

Київський національний університет імені Тараса Шевченка
факультет радіофізики, електроніки та комп'ютерних систем

Лабораторна робота 1

Роботу виконав
студент 3 курсу
Перегінець Маргарита

Київ 2019

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

1. 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

text_1

```
Файл для аналізу: text_1.txt
Загальна кількість символів файлу: 744

Відносна частота появи літери "А" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Б" у тексті = 0,00134408602150538 ; Літера присутня у тексті: 1 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 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Е" у тексті = 0 ; Літера присутня у тексті: 0 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 ; Літера присутня у тексті: 0 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 = 0,00134408602150538 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Л" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "М" у тексті = 0,00268817204301075 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "Н" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "О" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "П" у тексті = 0,00134408602150538 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Р" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "С" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Т" у тексті = 0,00134408602150538 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "У" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "Ф" у тексті = 0,00134408602150538 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Х" у тексті = 0,00134408602150538 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Ц" у тексті = 0,00134408602150538 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "Ч" у тексті = 0 ; Літера присутня у тексті: 0 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 "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 = 0,00268817204301075 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "а" у тексті = 0,0631720430107527 ; Літера присутня у тексті: 47 разів.
Відносна частота появи літери "б" у тексті = 0,0134408602150538 ; Літера присутня у тексті: 10 разів.
Відносна частота появи літери "в" у тексті = 0,0389784946236559 ; Літера присутня у тексті: 29 разів.
Відносна частота появи літери "г" у тексті = 0,0161290322580645 ; 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 = 0,0255376344086022 ; Літера присутня у тексті: 19 разів.
Відносна частота появи літери "е" у тексті = 0,0443548387096774 ; Літера присутня у тексті: 33 разів.
Відносна частота появи літери "є" у тексті = 0,00537634408602151 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "ж" у тексті = 0,00806451612903226 ; Літера присутня у тексті: 6 разів.
Відносна частота появи літери "з" у тексті = 0,0147849462365591 ; Літера присутня у тексті: 11 разів.
Відносна частота появи літери "и" у тексті = 0,0510752688172043 ; Літера присутня у тексті: 38 разів.
Відносна частота появи літери "і" у тексті = 0,043010752688172 ; Літера присутня у тексті: 32 разів.
Відносна частота появи літери "ї" у тексті = 0,00134408602150538 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "й" у тексті = 0,010752688172043 ; Літера присутня у тексті: 8 разів.
```

[illegible]

Середня ентропія алфавіту для данного тексту: 4,65741436357988

Кількість інформації у тексті: 433,139535812929

text_2

Файл для аналізу: text_2.txt

Загальна кількість символів файлу: 309

[illegible]

Середня ентропія алфавіту для данного тексту: 4,61505423808293
Кількість інформації у тексті: 178,256469945953

```
Файл для аналізу: text_3.txt
Загальна кількість символів файлу: 141
```

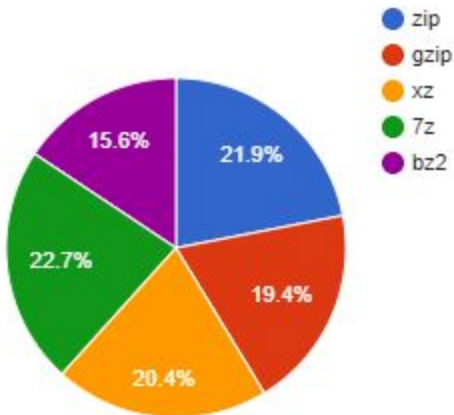
Відносна частота появи літери "к" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "л" у тексті = 0,0425531914893617 ; Літера присутня у тексті: 6 разів.
Відносна частота появи літери "м" у тексті = 0,00709219858156028 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "н" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "о" у тексті = 0,049645390070922 ; Літера присутня у тексті: 7 разів.
Відносна частота появи літери "п" у тексті = 0,00709219858156028 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "р" у тексті = 0,0283687943262411 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "с" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "т" у тексті = 0,0283687943262411 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "у" у тексті = 0,0141843971631206 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "ф" у тексті = 0,00709219858156028 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "х" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "ц" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "ч" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "ш" у тексті = 0,0141843971631206 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "щ" у тексті = 0,00709219858156028 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "ъ" у тексті = 0,0141843971631206 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "ю" у тексті = 0,00709219858156028 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "я" 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 "з" у тексті = 0,0212765957446809 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "(" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери ")" у тексті = 0,0141843971631206 ; 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 "-" 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 = 0,00709219858156028 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "1" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "2" у тексті = 0,0283687943262411 ; Літера присутня у тексті: 4 разів.
Відносна частота появи літери "3" у тексті = 0,0212765957446809 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "4" у тексті = 0,0212765957446809 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "5" у тексті = 0 ; Літера присутня у тексті: 0 разів.
Відносна частота появи літери "6" у тексті = 0,0141843971631206 ; Літера присутня у тексті: 2 разів.
Відносна частота появи літери "7" у тексті = 0,0212765957446809 ; Літера присутня у тексті: 3 разів.
Відносна частота появи літери "8" у тексті = 0,00709219858156028 ; Літера присутня у тексті: 1 разів.
Відносна частота появи літери "9" у тексті = 0 ; Літера присутня у тексті: 0 разів.

Середня ентропія алфавіту для данного тексту: 4,3829387296871
Кількість інформації у тексті: 77,2492951107351

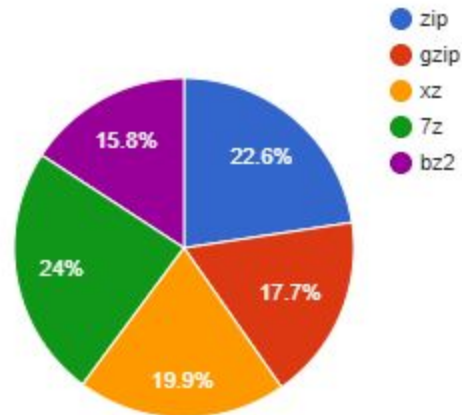
- Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення).

Файл	text_1.txt, Байт	text_2.txt, Байт	text_3.txt, Байт
zip	757	409	285
gzip	668	320	196
xz	704	360	220
7z	782	435	298
bz2	540	286	172
Розмір файлу	1 364	586	267
К-ть інформації	433	178	77

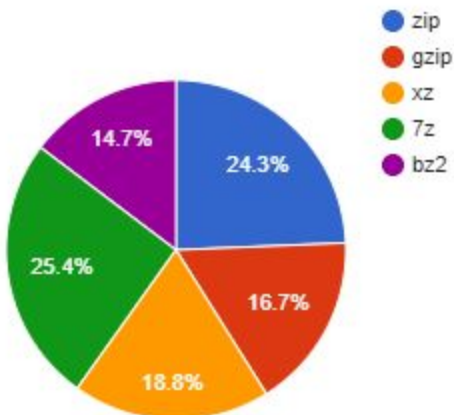
text_1



text_2



text_3



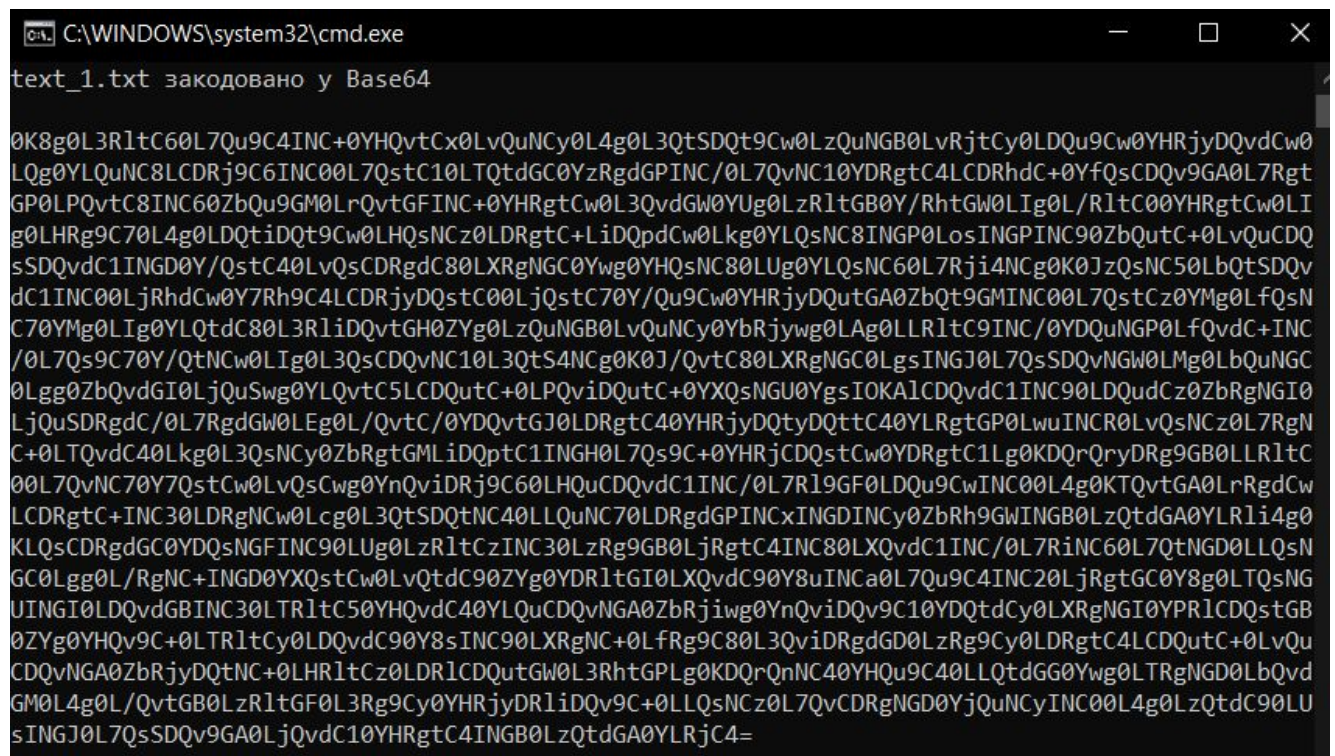
3. Порівняйте результуючі обсяги архівів з обчисленою кількістю інформації та **наведіть у звіті висновки** щодо кореляції цих величин для обраних вами файлів (яка відмінність, що вийшло більше і чому)¹

Щоб відбулося ідеальне стиснення розмір файлу повинен бути рівним кількості інформації. У реальності виходить, що розміри архівованих файлів більші за кількість інформації. Алгоритми архіваторів побудовані таким чином аби використати повторювані частини тексту (в мене це не видно, напевно тому що тексти недостатньо великі). Виходячи з цього, формула

розрахунку кількості інформації, використана для програми, не є досконалою, бо вона не враховує передбачення наступної частини тексту. Алгоритм bzip2 виявився найефективнішим у всіх випадках, а 7z найгірший. «Ідеального стисненням» файлу не відбулося.

2. Дослідження способів кодування інформації на прикладі Base64

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

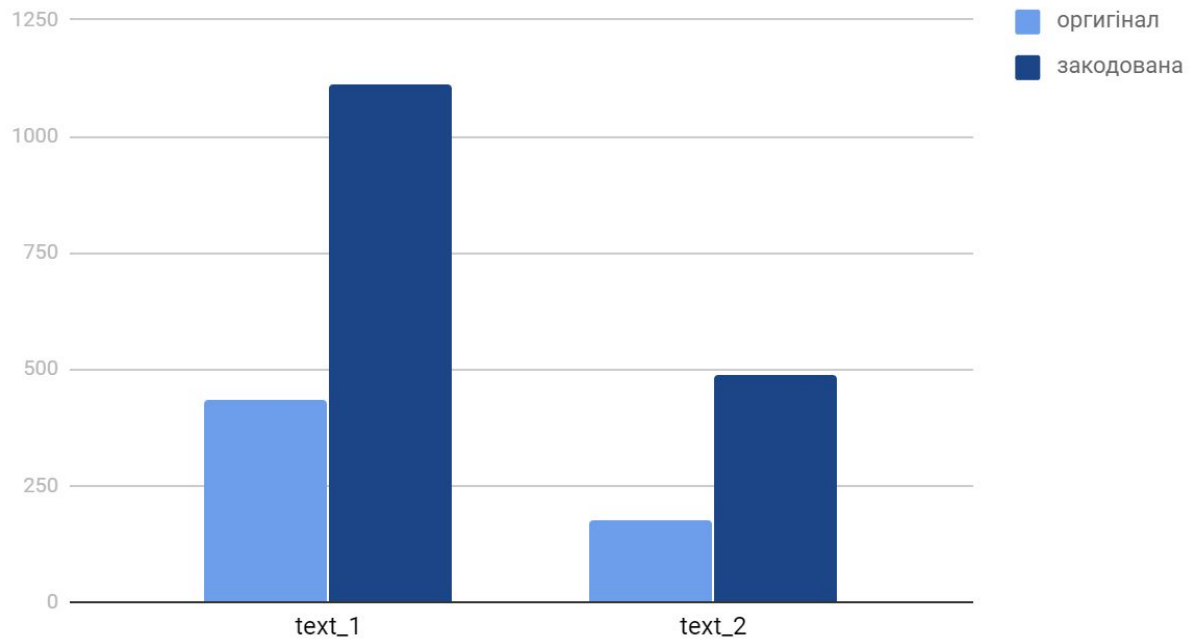


```
C:\WINDOWS\system32\cmd.exe
text_1.txt закодовано у Base64

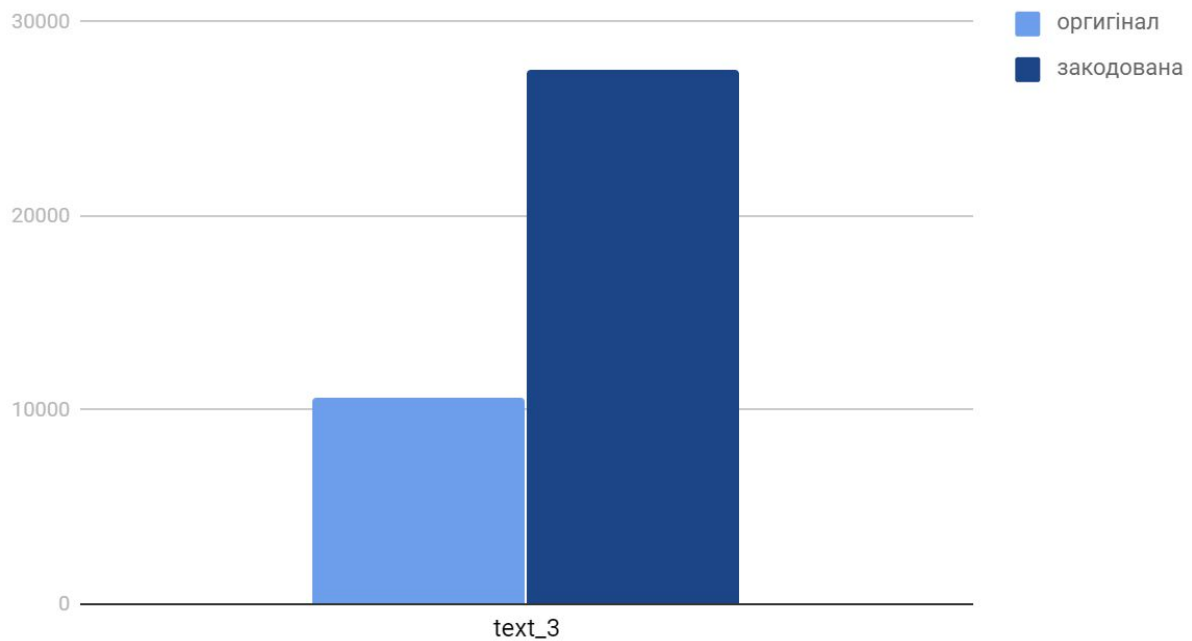
0K8g0L3RltC60L7Qu9C4INC+0YHQvtCx0LvQuNCy0L4g0L3QtSDQt9Cw0LzQuNGB0LvRjtCy0LDQu9Cw0YHRjyDQvdCw0
LQg0YLQuNC8LCDRj9C6INC00L7QstC10LTQtDGC0YzRgdGPINC/0L7QvNC10YDRgtC4LCDRhdc+0YfQsCDQv9GA0L7Rgt
GP0LPQvtC8INC60ZbQu9GM0LrQvtGFINC+0YHRgtCw0L3QvdGW0YUg0LzRltGB0Y/RhtGW0LIg0L/RltC00YHRgtCw0LI
g0LHRg9C70L4g0LDQtIDQt9Cw0LHQsNCz0LDRgtC+LiDQpdCw0Lkg0YLQsNC8INGP0LosINGPINC90ZbQutC+0LvQuCDQ
sSDQvdC1INGD0Y/QstC40LvQsCDRgdC80LXRgNGC0Ywg0YHQsNC80LUg0YLQsNC60L7Rji4NCg0K0JzQsNC50LbQtSDQv
dC1INC00LjRhdcw0Y7Rh9C4LCDRjyDQstC00LjQstC70Y/Qu9Cw0YHRjyDQutGA0ZbQt9GMINC00L7QstCz0YMG0LfQsN
C70YMG0LIg0YLQtDc80L3RliDQvtGH0ZYg0LzQuNGB0LvQuNCy0YbRjywg0LAG0LLRltC9INC/0YDQuNGP0LfQvdC+INC
/0L7Qs9C70Y/QtNCw0LIg0L3QsCDQvNC10L3QtS4NCg0K0J/QvtC80LXRgNGC0LgSINGJ0L7QsSDQvNGW0LMg0LbQuNGC
0Lgg0ZbQvdGI0LjQuSwg0YLQvtC5LCDQutC+0LPQviDQutC+0YXQsNGU0YgsIOKA1CDQvdC1INC90LDQudCz0ZbRgNGI0
LjQuSDRgdC/0L7RgdGW0LEg0L/QvtC/0YDQvtGJ0LDRgtC40YHRjyDQtyDQtC40YLRgtGP0LwuINCR0LvQsNCz0L7RgN
C+0LTQvdC40Lkg0L3QsNCy0ZbRgtGMLiDQptC1INGH0L7Qs9C+0YHRjCDQstCw0YDRgtC1Lg0KDQrQryDRg9GB0LLRltC
00L7QvNC70Y7QstCw0LvQsCwg0YnQviDRj9C60LHQUCDQvdC1INC/0L7Rl9GF0LDQu9CwINC00L4g0KTQvtGA0LrRgdCw
LCDRgtC+INC30LDRgNCw0Lcg0L3QtSDQtNC40LLQuNC70LDRgdGPINCxINGDINCy0ZbRh9GWINGB0LzQtdGA0YLRli4g0
KLQsCDRgdGC0YDQsNGFINC90LUg0LzRltCzINC30LzRg9GB0LjRgtC4INC80LXQvdC1INC/0L7RiNC60L7QtNGD0LLQsN
GC0Lgg0L/RgNC+INGD0YXQstCw0LvQtdC90ZYg0YDRltGI0LXQvdC90Y8uINCa0L7Qu9C4INC20LjRgtGC0Y8g0LTQsNG
UINGI0LDQvdGBINC30LTRltC50YHQvdC40YLQuCDQvNGA0ZbRjiwg0YnQviDQv9C10YDQtdCy0LXRgNGI0YPRICDQstGB
0ZYg0YHQv9C+0LTRltCy0LDQvdC90Y8sINC90LXRgNC+0LfRg9C80L3QviDRgdGD0LzRg9Cy0LDRgtC4LCDQutC+0LvQu
CDQvNGA0ZbRjyDQtNC+0LHRltCz0LDRICDQutGW0L3RhtGPLg0KDQrQnNC40YHQv9C40LLQtdGG0Ywg0LTRgNGD0LbQvd
GM0L4g0L/QvtGB0LzRltGF0L3Rg9Cy0YHRjyDRliDQv9C+0LLQsNCz0L7QvCDRgNGD0YjQuNCyINC00L4g0LzQtdC90LU
sINGJ0L7QsSDQv9GA0LjQvdC10YHRgtC4INGB0LzQtdGA0YLRjC4=
```


[illegible]

Points scored

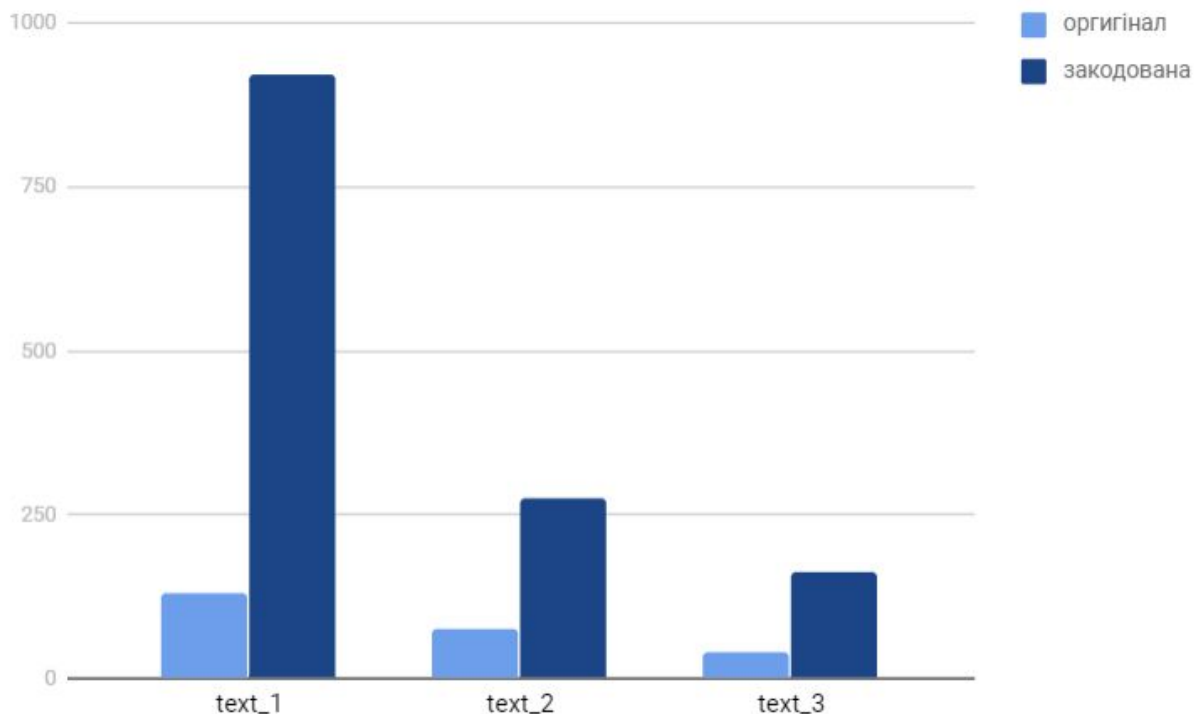


Points scored



За допомогою стовпчикової діаграми видно, що розмір закодованих файлах зріс. Це пов'язано з алгоритмом кодування base64.

3. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
 - а. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
 - б. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу²
 - с. Зробіть висновки з отриманого результату



Теж саме, що і в попередньому пункті.

Висновок: під час лабораторної роботи було досліджено імовірнісні параметри української мови для оцінки кількості інформації текстів. Досліджено вплив різних методів кодування інформації на її кількість. Також було порівняннюно алгоритми стиснення та обраго кращий з них для випадків, коли треба буде зекономити місце на носії. Теоретично та практично ознайомилась з алгоритмом кодування Base64, та дослідила переваги та недоліки.