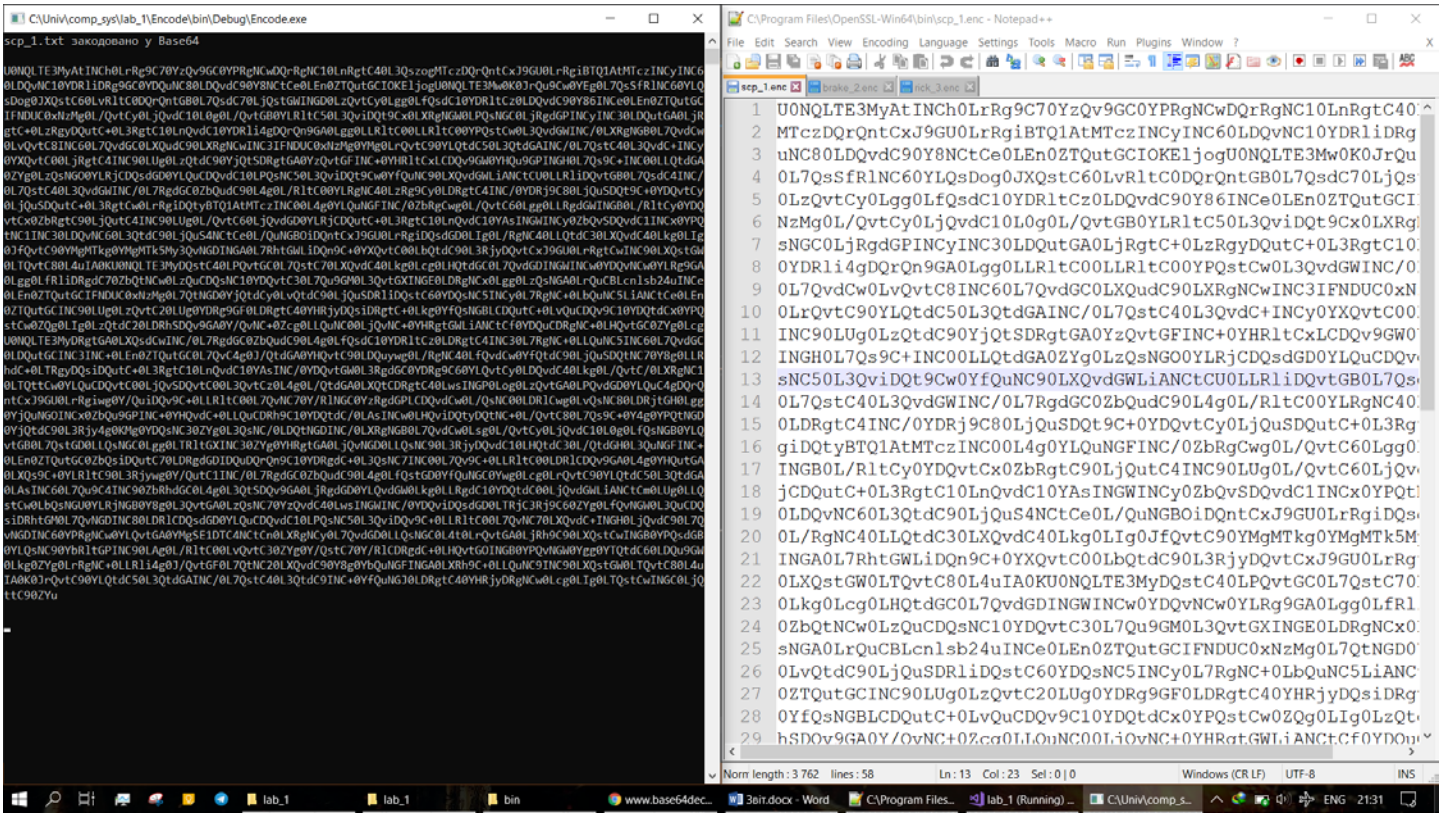
Файл	scp_1.txt, Байт	brake_2.txt, Байт	rick_3.txt, Байт
rar	1181	4083	544
zip	1216	4680	600
gzip	1092	4554	475
xz	1112	4436	516
bz2	932	3685	466
Оригінал	2736	16174	1977
Кількість інформації	659	4095	450

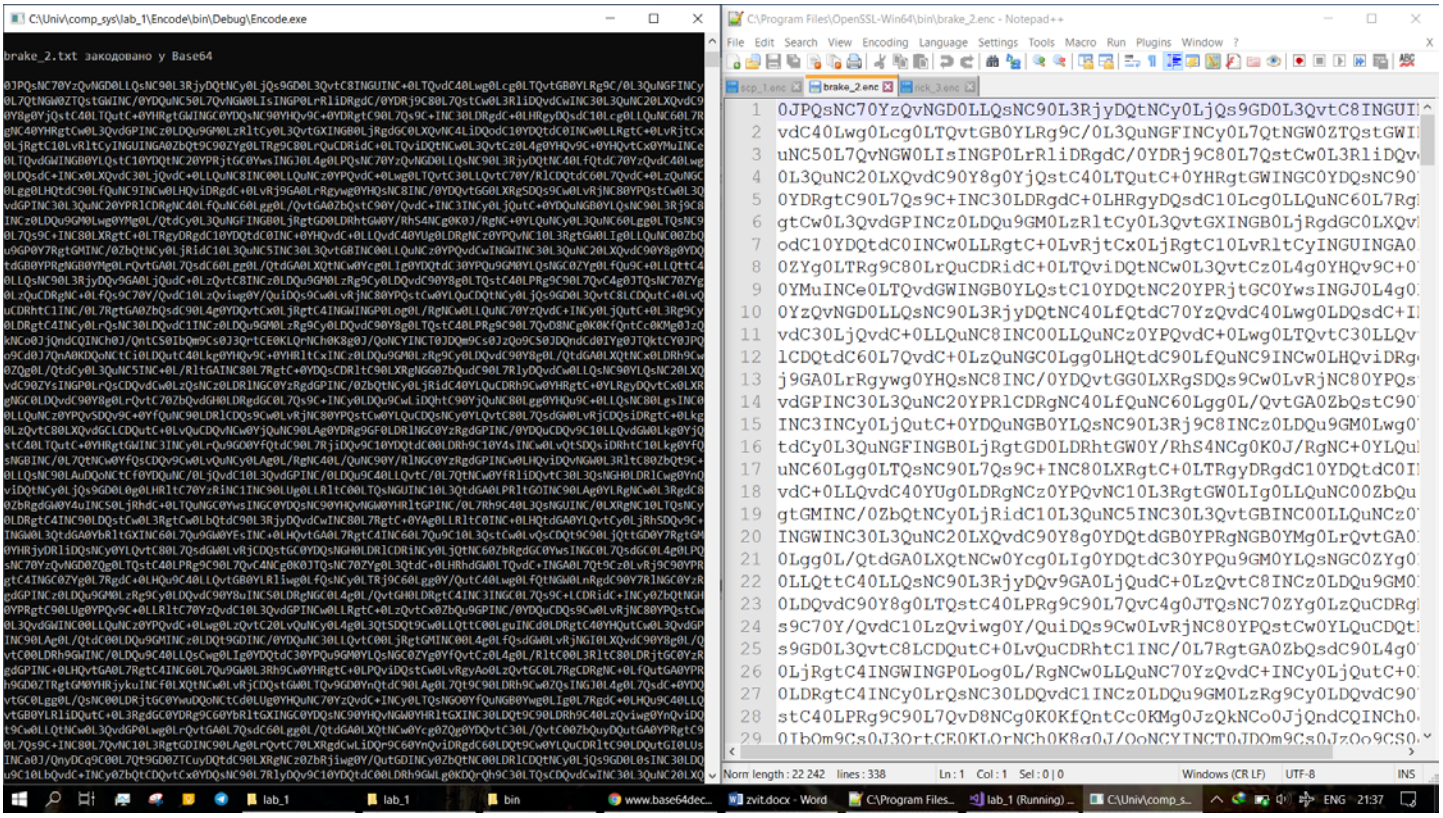
5. У результаті ідеального стиснення розмір файлу повинен бути рівним кількості інформації. Але у реальності розміри архівованих файлів у більшості випадків дещо більші за кількість інформації. Проте, у випадку з великим текстом, помітно сильніше стиснення тексту.
   Це відбувається тому, що алгоритми архіваторів побудовані таким чином аби використати повторювані частини тексту. Виходячи з цього, формула розрахунку кількості інформації, використана для програми, не є досконалою, бо вона не враховує передбачення наступного шматочку текста.
   - Хочу звернути уваги на алгоритм bzip2, який виявився найефективнішим у всіх випадках. А також навіть упорався «ідеальним стисненням» у випадку великого файлу, тобто стиснений архів має розмір навіть менший, ніж кі-сть інформації (у випадку великого файлу справився з цим і RAR).

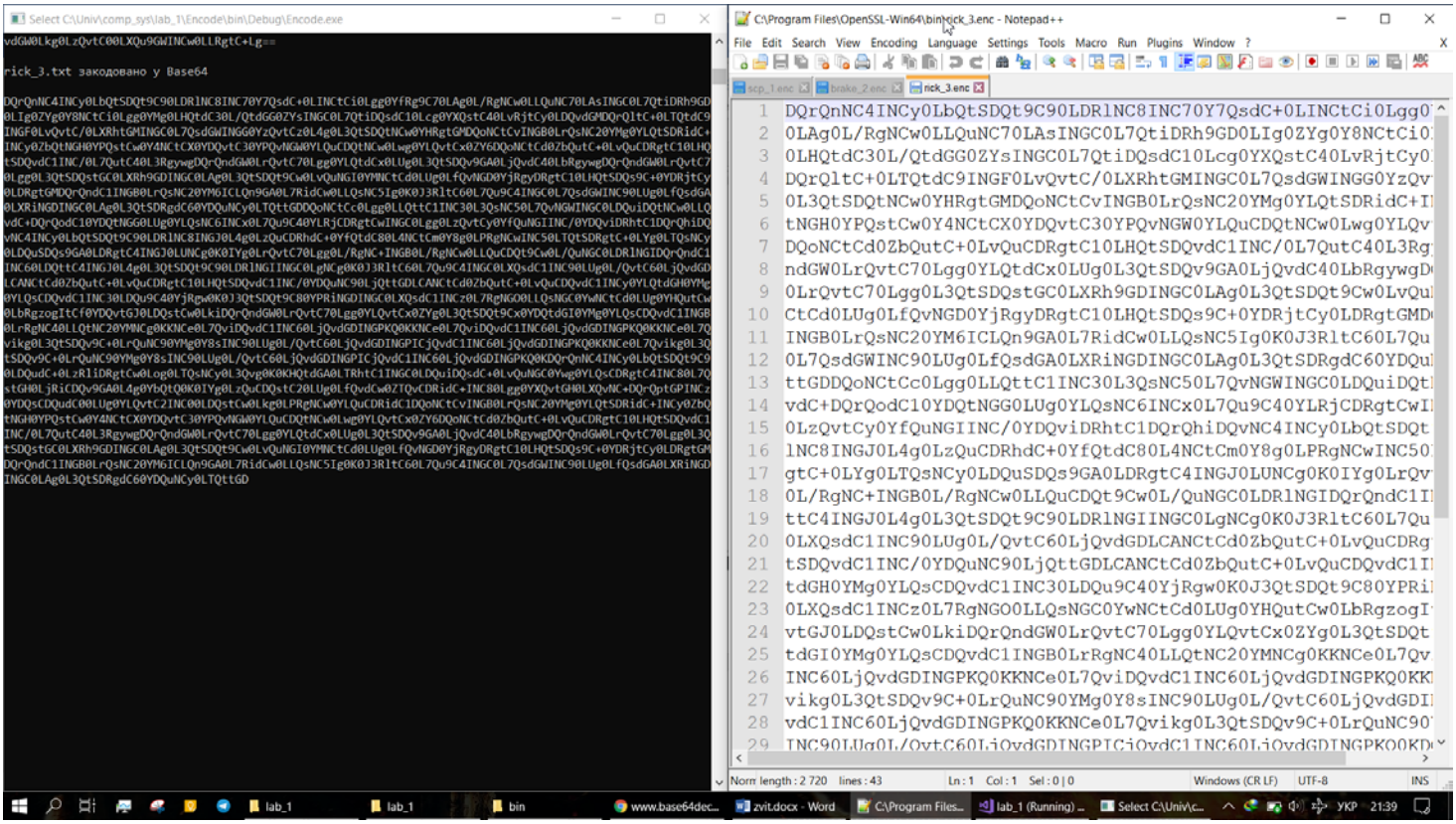
Дослідження способів кодування інформації на прикладі Base64

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)

Зліва – закодовано моєю програмою. Справа – використовуючи OpenSSL для Windows

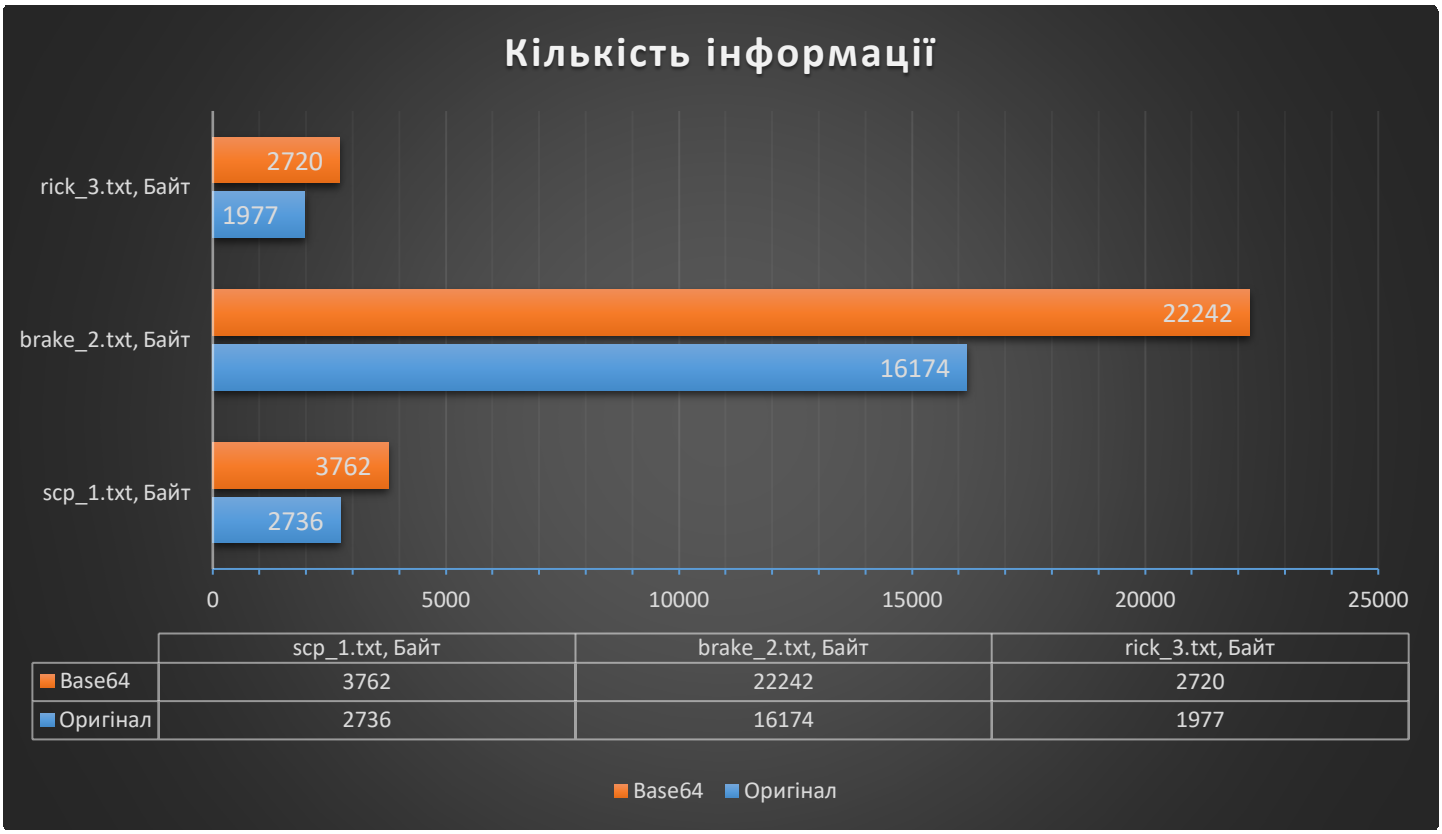






Коректність мого алгоритму також перевірена онлайн кодерами/декодерами.

3. Порівняння текстових файлів та їх версій, закодованих у base64:

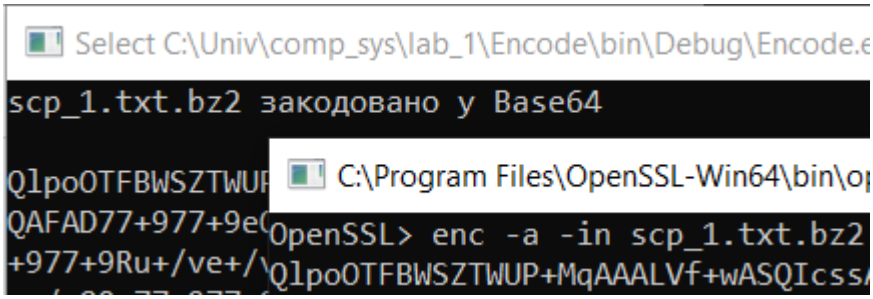


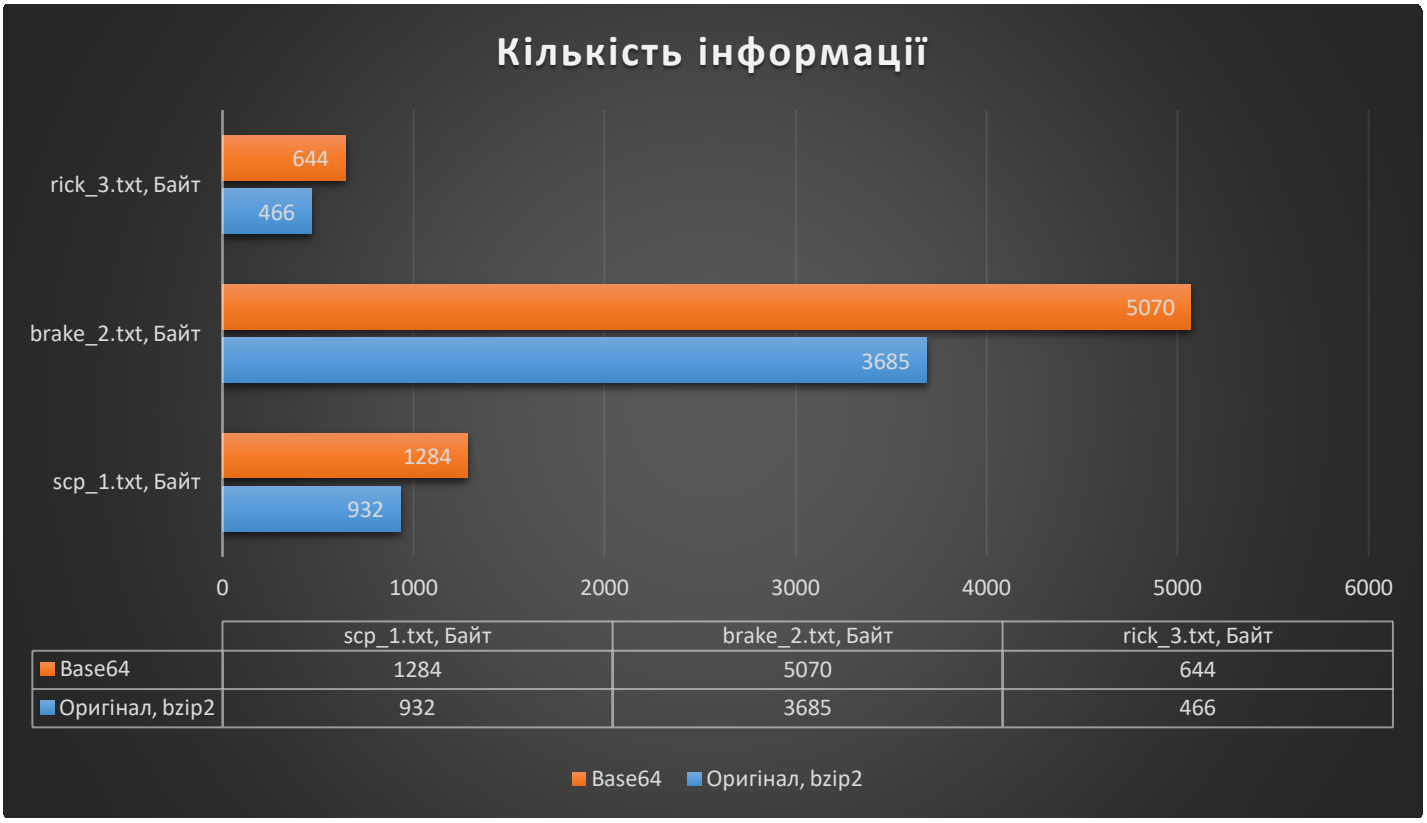
	scp_1.txt, Байт	brake_2.txt, Байт	rick_3.txt, Байт
Base64	3762	22242	2720
Оригінал	2736	16174	1977

Можна легко помітити, що кількість інформації у закодованих файлах зросла. Це пов'язано з алгоритмом кодування base64 – перетворення, наприклад 3 октетів (по 8 біт) у 4 секстети (по 6 біт), що збільшує розмір на третину.

4. Найефективнішим алгоритмом стиснення виявився bzip2, тому тепер порівняння його з кодування у base64:

Перевірка коректності:





Отже, ситуація аналогічна з попереднім пунктом, тобто розмір зріс на 33% (4/3 або ж третину).

Висновок

У ході виконання лабораторної роботи ознайомився з поняттям ентропії інформації та пов’язаних понять. Теоретичні знання закріпив практично. Також порівняв алгоритми стиснення – обрав кращий з них для випадків, коли треба буде зекономити місце на носії. Теоретично та практично ознайомився з алгоритмом кодування Base64, його перевагами та недоліками.

Код програм, звіт та текстові файли, використані у роботі містяться у репозиторії за [цим посиланням (натисніть мене).]