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 3 курсу
КІ - СА
Гулівець Владислав
Андрійович

Київ 2020

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

Теоретичні відомості

Відносна частота появи символу *- імовірність появи певного символу в певному місці тексту - відношення числа появи символу в тексті до загальної кількості символів.*

Середня ентропія нерівноймовірного алфавіту*:*

$$H = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{m} p_i \log_2 p_i$$

де m - кількість символів алфавіту, p - імовірність появи символу
Ентропія вимірюється в ***БІТАХ*** *(як представлення кількості можливих варіантів).*

Кількість інформації в тексті *- середня ентропія вихідного алфавіту помножена на кількість символів тексту. (****HINT****: результат обрахунку для порівняння значення з розміром файлів треба перевести з бітів в байти)*

Хід виконання роботи:

Дослідження кількості інформації в тексті

1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI)

    - text1.txt – куплет пісні Скрябіна.
    - text2.txt – історія України за 1917 рік.
    - text3.txt – рецензія на фільм «Захар Беркут».

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

Analyze of ***text1.txt***

```
Цей символ А зустр?чається 14 раз з ймов?рн?стю 0,06034483. Ентроп?я символа = 0,24443433
Цей символ Б зустр?чається 3 раз з ймов?рн?стю 0,01293103. Ентроп?я символа = 0,08111662
Цей символ В зустр?чається 11 раз з ймов?рн?стю 0,04741379. Ентроп?я символа = 0,20855191
Цей символ Г зустр?чається 2 раз з ймов?рн?стю 0,00862069. Ентроп?я символа = 0,05912053
Цей символ ? зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Д зустр?чається 3 раз з ймов?рн?стю 0,01293103. Ентроп?я символа = 0,08111662
Цей символ Е зустр?чається 7 раз з ймов?рн?стю 0,03017241. Ентроп?я символа = 0,15238958
Цей символ Є зустр?чається 1 раз з ймов?рн?стю 0,00431034. Ентроп?я символа = 0,03387061
Цей символ Ж зустр?чається 4 раз з ймов?рн?стю 0,01724138. Ентроп?я символа = 0,10099967
Цей символ З зустр?чається 5 раз з ймов?рн?стю 0,02155172. Ентроп?я символа = 0,11931148
Цей символ И зустр?чається 15 раз з ймов?рн?стю 0,06465517. Ентроп?я символа = 0,25545843
Цей символ ? зустр?чається 15 раз з ймов?рн?стю 0,06465517. Ентроп?я символа = 0,25545843
Цей символ Ї зустр?чається 2 раз з ймов?рн?стю 0,00862069. Ентроп?я символа = 0,05912053
Цей символ Й зустр?чається 1 раз з ймов?рн?стю 0,00431034. Ентроп?я символа = 0,03387061
Цей символ К зустр?чається 9 раз з ймов?рн?стю 0,03879310. Ентроп?я символа = 0,18186424
Цей символ Л зустр?чається 8 раз з ймов?рн?стю 0,03448276. Ентроп?я символа = 0,16751659
Цей символ М зустр?чається 9 раз з ймов?рн?стю 0,03879310. Ентроп?я символа = 0,18186424
Цей символ Н зустр?чається 8 раз з ймов?рн?стю 0,03448276. Ентроп?я символа = 0,16751659
Цей символ О зустр?чається 14 раз з ймов?рн?стю 0,06034483. Ентроп?я символа = 0,24443433
Цей символ П зустр?чається 8 раз з ймов?рн?стю 0,03448276. Ентроп?я символа = 0,16751659
Цей символ Р зустр?чається 6 раз з ймов?рн?стю 0,02586207. Ентроп?я символа = 0,13637117
Цей символ С зустр?чається 10 раз з ймов?рн?стю 0,04310345. Ентроп?я символа = 0,19551952
Цей символ Т зустр?чається 7 раз з ймов?рн?стю 0,03017241. Ентроп?я символа = 0,15238958
Цей символ У зустр?чається 5 раз з ймов?рн?стю 0,02155172. Ентроп?я символа = 0,11931148
Цей символ Ф зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Х зустр?чається 2 раз з ймов?рн?стю 0,00862069. Ентроп?я символа = 0,05912053
Цей символ Ц зустр?чається 1 раз з ймов?рн?стю 0,00431034. Ентроп?я символа = 0,03387061
Цей символ Ч зустр?чається 1 раз з ймов?рн?стю 0,00431034. Ентроп?я символа = 0,03387061
Цей символ Ш зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Щ зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Ь зустр?чається 4 раз з ймов?рн?стю 0,01724138. Ентроп?я символа = 0,10099967
Цей символ Ю зустр?чається 1 раз з ймов?рн?стю 0,00431034. Ентроп?я символа = 0,03387061
Цей символ Я зустр?чається 8 раз з ймов?рн?стю 0,03448276. Ентроп?я символа = 0,16751659
Цей символ   зустр?чається 38 раз з ймов?рн?стю 0,16379310. Ентроп?я символа = 0,42750876
Загальна ентроп?я: 4,25588104
HINT файлу: 987,36440108
```

Analyze of ***text2.txt***

```
Ця програма рахує в?дносну частоту появи символу в текст?.
В текст? було знайдено наступн? букви:

Цей символ А зустр?чається 14 раз з ймов?рн?стю 0,02513465. Ентроп?я символа = 0,13357002
Цей символ Б зустр?чається 3 раз з ймов?рн?стю 0,00538600. Ентроп?я символа = 0,04059194
Цей символ В зустр?чається 11 раз з ймов?рн?стю 0,01974865. Ентроп?я символа = 0,11181889
Цей символ Г зустр?чається 2 раз з ймов?рн?стю 0,00359066. Ентроп?я символа = 0,02916170
Цей символ ? зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Д зустр?чається 3 раз з ймов?рн?стю 0,00538600. Ентроп?я символа = 0,04059194
Цей символ Е зустр?чається 7 раз з ймов?рн?стю 0,01256732. Ентроп?я символа = 0,07935233
Цей символ Є зустр?чається 1 раз з ймов?рн?стю 0,00179533. Ентроп?я символа = 0,01637618
Цей символ Ж зустр?чається 4 раз з ймов?рн?стю 0,00718133. Ентроп?я символа = 0,05114207
Цей символ З зустр?чається 5 раз з ймов?рн?стю 0,00897666. Ентроп?я символа = 0,06103775
Цей символ И зустр?чається 15 раз з ймов?рн?стю 0,02692998. Ентроп?я символа = 0,14043024
Цей символ ? зустр?чається 15 раз з ймов?рн?стю 0,02692998. Ентроп?я символа = 0,14043024
Цей символ Ї зустр?чається 2 раз з ймов?рн?стю 0,00359066. Ентроп?я символа = 0,02916170
Цей символ Й зустр?чається 1 раз з ймов?рн?стю 0,00179533. Ентроп?я символа = 0,01637618
Цей символ К зустр?чається 9 раз з ймов?рн?стю 0,01615799. Ентроп?я символа = 0,09616603
Цей символ Л зустр?чається 8 раз з ймов?рн?стю 0,01436266. Ентроп?я символа = 0,08792149
Цей символ М зустр?чається 9 раз з ймов?рн?стю 0,01615799. Ентроп?я символа = 0,09616603
Цей символ Н зустр?чається 8 раз з ймов?рн?стю 0,01436266. Ентроп?я символа = 0,08792149
Цей символ О зустр?чається 14 раз з ймов?рн?стю 0,02513465. Ентроп?я символа = 0,13357002
Цей символ П зустр?чається 8 раз з ймов?рн?стю 0,01436266. Ентроп?я символа = 0,08792149
Цей символ Р зустр?чається 6 раз з ймов?рн?стю 0,01077199. Ентроп?я символа = 0,07041190
Цей символ С зустр?чається 10 раз з ймов?рн?стю 0,01795332. Ентроп?я символа = 0,10412218
Цей символ Т зустр?чається 7 раз з ймов?рн?стю 0,01256732. Ентроп?я символа = 0,07935233
Цей символ У зустр?чається 5 раз з ймов?рн?стю 0,00897666. Ентроп?я символа = 0,06103775
Цей символ Ф зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Х зустр?чається 2 раз з ймов?рн?стю 0,00359066. Ентроп?я символа = 0,02916170
Цей символ Ц зустр?чається 1 раз з ймов?рн?стю 0,00179533. Ентроп?я символа = 0,01637618
Цей символ Ч зустр?чається 1 раз з ймов?рн?стю 0,00179533. Ентроп?я символа = 0,01637618
Цей символ Ш зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Щ зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Ь зустр?чається 4 раз з ймов?рн?стю 0,00718133. Ентроп?я символа = 0,05114207
Цей символ Ю зустр?чається 1 раз з ймов?рн?стю 0,00179533. Ентроп?я символа = 0,01637618
Цей символ Я зустр?чається 8 раз з ймов?рн?стю 0,01436266. Ентроп?я символа = 0,08792149
Цей символ   зустр?чається 38 раз з ймов?рн?стю 0,06822262. Ентроп?я символа = 0,26426756
Загальна ентроп?я: 2,27625325
HINT файлу: 1267,87306101
```

Analyze of ***text3.txt***

```
Ця програма рахує в?дносну частоту появи символу в текст?.
В текст? було знайдено наступн? букви:

Цей символ А зустр?чається 14 раз з ймов?рн?стю 0,03063457. Ентроп?я символа = 0,15405194
Цей символ Б зустр?чається 3 раз з ймов?рн?стю 0,00656455. Ентроп?я символа = 0,04760014
Цей символ В зустр?чається 11 раз з ймов?рн?стю 0,02407002. Ентроп?я символа = 0,12941533
Цей символ Г зустр?чається 2 раз з ймов?рн?стю 0,00437637. Ентроп?я символа = 0,03429344
Цей символ ? зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Д зустр?чається 3 раз з ймов?рн?стю 0,00656455. Ентроп?я символа = 0,04760014
Цей символ Е зустр?чається 7 раз з ймов?рн?стю 0,01531729. Ентроп?я символа = 0,09234326
Цей символ Є зустр?чається 1 раз з ймов?рн?стю 0,00218818. Ентроп?я символа = 0,01933490
Цей символ Ж зустр?чається 4 раз з ймов?рн?стю 0,00875274. Ентроп?я символа = 0,05983414
Цей символ З зустр?чається 5 раз з ймов?рн?стю 0,01094092. Ентроп?я символа = 0,07127048
Цей символ И зустр?чається 15 раз з ймов?рн?стю 0,03282276. Ентроп?я символа = 0,16178861
Цей символ ? зустр?чається 15 раз з ймов?рн?стю 0,03282276. Ентроп?я символа = 0,16178861
Цей символ Ї зустр?чається 2 раз з ймов?рн?стю 0,00437637. Ентроп?я символа = 0,03429344
Цей символ Й зустр?чається 1 раз з ймов?рн?стю 0,00218818. Ентроп?я символа = 0,01933490
Цей символ К зустр?чається 9 раз з ймов?рн?стю 0,01969365. Ентроп?я символа = 0,11158671
Цей символ Л зустр?чається 8 раз з ймов?рн?стю 0,01750547. Ентроп?я символа = 0,10216281
Цей символ М зустр?чається 9 раз з ймов?рн?стю 0,01969365. Ентроп?я символа = 0,11158671
Цей символ Н зустр?чається 8 раз з ймов?рн?стю 0,01750547. Ентроп?я символа = 0,10216281
Цей символ О зустр?чається 14 раз з ймов?рн?стю 0,03063457. Ентроп?я символа = 0,15405194
Цей символ П зустр?чається 8 раз з ймов?рн?стю 0,01750547. Ентроп?я символа = 0,10216281
Цей символ Р зустр?чається 6 раз з ймов?рн?стю 0,01312910. Ентроп?я символа = 0,08207118
Цей символ С зустр?чається 10 раз з ймов?рн?стю 0,02188184. Ентроп?я символа = 0,12065913
Цей символ Т зустр?чається 7 раз з ймов?рн?стю 0,01531729. Ентроп?я символа = 0,09234326
Цей символ У зустр?чається 5 раз з ймов?рн?стю 0,01094092. Ентроп?я символа = 0,07127048
Цей символ Ф зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Х зустр?чається 2 раз з ймов?рн?стю 0,00437637. Ентроп?я символа = 0,03429344
Цей символ Ц зустр?чається 1 раз з ймов?рн?стю 0,00218818. Ентроп?я символа = 0,01933490
Цей символ Ч зустр?чається 1 раз з ймов?рн?стю 0,00218818. Ентроп?я символа = 0,01933490
Цей символ Ш зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Щ зустр?чається 0 раз з ймов?рн?стю 0,00000000. Ентроп?я символа = 0,00000000
Цей символ Ь зустр?чається 4 раз з ймов?рн?стю 0,00875274. Ентроп?я символа = 0,05983414
Цей символ Ю зустр?чається 1 раз з ймов?рн?стю 0,00218818. Ентроп?я символа = 0,01933490
Цей символ Я зустр?чається 8 раз з ймов?рн?стю 0,01750547. Ентроп?я символа = 0,10216281
Цей символ   зустр?чається 38 раз з ймов?рн?стю 0,08315098. Ентроп?я символа = 0,29835595
Загальна ентроп?я: 2,63565820
HINT файлу: 1204,49579898
```

Entropy:

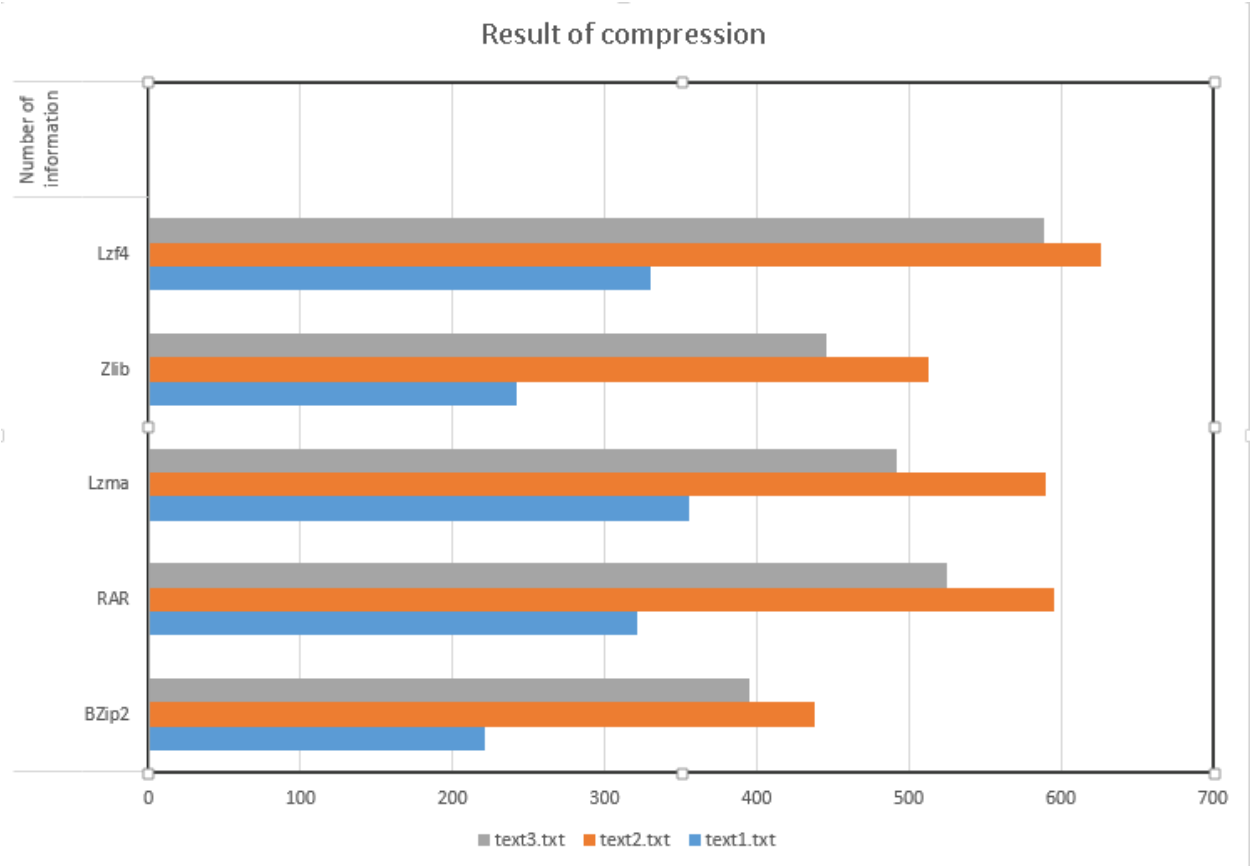
text1.txt – 4.25588104
text2.txt – 2.27625325
text3.txt – 2.63565820

Size of file (bytes):

text1.txt – 987bit/8 = 123.375bytes
text2.txt – 1277bit/8 = 159.625bytes
text3.txt – 1204bit/8 = 150.5bytes

Result of compression

Назва файлу	BZip2	RAR	Lzma	Zlib	Lzf4	Number of information
text1.txt	221	321	356	242	330	123.375
text2.txt	438	595	590	513	626	159.625
text3.txt	395	525	492	446	589	150.5



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).
   Перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)
3. Закодуйте в Base64 обрані вами текстові файли
   Обрахуйте кількість інформації в base64-закодованому варіанті файлу
   Порівняйте отримане значення з кількістю інформації вихідного файлу
   Зробіть висновки з отриманого результату
4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
   Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
   Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу
   Зробіть висновки з отриманого результату

Хід виконання роботи:

Base64 compression of ***text1.txt***

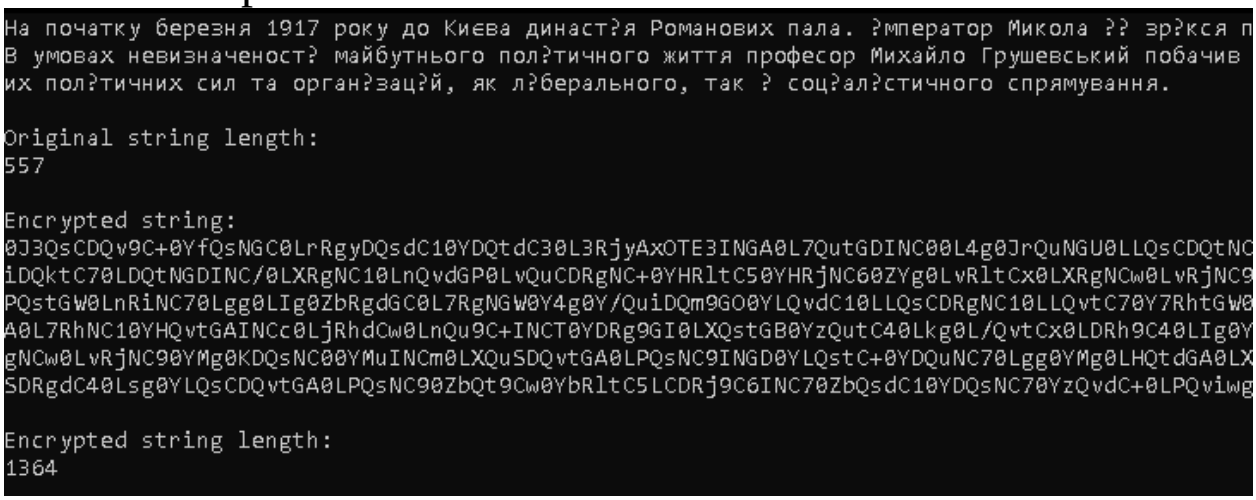
```
Original string:
Я завжди мр?яв написати п?сню про маму
Але р?зн? поети вс? слова вже сказали
? я не хот?в повторити когось ?з них
Я б?ля своєї мами буду завжди маленьким
? як т?льки покличе приб?жу скоренько
Тому я їй ? написав ц? слова як м?г

Original string length:
232

Encrypted string:
0K8g0LfQsNCy0LbQtNC4INC80YDRltGP0LIg0L3QsNC/0LjRgdCw0YLQuCDQv9GW0YHQvdGOINC/0YDQviDQvNCw0LzRgw0K0JD
NGC0Lgg0LrQvtCz0L7RgdGMINGW0Lcg0L3QuNGFDQrQryDQsdGW0LvRjyDRgdCy0L7RlNGXINC80LDQvNC4INCx0YPQtNGDINC3
LQvtC80YMg0Y8g0ZfQuSDRliDQvdCw0L/QuNGB0LDQsiDRhtGWINGB0LvQvtCy0LAg0Y/QuiDQvNGW0LM=

Encrypted string length:
556
```

Base64 compression of ***text2.txt***



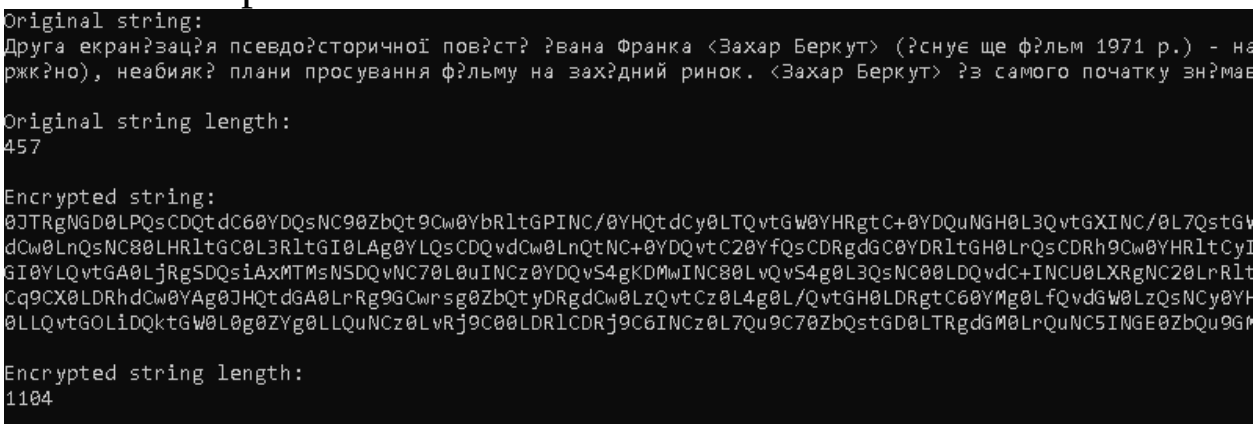
```
На початку березня 1917 року до Києва династ?я Романових пала. ?мператор Микола ?? зр?кся п
В умовах невизначеност? майбутнього пол?тичного життя професор Михайло Грушевський побачив
их пол?тичних сил та орган?зац?й, як л?берального, так ? соц?ал?стичного спрямування.

Original string length:
557

Encrypted string:
0J3QsCDQv9C+0YfQsNGC0LrRgyDQsdC10YDQtdC30L3RjyAxOTE3INGA0L7QutGDINC00L4g0JrQuNGU0LLQsCDQtNC
iDQktC70LDQtNGDINC/0LXRgNC10LnQvdGP0LvQuCDRgNC+0YHRltC50YHRjNC60ZYg0LvRltCx0LXRgNCw0LvRjNC9
PQstGW0LnRiNC70Lgg0LIg0ZbRgdGC0L7RgNGW0Y4g0Y/QuiDQm9GO0YLQvdC10LLQsCDRgNC10LLQvtC70Y7RhtGW0
A0L7RhNC10YHQvtGAINCc0LjRhdCw0LnQu9C+INCT0YDRg9GI0LXQutC40Lkg0L/QvtCx0LDRh9C40LIg0Y
gNCw0LvRjNC90YMg0KDQsNC00YMuINCm0LXQuSDQvtGA0LltC+0YDQuNC70Lgg0YMg0LHQtdGA0LXQ
SDRgdC40Lsg0YLQsCDQvtGA0LPQsNC90ZbQt9Cw0YbRltC5LCDRj9C6INC70ZbQsdC10YDQsNC70YzQvdC+0LPQviwg

Encrypted string length:
1364
```

Base64 compression of ***text3.txt***



```
Original string:
Друга екран?зац?я псевдо?сторичної пов?ст? ?вана Франка <Захар Беркут> (?снує ще ф?льм 1971 р.) - на
ржк?но), неабияк? плани просування ф?льму на зах?дний ринок. <Захар Беркут> ?з самого початку зн?мав

Original string length:
457

Encrypted string:
0JTRgNGD0LPQsCDQtdC60YDQsNC90ZbQt9Cw0YbRltGPINC/0YHQtdCy0LTQvtGW0YHRgtC+0YDQuNGH0L3QvtGXINC/0L7QstGv
dCw0LnQsNC80LHRltGC0L3RltGI0LAg0YLQsCDQvdCw0LnQtC+0YDQvtC20YfQsCDRgdC0YDltGH0LrQsCDRh9Cw0YHRltCyI
GI0YLQvtGA0LjRgSDQsiAxMTMsNSDQvNC70L0uINCz0YDQvS4gKDMwINC80LvQvS4g0L3QsNC00LDQvdC+INCU0LXRgNC20LRlt
Cq9CX0LDRhdCw0YAg0JHQtdGA0LrRg9GCwrsg0ZbQtyDRgdCw0LzQvtCz0L4g0L/QvtGH0LDRgtC60YMg0LfQsNCy0YH
0LLQvtGOLiDQktGW0L0g0ZYg0LLQuNCz0LvRj9C00LDRlCDRj9C6INCz0L7RltGD0LTRgdGM0LrQvtGXINC5INGE0ZbQu9GM

Encrypted string length:
1104
```

Size of file (bytes):
text1encrypted.txt = 1267bit/8 = 158.375 bytes
text2encrypted.txt = 1554bit/8 = 194.25 bytes
text3encrypted.txt = 1486bit/8 = 185.75 bytes

Entropy:
text1.txt – 4.25588104
text2.txt – 2.27625325
text3.txt – 2.63565820

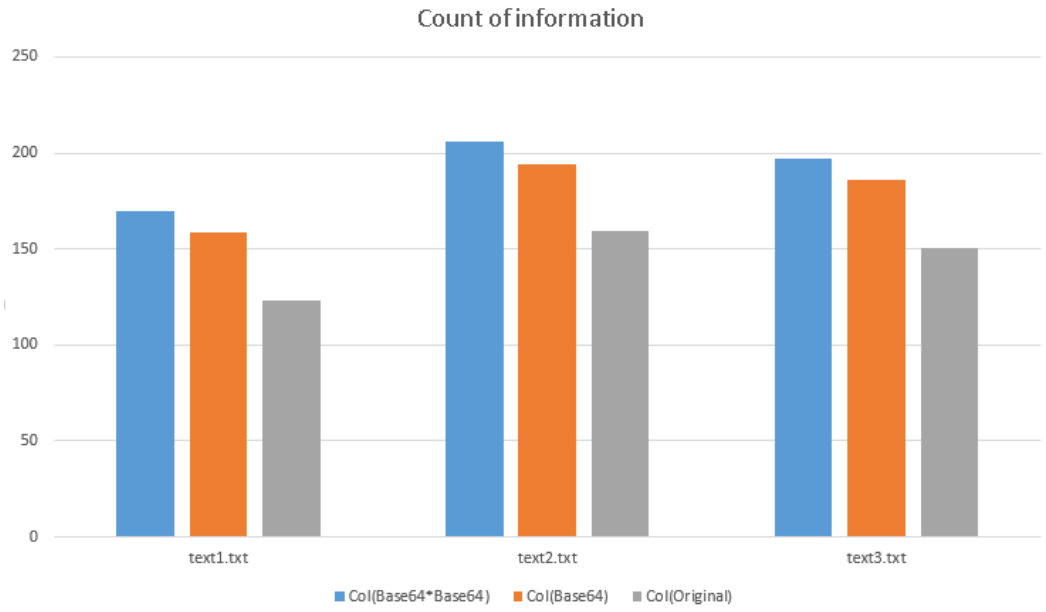
Size of file (bytes):
text1.txt – 987bit/8 = 123.375bytes
text2.txt – 1277bit/8 = 159.625bytes
text3.txt – 1204bit/8 = 150.5bytes

Назва файлу	Entropy	Length	Count of information
text1encrypted.txt	2,2793	556	158.375
text2encrypted.txt	1,1398	1364	194.25
text3encrypted.txt	1,3469	1104	185.75

Назва файлу	BZip2	RAR	Lzma	Zlib	Lzf4
text1.txt	221	321	356	242	330
text2.txt	438	595	590	513	603
text3.txt	395	525	492	446	541

Назва файлу	BZip(BASE64)	Count of information	Number of information
text1encrypted.txt	335	158.375	123.375
text2encrypted.txt	630	194.25	159.625
text3encrypted.txt	577	185.75	150.5

Назва файлу	CoI(BASE64)^2	Number of information
text1encrypted.txt	170	123.375
text2encrypted.txt	205.875	159.625
text3encrypted.txt	197.375	150.5



Висновок: В цій лабораторній роботі мною була зроблена програма що аналізує текст та обчислює ймовірності появи букви в тексті, ентропію тексту, загальний обсяг інформації. Мною було проаналізований обсяг зайнятого простору на диску файлів з різним типом стиснення. Як результат аналізу можна сказати, що обсяг інформації є значно меншим аніж обсяг який займає стиснутий цей же файл на диску. Було встановлено що найкращим з перевірених алгоритмом стиснення є BZip2. Об'єм його файлів є найближчим до кількості інформації. Також були опрацьовані навички в кодуванні ***Base64***. Як результат можна сказати, що с кожним повторним кодуванням файлу обсяг інформації збільшується. Це зумовлено тим, що сама кількість символів в файлі збільшується, а отже й сам обсяг збільшується також.

My personal GitHub link: ***https://github.com/gulivec84***