

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

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 \(натисніть, щоб пійматися на рікролл українською\)](#)
 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:

 C:\Univ\comp_sys\lab_1\lab_1\bin\Debug\lab_1.exe

```
Файл для аналізу: scp_1.txt
Загальна кількість символів файлу: 1523

Відносна частота появи літери "А" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Б" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "В" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Г" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Ґ" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Д" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Е" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Є" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Ж" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "З" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "И" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "І" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Ї" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Й" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "К" у тексті = 0,00131319763624425 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "Л" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "М" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Н" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "О" у тексті = 0,00590938936309915 ; Літера присутня у тексті: 9 разів.
Відносна частота появи літери "П" у тексті = 0,00393959290873276 ; Літера присутня у тексті: 6 разів.
Відносна частота появи літери "Р" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "С" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Т" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "У" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Ф" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Х" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Ц" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Ч" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Ш" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Щ" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Ъ" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "ь" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "ю" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "я" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "а" у тексті = 0,0446487196323047 ; Літера присутня у тексті: 68 разів.
Відносна частота появи літери "б" у тексті = 0,0216677609980302 ; Літера присутня у тексті: 33 разів.
Відносна частота появи літери "в" у тексті = 0,0393959290873276 ; Літера присутня у тексті: 60 разів.
Відносна частота появи літери "г" у тексті = 0,00787918581746553 ; Літера присутня у тексті: 12 разів.
```

Відносна частота появи літери "г" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "д" у тексті = 0,0223243598161523 ; Літера присутня у тексті: 34 разів.
Відносна частота появи літери "е" у тексті = 0,0499015101772817 ; Літера присутня у тексті: 76 разів.
Відносна частота появи літери "є" у тексті = 0,0118187787261983 ; Літера присутня у тексті: 18 разів.
Відносна частота появи літери "ж" у тексті = 0,00525279054497702 ; Літера присутня у тексті: 8 разів.
Відносна частота появи літери "з" у тексті = 0,0177281680892974 ; Літера присутня у тексті: 27 разів.
Відносна частота появи літери "и" у тексті = 0,0472751149047932 ; Літера присутня у тексті: 72 разів.
Відносна частота появи літери "і" у тексті = 0,0393959290873276 ; Літера присутня у тексті: 60 разів.
Відносна частота появи літери "ї" у тексті = 0,00196979645436638 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "й" у тексті = 0,0183847669074196 ; Літера присутня у тексті: 28 разів.
Відносна частота появи літери "к" у тексті = 0,0288903479973736 ; Літера присутня у тексті: 44 разів.
Відносна частота появи літери "л" у тексті = 0,0164149704530532 ; Літера присутня у тексті: 25 разів.
Відносна частота появи літери "м" у тексті = 0,0229809586342745 ; Літера присутня у тексті: 35 разів.
Відносна частота появи літери "н" у тексті = 0,0676296782665791 ; Літера присутня у тексті: 103 разів.
Відносна частота появи літери "о" у тексті = 0,0801050558108995 ; Літера присутня у тексті: 122 разів.
Відносна частота появи літери "п" у тексті = 0,0262639527248851 ; Літера присутня у тексті: 40 разів.
Відносна частота появи літери "р" у тексті = 0,0426789231779383 ; Літера присутня у тексті: 65 разів.
Відносна частота появи літери "с" у тексті = 0,0242941562705187 ; Літера присутня у тексті: 37 разів.
Відносна частота появи літери "т" у тексті = 0,0453053184504268 ; Літера присутня у тексті: 69 разів.
Відносна частота появи літери "у" у тексті = 0,0262639527248851 ; Літера присутня у тексті: 40 разів.
Відносна частота появи літери "ф" у тексті = 0,00131319763624425 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "х" у тексті = 0,0072225869993434 ; Літера присутня у тексті: 11 разів.
Відносна частота появи літери "ц" у тексті = 0,00262639527248851 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "ч" у тексті = 0,00787918581746553 ; Літера присутня у тексті: 12 разів.
Відносна частота появи літери "ш" у тексті = 0,00328299409061064 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "щ" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "ъ" у тексті = 0,0072225869993434 ; Літера присутня у тексті: 11 разів.
Відносна частота появи літери "ю" у тексті = 0,00328299409061064 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "я" у тексті = 0,0164149704530532 ; Літера присутня у тексті: 25 разів.
Відносна частота появи літери ", " у тексті = 0,00919238345370978 ; Літера присутня у тексті: 14 разів.
Відносна частота появи літери " " у тексті = 0,137229152987525 ; Літера присутня у тексті: 209 разів.
Відносна частота появи літери "(" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери ")" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "." у тексті = 0,0111621799080762 ; Літера присутня у тексті: 17 разів.
Відносна частота появи літери ":" у тексті = 0,00328299409061064 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "-" у тексті = 0,00853578463558766 ; Літера присутня у тексті: 13 разів.
Відносна частота появи літери "" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "0" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "1" у тексті = 0,00787918581746553 ; Літера присутня у тексті: 12 разів.
Відносна частота появи літери "2" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "3" у тексті = 0,0072225869993434 ; Літера присутня у тексті: 11 разів.
Відносна частота появи літери "4" у тексті = 0,000656598818122127 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "5" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "6" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "7" у тексті = 0,00656598818122127 ; Літера присутня у тексті: 10 разів.
Відносна частота появи літери "8" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "9" у тексті = 0,00196979645436638 ; Літера присутня у тексті: 3 разів.

Середня ентропія нерівномірного алфавіту у заданому тексті: 4,63863065583817

Кількість інформації у тексті: 883,079311105191


```
Файл для аналізу: brake_2.txt
Загальна кількість символів файлу: 8679
```

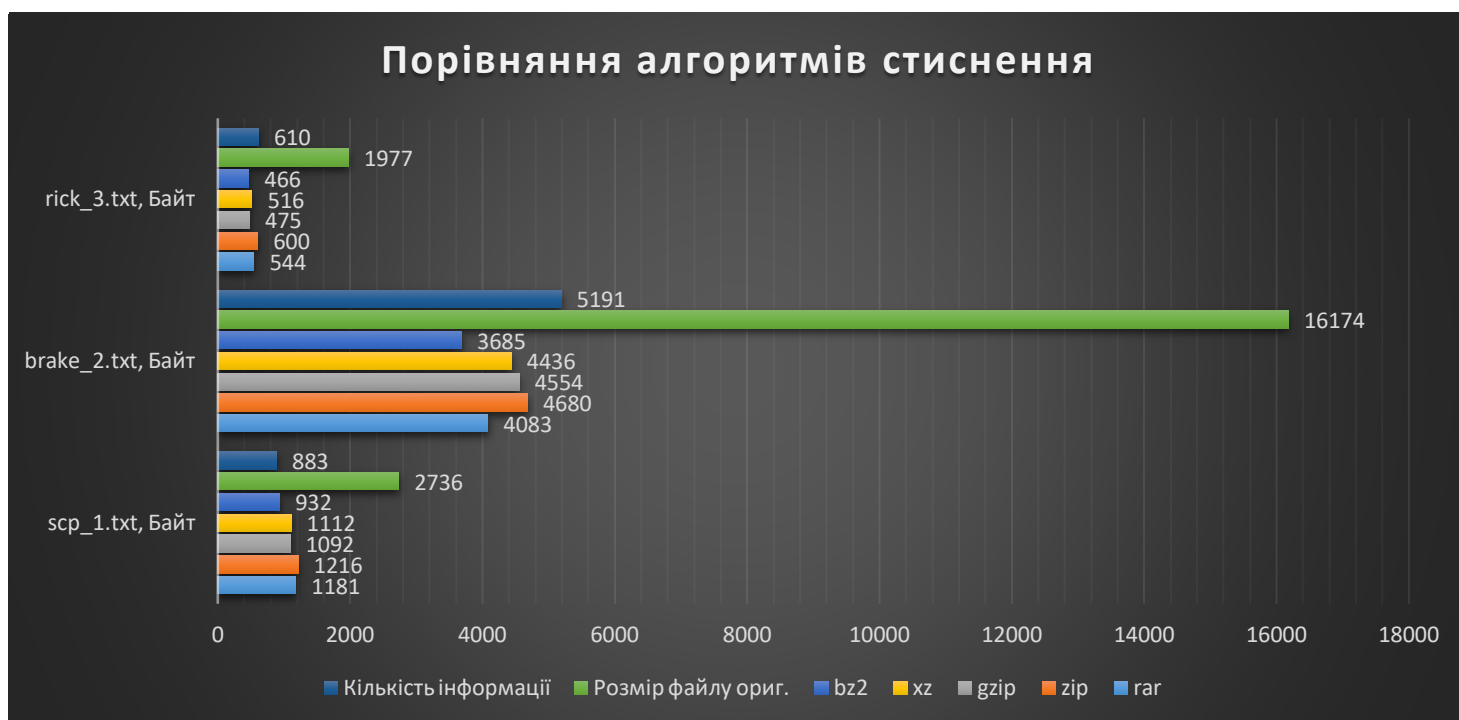
[illegible]

Середня ентропія нерівномірного алфавіту у заданому тексті: 4,78498206501019
Кількість інформації у тексті: 5191,10741777793

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
Розмір файлу ориг.	2736	16174	1977
Кількість інформації	883	5191	610



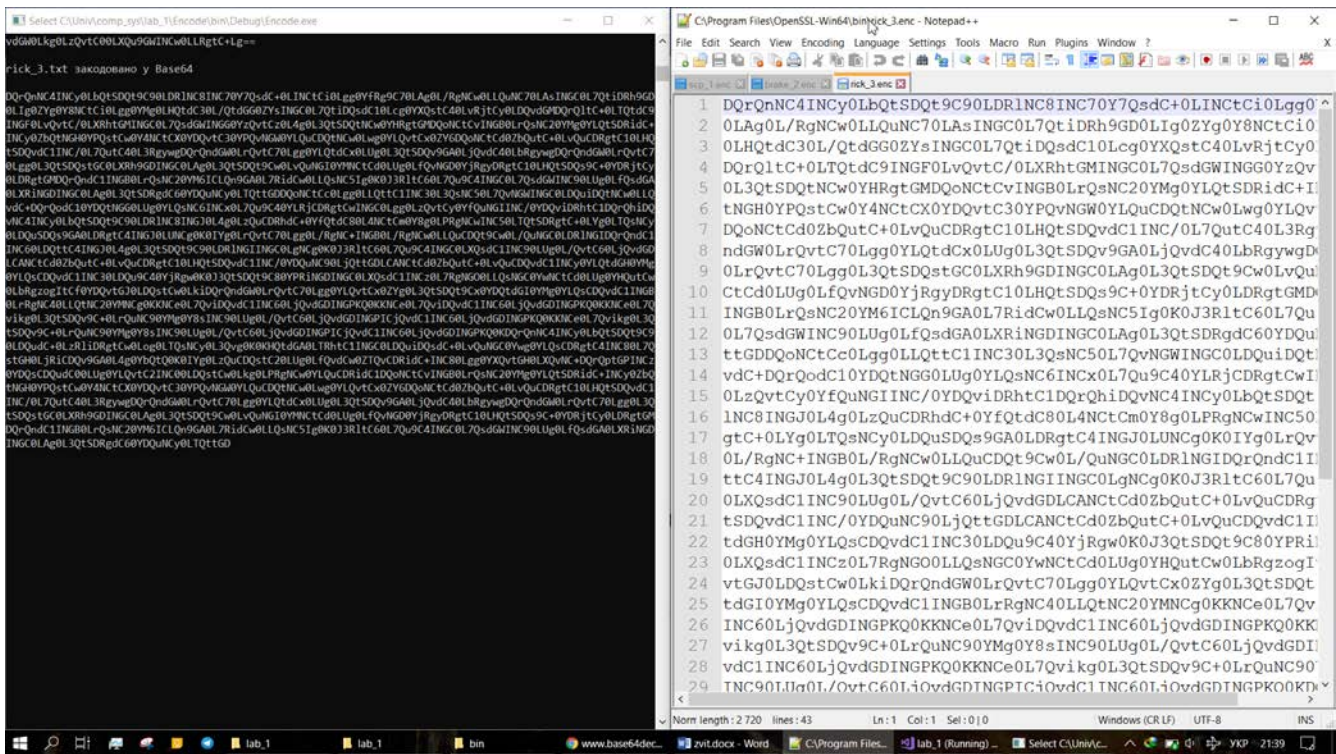
5. У результаті ідеального стиснення розмір файлу повинен бути рівним кількості інформації. Але у реальності розміри архівованих файлів у більшості випадків дещо більші за кількість інформації, окрім випадку з великим текстом. Це відбувається тому, що алгоритми архіваторів побудовані таким чином аби використати повторювані частини тексту. Виходячи з цього, формула розрахунку кількості інформації, використана для програми, не є досконалою, бо вона не враховує передбачення наступного шматочку тексту.
- Хочу звернути уваги на алгоритм bz2, який виявився найефективнішим у всіх випадках. А також навіть упорався «ідеальним стисненням» у випадку великого файлу, тобто стиснений архів має розмір навіть менший, ніж кількість інформації. У випадку з піснею, хоч файл і меншого розміру, ніж текст про SCP-173, але є багато частин повторюваного тексту (як-от приспів)

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)

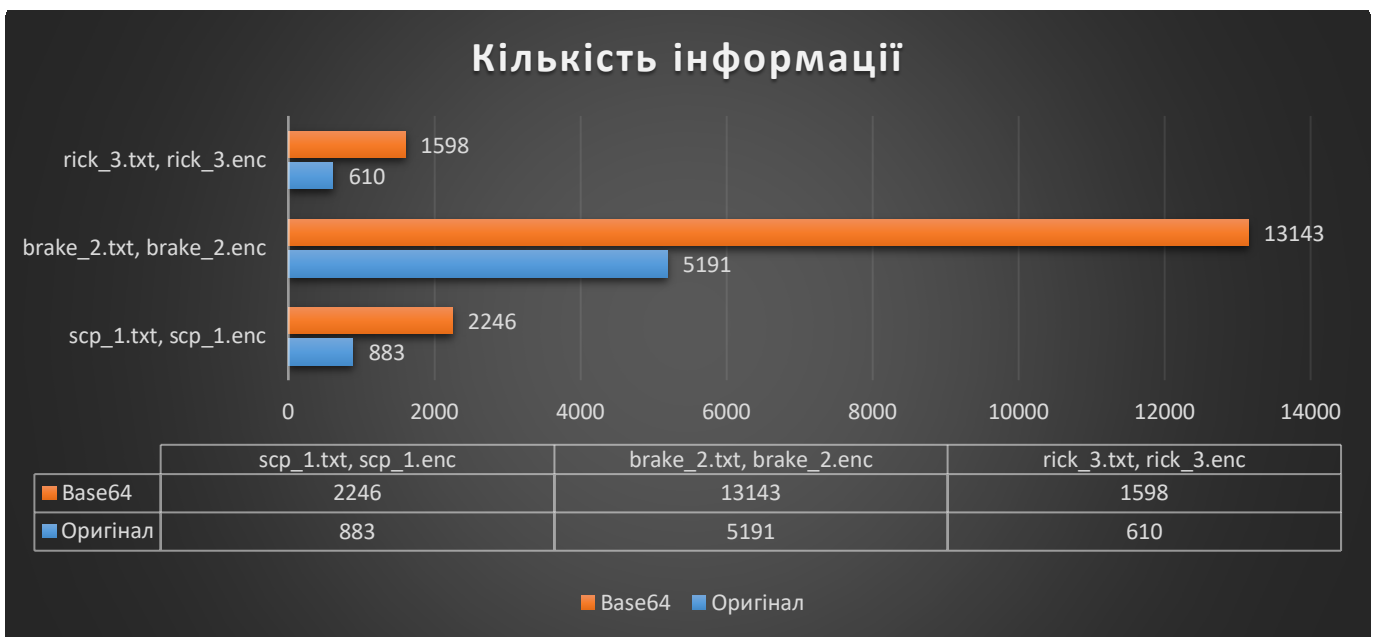
Зліва – закодовано моєю програмою. Справа – використовуючи OpenSSL для Windows

```
U00NQLTE3MyAtInCh0LrRg9C70YzQv9G0C0YPRgNCwDQrRgNC10LnRgtC40...
1 U00NQLTE3MyAtInCh0LrRg9C70YzQv9G0C0YPRgNCwDQrRgNC10LnRgtC40
2 MTczDQrQntCxCJ9GU0LrRgIBTQ1AtMtCzInCYINC60LDQvNC10YDR1iDRg
3 uNC80LDQvdc90Y8NctCe0Len0ZTQtGCIKE1jogU0NQLTE3Mw0K0JrQu
4 0L7QsSfr1NC60YLQsDog0JXQstC60LvR1tC0DQrQntGB0L7QsdC70LjQs
5 0LzQvtCy0Lgg0LQsdC10YDR1tCz0LDQvdc90Y8INc0LEn0ZTQtGCI
6 NzMg0L/QvtCy0LjQvdc10L0g0L/QvtGB0YLRLtC50L3QvDQt9CxC0LXRg
7 sNGC0LjRgdGPINcYINC30LDQtGA0LjRgtC+0LzRgyDQtC+0L3RgtC10
8 0YDR1i4gDvRqN9GA0Lgg0LLR1tC00LLR1tC00YPQstCw0L3QvdcGwINC/0
9 0L7QvdcW0LQvtC8INC60L7QvdcG0LXQudC90LXRGNCwINC3IFNDUC0xN
10 0LrQvtC90YLQtdC50L3QtGdAINC/0L7QstC40L3Qvdc+INcY0YXQvtC00
11 INC90LUG0LzQtdC90YJQtSDRgtGA0YzQvtGFINC+0YHRLtCxCdC9v9GW0
12 INGH0L7Qs9C+INC00LLQtdGA0ZYg0LzQsNG00YLrjCDQsGd0YLQvdcv
13 sNC50L3QvDQt9Cw0YfQuNC90LXQvdcGwLiANctCU0LLR1iDQvtGB0L7Qs
14 0L7QstC40L3QvdcGwINC/0L7RgdG0C0ZbQudC90L4g0L/R1tC00YLrNC40
15 0LDRgtC4INC/0YDRj9C80LjQsUDQt9C+0YDQvtCy0LjQsUDQtC+0L3Rg
16 giDQtYBTQ1AtMtCzINC00L4g0YLQuNGFINC/0zBRgCwG0L/QvtC60Lgg0
17 INGB0L/R1tCy0YDQvtCxC0ZbRgtC90LjQtC4INC90LUG0L/QvtC60LjQv
18 jCDQtC+0L3RgtC10LnQvdc10YAsINGwINCy0ZbQvSDQvdc1INCx0YPQt
19 0LDQvNC60L3QtdC90LjQuS4NctCe0L/QuNGBoiDQtCxCJ9GU0LrRgiDQs
20 0L/RgNC40LLQtdC30LXQvdc40Lkg0Lig0JfQvtC90YmGmTkg0YmGmTK5M
21 INGA0L7RhtGwLiDqn9C+0YXQvtC00LbQtdC90L3RjYDQvtCxCJ9GU0LrRg
22 0LXQstGW0LTQvtC80L4uIA0KU0NQLTE3MYDQstC40LPQvtGC0L7QstC70
23 0Lkg0Lcg0LHQtdGc0L7QvdcGdINGwINCw0YDQvNCw0YLrRg9GA0Lgg0Lfr1
24 0ZbQtnCw0LzQvdcQsNC10YDQvtC30L7Qv6M0L3QvtGXNGE0LDRgNCx0
25 sNGA0LrQuCBCLn1sb24uINc0LEn0ZTQtGCIFNDC0xNzMG0L7QtdNGD0
26 0LQvtdC90LjQvSDR1iDQtC60YDQsNC5INCy0L7RgNC+0LbQuNC5LiANC
27 0ZTQtGcINC90LUG0LzQvtC20LUG0YDRg9G0LDRgtC40YHRjYDQsIdRg
28 0YfQsNGBLCDQtC+0LwQuCdQv9C10YDQtdCxC0YPQstCw0ZQg0Lig0LzQt
29 hSD0v9GA0Y/vNC+0Zc0L10uNC00Li0vNC+0YHRAfGwLiANctCf0YDQm...
```

Коректність мого алгоритму також перевірена онлайн кодерами/декодерами.

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



Можна легко помітити, що розмір закодованих файлах зріс. 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 bzir2, тому тепер порівняння його з кодування у base64:

Перевірка коректності:



Отже, ситуація аналогічна з попереднім пунктом.

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 \(натисніть мене\)](#).