

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 МА
Мормуль Олексій
Володимирович

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 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:

Загальна кількість символів файлу: 338

Загальна кількість символів файлу: 338

[illegible]

Відносна частота появи літери "щ" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "ь" у тексті = 0,0177514792899408 ; Літера присутня у тексті: 6 разів.
Відносна частота появи літери "ю" у тексті = 0,0236686390532544 ; Літера присутня у тексті: 8 разів.
Відносна частота появи літери "я" у тексті = 0,0177514792899408 ; Літера присутня у тексті: 6 разів.
Відносна частота появи літери "і" у тексті = 0,029585798816568 ; Літера присутня у тексті: 10 разів.
Відносна частота появи літери "ї" у тексті = 0,14792899408284 ; Літера присутня у тексті: 50 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 = 0,00887573964497041 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери ":" у тексті = 0,00591715976331361 ; Літера присутня у тексті: 2 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 "0" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "1" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "2" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "3" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "4" у тексті = 0 ; Літера присутня у тексті: 0 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 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "8" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "9" у тексті = 0 ; Літера присутня у тексті: 0 разів.

Середня ентропія нерівномірного алфавіту у заданому тексті: 4,6186527830823
Кількість інформації у тексті: 195,138080085227

Загальна кількість символів файлу: 4694

[illegible]

Відносна частота появи літери "ш" у тексті = 0,00191734128674904 ; Літера присутня у тексті: 9 разів.
Відносна частота появи літери "б" у тексті = 0,0249254367277375 ; Літера присутня у тексті: 117 разів.
Відносна частота появи літери "ю" у тексті = 0,00511291009799744 ; Літера присутня у тексті: 24 разів.
Відносна частота появи літери "я" у тексті = 0,0157648061354921 ; Літера присутня у тексті: 74 разів.
Відносна частота появи літери "і" у тексті = 0,0136344269279932 ; Літера присутня у тексті: 64 разів.
Відносна частота появи літери "ї" у тексті = 0,137835534725181 ; Літера присутня у тексті: 647 разів.
Відносна частота появи літери "(" у тексті = 0,00489987217724755 ; Літера присутня у тексті: 23 разів.
Відносна частота появи літери ")" у тексті = 0,00489987217724755 ; Літера присутня у тексті: 23 разів.
Відносна частота появи літери ":" у тексті = 0,0129953131657435 ; Літера присутня у тексті: 61 разів.
Відносна частота появи літери ";" у тексті = 0,00063911376224968 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "-" у тексті = 0,00106518960374947 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "''" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "0" у тексті = 0,000852151682999574 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "1" у тексті = 0,0034086067319983 ; Літера присутня у тексті: 16 разів.
Відносна частота появи літери "2" у тексті = 0,00127822752449936 ; Літера присутня у тексті: 6 разів.
Відносна частота появи літери "3" у тексті = 0,00063911376224968 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "4" у тексті = 0,00127822752449936 ; Літера присутня у тексті: 6 разів.
Відносна частота появи літери "5" у тексті = 0,00149126544524925 ; Літера присутня у тексті: 7 разів.
Відносна частота появи літери "6" у тексті = 0,00063911376224968 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "7" у тексті = 0,000852151682999574 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "8" у тексті = 0,000852151682999574 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "9" у тексті = 0,00106518960374947 ; Літера присутня у тексті: 5 разів.

Середня ентропія нерівномірного алфавіту у заданому тексті: 4,78265636880096

Кількість інформації у тексті: 2806,22362439397

Загальна кількість символів файлу: 3584

[illegible]

Відносна частота появи літери "ь" у тексті = 0,0139508928571429 ; Літера присутня у тексті: 50 разів.
Відносна частота появи літери "ю" у тексті = 0,00837053571428571 ; Літера присутня у тексті: 30 разів.
Відносна частота появи літери "я" у тексті = 0,0189732142857143 ; Літера присутня у тексті: 68 разів.
Відносна частота появи літери "і" у тексті = 0,00864955357142857 ; Літера присутня у тексті: 31 разів.
Відносна частота появи літери " " у тексті = 0,132254464285714 ; Літера присутня у тексті: 474 разів.
Відносна частота появи літери "(" у тексті = 0,001953125 ; Літера присутня у тексті: 7 разів.
Відносна частота появи літери ")" у тексті = 0,001953125 ; Літера присутня у тексті: 7 разів.
Відносна частота появи літери "." у тексті = 0,0108816964285714 ; Літера присутня у тексті: 39 разів.
Відносна частота появи літери ":" у тексті = 0,000837053571428571 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "-" у тексті = 0,00139508928571429 ; Літера присутня у тексті: 5 разів.
Відносна частота появи літери "" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "0" у тексті = 0,00279017857142857 ; Літера присутня у тексті: 10 разів.
Відносна частота появи літери "1" у тексті = 0,000279017857142857 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "2" у тексті = 0,00111607142857143 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "3" у тексті = 0,00111607142857143 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "4" у тексті = 0,000558035714285714 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "5" у тексті = 0,000279017857142857 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "6" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "7" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "8" у тексті = 0,000279017857142857 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "9" у тексті = 0 ; Літера присутня у тексті: 0 разів.

Середня ентропія нерівномірного алфавіту у заданому тексті: 4,68133589749239
Кількість інформації у тексті: 2097,23848207659

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

Для цього завдання я використовував WinRAR та 7-Zip. Для кожного алгоритму я використовував рівень стиснення Normal.

Файл	Text1.txt Байт	Text2.txt Байт	Text3.txt Байт
Rar	419	3040	2157
Zip	477	3102	2216
Gzip	353	2911	2047
Xz	392	2872	2036
Bz2	306	2454	1754
Кількість інформації	309	3930	3002
Розмір файлу ор:	650	8616	6569



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.

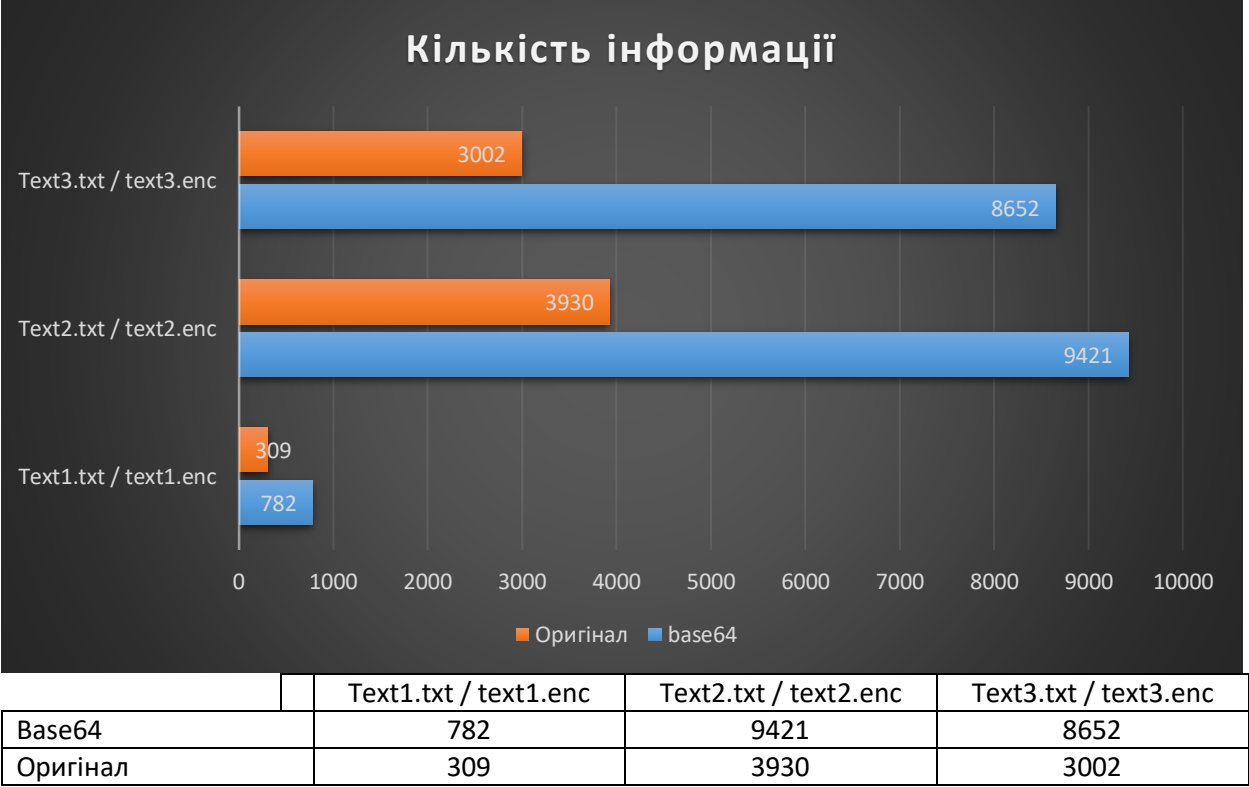
Це відбувається тому, що алгоритми архіваторів побудовані таким чином аби використати повторювані частини тексту. Виходячи з цього, формула розрахунку кількості інформації, використана для програми, не є досконалою, бо вона не враховує передбачення наступного шматочку тексту.

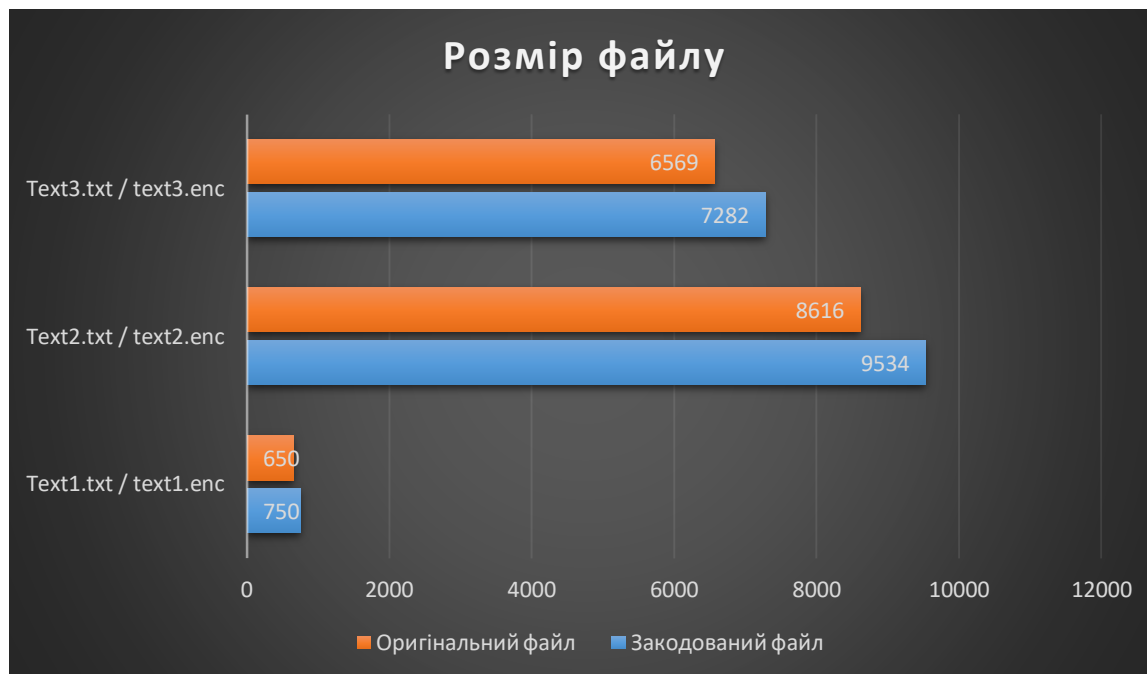
- Хочу звернути уваги на алгоритм bz2, який виявився найефективнішим у всіх випадках. А також навіть упорався «ідеальним стисненням» у випадку великого файлу, тобто стиснений архів має розмір навіть менший, ніж кі-сть інформації. У випадку з піснею, хоч файл і меншого розміру, ніж текст про SCP-173, але є багато частин повторюваного тексту (як-от приспів)

[illegible]

```
text3.txt.bz2 закодовано у Base64
Q1poTF8WSZTWXv73vv73vv70ADo+/v77vv70AEkdvv71u+Uo+/vUIQAAQF+/vdWt77+977+977+977+977+9AAAA77+9B39NQo+/ve+/ve+/vV3vv71s
Vu+/ve+/ve+/vR0077+9A1NBDEZTaO+/vqh677+9AGnv71NA9TQGHv73vv73vv73EJSbv70j77+9eu+/ve+/vwhoDQADQAx8Ke+/ve+/vUGaAAAA
AAAAABnv71I77+9ae+/ve+/ve+/vTVPyKPv73vv71PUO+/vT1AAAN77+977+9AAAWAAAAAAAEiTv73RoTTIk7vv73vv73vv708Ue+/vTQNAe+/vQOAA
Bu+/ve+/vS1LCu+/vT7vv73vv73vv73vv70yPe+/vXbv73vv73dtV4Xc+/v73vv73vv70f77+977+9MwAsA1Fvv73Iqkjvv705ETLvv73vv73vv73eme+/
v54FMu+/vvdKZe+/ve+/vWN677+977+9FBTVv73vv70T77+977+9J1rvv70o77+9S2QzZjnv73vv73vv70dUhhvv73vv73vv70U77+9XS3vv73vv70EWu+/
vQ/vv73Tqu+/vVd9IdmNT+/ve+/ve+/ve+/vD2V77+9PmtJCx7vv71F77+9Uhdvv71j77+9ch5ESkLv71kKj+/vULvv71ovD1H77+977+9ZHU377+9
Eyrvv71fW+/vTnv71fDe+/vSM877+977+9dw19Exbv71P77+977+9UQjvv71gf+/vQbv70c77+980+/ve+/vRrv73vv73vv70VPT1Z77+9bu+/
vXlw77+977+9Fo+/ve+/ve+/vW0WE+/vSrv70mb3E1YRV077+9eo+/vU0677+9OwHvv73ZuHB177+9Wvjvv73vv70bF1Pvv73vv73vv70Ww1MSu+/vTBr
ce+/ve+/ve+/vUgbeO+/vVEoce+/vTXvv70d0O+/ve+/vQ5XBj15LXd77+977+9agZ6Ve+/ve+/vVIk77+977+977+9FBIgC+/ve+/vUVMNu+/ve+/vut9
bnHvv73vv70s77+977+9NwXvv713G+/ve+/vS4A77+9ou+/vWf77+9bBu+/ve+/ve+/ve+/vTKpCnr73vv714Envv73vv70SeFp6Gu+/vTo0JO+/ve+/
ve+/ve+/vQXvv73vv73d00IQMu+/vUTvv70T77+9Ee+/vWczD1o+/ve+/vSjvv73vv71477+9JAJAeTEJJO+/ve+/vQ3vv70kDe+/vRoaaEgEhm77+9DZbv
v71C77+9Vv+/vURCwe+/vQATEhdv73vv70IUO+/vX1D77+9fn9EeHvv73vv73vv70cdTqgg1IJC7vv73vv70DMO+/vS1dO+/vRJMJO+/vQXvv71D77+9
77+977+9C+/vT94X+/ve+/ve+/ve+/vSjvv73vv71IJe+/ve+/vS5677+9Au+/vRggOu+/vQ1IXiXvv73vv71977+977+977+9Fe+/vVvXfU+/ve+/vwdj
JO+/vQYsP+/vRDvv71I77+9wo+/vV73IgQNCsO+/vR3vv70L77+977+9au+/vUnv71JF+/vQIGUHQm77+9I+/ve+/ve+/ve+/vWvvv71TVZ3vv73v
v73vv73vv70X77+9EADvv71i77+977+92oJgPsOYao+/vQzvv71m77+9d+/vR7WoS1g77+9cTLvv73vv71S77+90+/ve+/ve+/vW3vv70+Ee+/vWHD
77+9CqhiZCvv71pRO+/ve+/ve+/ve+/vS7vv71ZVe+/vRPvv714I2/vv701Xe+/vWjvv71YQnMm77+9Ut6TBu+/ve+/ve+/vQ1o77+977+977+9WFnvv715
77+977+977+977+977+9ZTw77+9P+/ve+/vSrv70V77+977+977+977+9eO+/vTLvv73vv71Re+/ve+/ve+/ve+/vXkEIu+/vV7vv70W77+9
LO+/ve+/vRjvv71a77+9De+/vCyF77+977+9Au+/vXASKO+/vFOos7jvv70z77+9M+/vS7vv70A77+9VvZvv70IG0Dvv73vv73vv73vv73vv707Re+/
vSVK77+9Z2F2Zu+/ve+/vQZBHUC0VvTv70C77+9Xe+/ve+/vSbv71c77+9UkIOce+/vU3vv70377+9JO+/vdyB00+/vUDvv717Me+/ve+/vR1z656B
aVpmYe+/vVHvv71C77+977+9Yu+/ve+/vWdv73vv70577+9OV/vv71e77+9TV8Ezr4g77+900+/ve+/vTJM77+9ZO+/vXMD77+977+9cu+/ve+/ve+/ve+/
ve+/ve+/vSZWmu+/vSj777+977+977+977+9oe+/vTbv717Du+/vRgiTO+/vRQpMu+/ve+/ve+/vTPvv73Xne+/vWofSw1177+977+9I3bv70S77+9
EAtw77+93gf77+9v71o77+9LjYxJmnvv70DUO+/vSf77+9Ljgalo77+9gf+/ve+/vVjEtO+/vDqYe+/ve+/ve+/vSnnv73vv73vv71mStvv710BWDv
v73vv73vv70kUEdv71q77+977+9FHNO77+9L2xr77+977+9aEvv73x17KwMmfnYBZF804Ve+/vXHGqx77+903bd2FLu+/ve+/ve+/vS2vv73vv71s77+9
MO+/vU3vv73vv70NKO+/vTvv73vv70wCw4M77+9U+/ve+/ve+/ve+/ve+/ve+/vQbv73vv73vv70V77+9EO+/vT3vv73vv71k77+9Q8tnCe+/
vUIQ77+9L+/vWtFyJrv73vv70V77+9G+/ve+/vQOX77+977+977+977+9KRZvv73NtSjN77+977+977+9Ee+/ve+/ve+/ve+/vW5Za3Lv71s77+9
77+9Yu+/ve+/vXbv73vv70577+9Be+/ve+/ve+/ve+/ve+/ve+/vW8jmr77+90477+9Ae+/ve+/ve+/vRvv707amf77+909Gec77+977+9ee+/vWjvv70L77+977+9
TO+/ve+/vWmIngsb77+9dmIhc37vv73vv73vv73vv71uEDrv70N77+977+9Q13vv73S1NOEwu+/vQLvv73vv71P77+9Vv+/ve+/ve+/vW8MFRbv70hGdkm
CALvv70Y77+977+9TU+/vUEIYi4hKyjvv73vv73vv70MMFTvv70BeEBKnbvv73vv71rJu+/ve+/ve+/vSgrLEguMO+/vSTvv73Equ+/vQwv77+9MO+/vVYY
XBg77+9UidaO+/ve+/vVzvv71jGxbZolbv719aAk377+9SdycUm8NOe+/vT1I77+90b4KY+/vSYcdO+/vVLvv73vv70477+9J+/vVbv70Q77+9DDPv
v71dyDRY77+9G0+/vQLvv71q77+977+977+977+977+90Bw4FO+/ve+/ve+/vTXw77+9Qu+/vURG77+977+9Luzvv73vv73vv73vv70f77+977+9c+/
ve+/vV9/77+9c1fvv70fQQU77+9Tt6z77+9V0+/ve+/vQ7vv71w77+977+9GydwCg5DD+/vTPvv73vv73vv70Da2VE77+9FDMC77+977+9Yu+/vXU4
Ie+/vQTvv71177+9c0+/vUXvv73vv70QRnHvv73vv70r77+9Nu+/vXhVve+/ve+/vXjvv71777+977+977+9HAgQNRagThjvv73vv70T77+9Xu+/ve+/vTFv
v70X77+9Ze+/ve+/vREJ77+9Me+/vWZzwqYIEe+/vWTVv71xe+/ve+/vWvOZmnvv709Ku+/ve+/vRQX77+977+9Yu+/vRhhJO+/ve+/vXLvv73vv73vv70T
Nu+/vSPvv70a77+9VS1ZGzBndInvv70cKO+/ve+/ve+/ve+/ve+/vX9j77+93ZyC77+9GDLvv71he+/vXMTV13Guu+/vQbv70U77+9Yo+/ve+/ve+/vVIV
dDIT77+977+977+9dhQRA+/vTXvv73vv73vv73vv73ckU4UJA1m77+977+977+9
```

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





Можна 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.

Найефективнішим алгоритмом стиснення виявився bzip2.

Висновок

У ході виконання лабораторної роботи ознайомився з поняттям ентропії інформації та пов'язаних з ентропією понять. Порівняв алгоритми стиснення – обрав кращий з них для випадків, коли треба буде зекономити місце на носії. Закріпив теоретичні знання практикою. 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 посиланням: [\(тут\)](#)