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
Мургашов Гліб

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 роботи

1. Дослідження кількості інформації в тексті

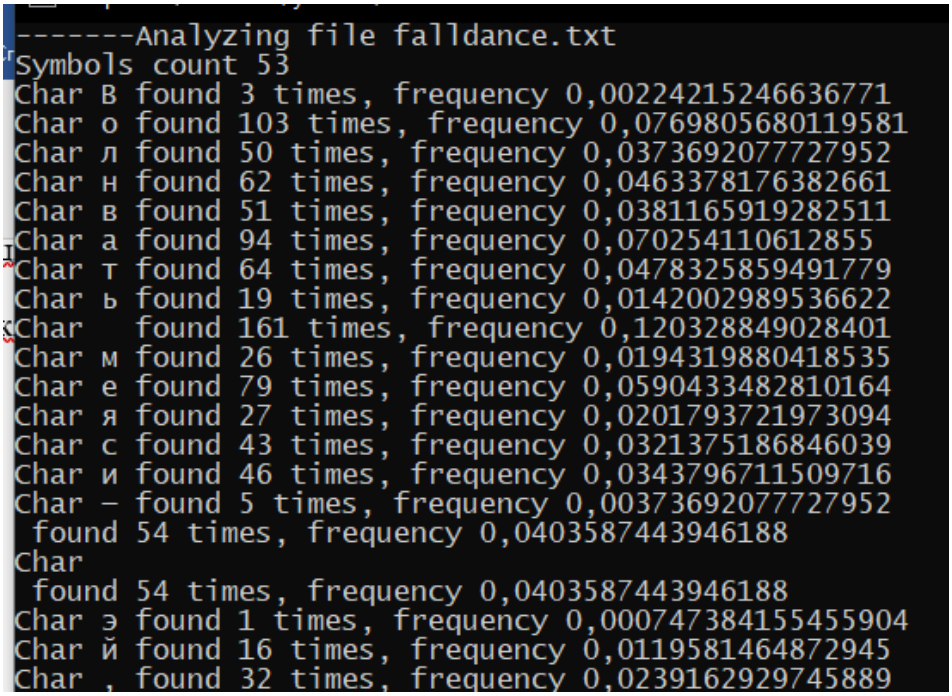
1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI)

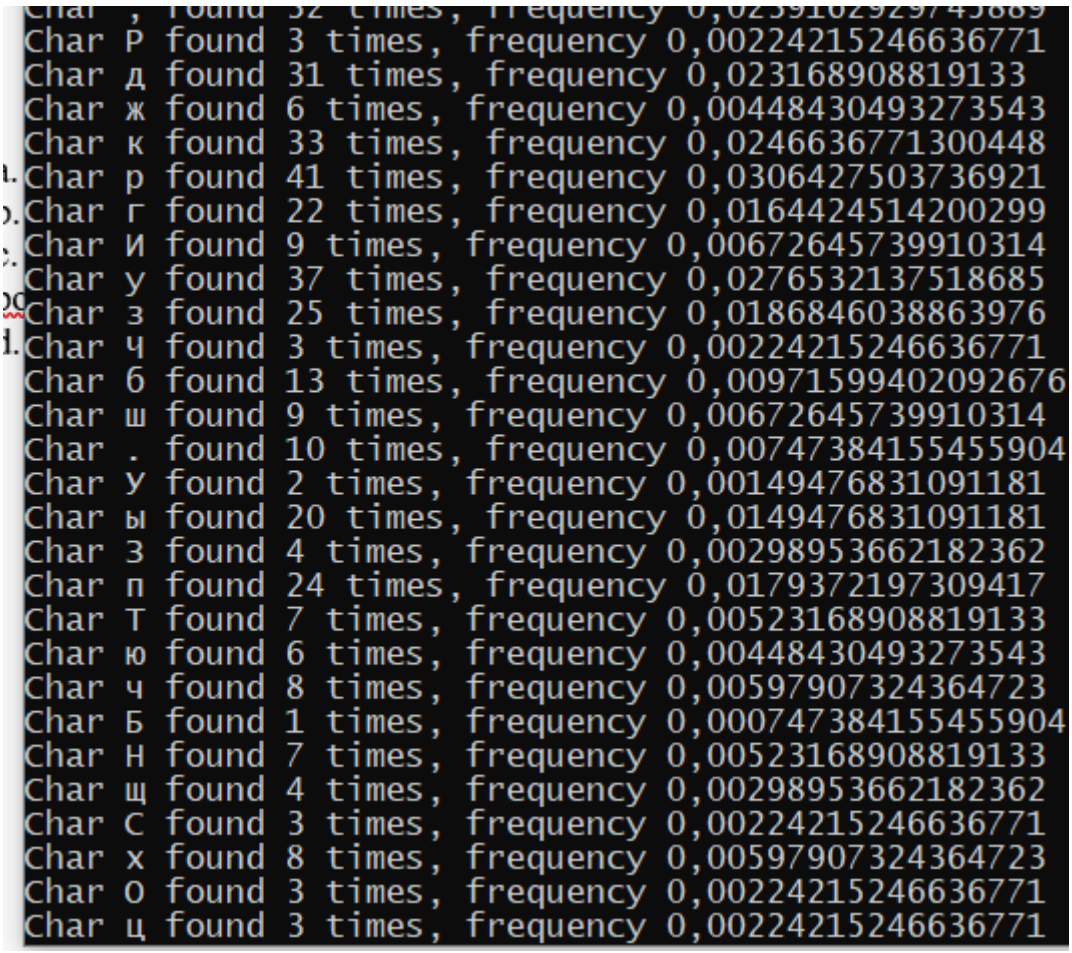
    1. **falldance.txt** – Вірш. Автор: Олександр Блок. «Пляски осенние»
    2. **briefHistoryOfTime.txt** – Уривок з книги. Автор: С. Хокинг. Назва: “Краткая история времени”.
    3. **constitution.txt –** уривок із Конституції України

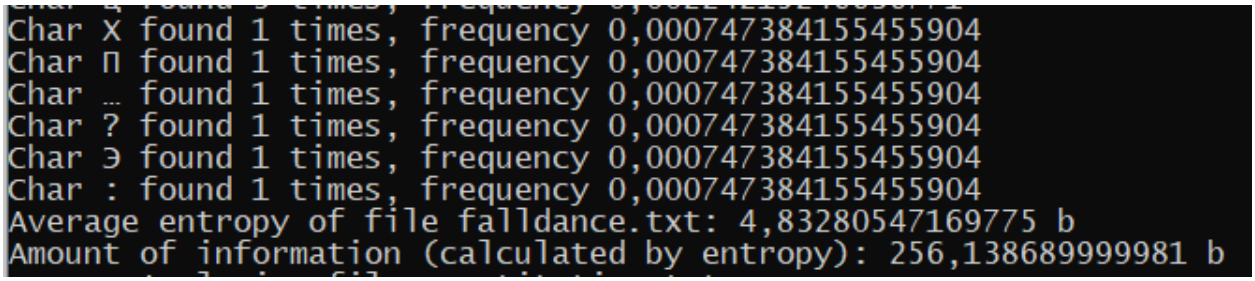
2. Переконайтесь, що тексти, які ви використовуєте є унікальними і не повторюються у ваших колег! Використовуйте наявні електронні засоби зв’язку та документообігу, щоб уникнути дублювання! Вдруге аналіз того самого тексту не зараховується!
3. Створіть програму (будь-якою зручною для вас мовою), яка 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

falldance.txt

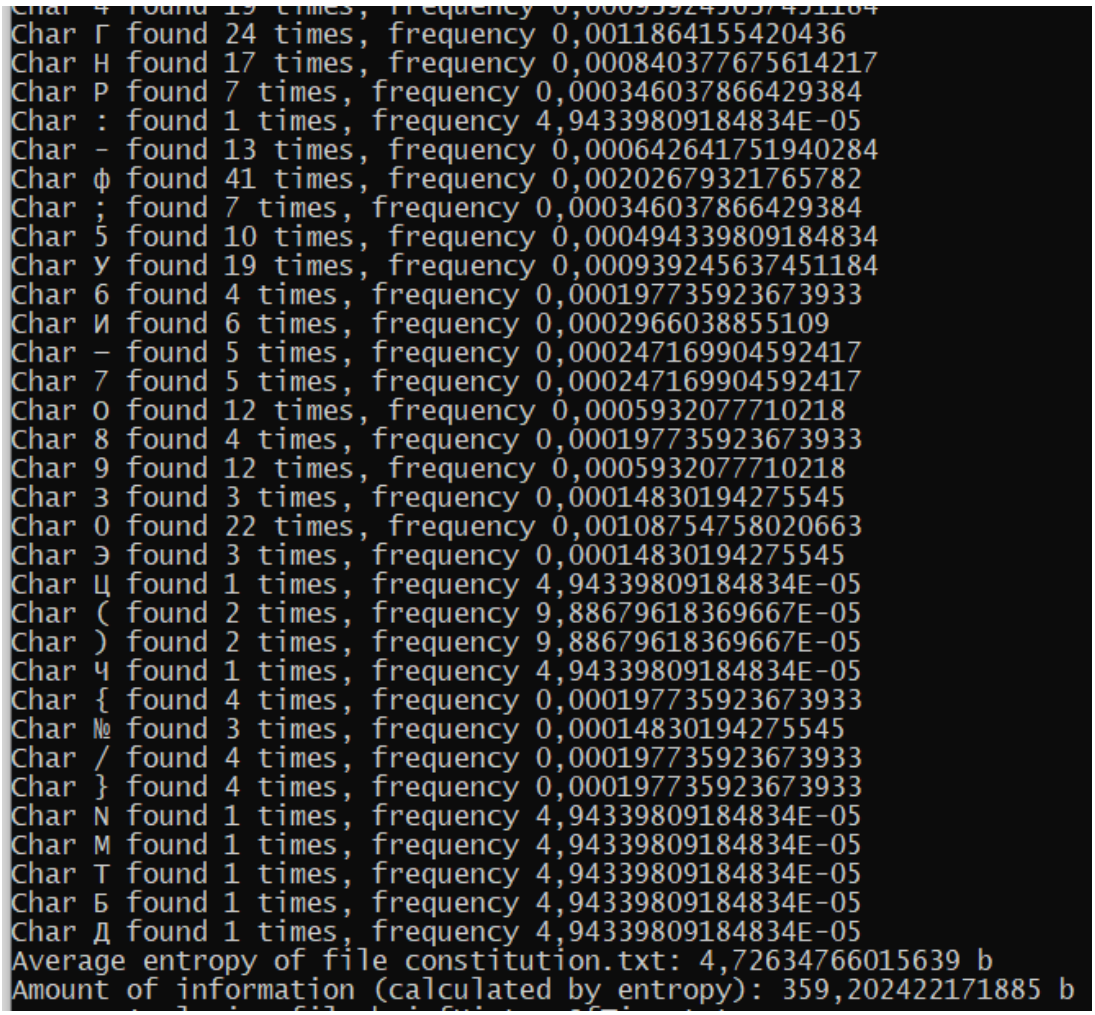



Char Р found 3 times, frequency 0,00224215246636771
Char д found 31 times, frequency 0,023168908819133
Char ж found 6 times, frequency 0,00448430493273543
Char к found 33 times, frequency 0,0246636771300448
Char р found 41 times, frequency 0,0306427503736921
Char г found 22 times, frequency 0,01644245142002 99
Char И found 9 times, frequency 0,00672645739910314
Char у found 37 times, frequency 0,0276532137518685
Char з found 25 times, frequency 0,0186846038863976
Char ч found 3 times, frequency 0,00224215246636771
Char б found 13 times, frequency 0,00971599402092676
Char ш found 9 times, frequency 0,00672645739910314
Char . found 10 times, frequency 0,00747384155455904
Char У found 2 times, frequency 0,00149476831091181
Char ы found 20 times, frequency 0,0149476831091181
Char З found 4 times, frequency 0,00298953662182362
Char п found 24 times, frequency 0,0179372197309417
Char Т found 7 times, frequency 0,00523168908819133
Char ю found 6 times, frequency 0,00448430493273543
Char ч found 8 times, frequency 0,00597907324364723
Char Б found 1 times, frequency 0,000747384155455904
Char Н found 7 times, frequency 0,00523168908819133
Char щ found 4 times, frequency 0,00298953662182362
Char С found 3 times, frequency 0,00224215246636771
Char х found 8 times, frequency 0,00597907324364723
Char О found 3 times, frequency 0,00224215246636771
Char ц found 3 times, frequency 0,00224215246636771


Char Х found 1 times, frequency 0,000747384155455904
Char П found 1 times, frequency 0,000747384155455904
Char … found 1 times, frequency 0,000747384155455904
Char ? found 1 times, frequency 0,000747384155455904
Char Э found 1 times, frequency 0,000747384155455904
Char : found 1 times, frequency 0,000747384155455904
Average entropy of file falldance.txt: 4,83280547169775 b
Amount of information (calculated by entropy): 256,138689999981 b

Constitution.txt

```
-------Analyzing file constitution.txt
Symbols count 76
Char С found 41 times, frequency 0,002026793217657 82
Char т found 1048 times, frequency 0,05180681200257 06
Char а found 1323 times, frequency 0,06540115675515 35
Char ь found 327 times, frequency 0,01616491176034 41
Char я found 362 times, frequency 0,017895101092491
Char   found 2167 times, frequency 0,107123436650353
Char 2 found 22 times, frequency 0,001087547580206 63
Char 1 found 9 times, frequency 0,000444905828266 35
Char . found 175 times, frequency 0,008650946660734 59
Char В found 9 times, frequency 0,000444905828266 35
Char с found 924 times, frequency 0,04567699836867 86
Char е found 1461 times, frequency 0,07222304612190 42
Char л found 550 times, frequency 0,02718868950516 59
Char ю found 122 times, frequency 0,006030945672054 97
Char д found 524 times, frequency 0,02590340600128 53
Char и found 1427 times, frequency 0,07054229077067 58
Char в found 743 times, frequency 0,03672944782243 31
Char о found 1788 times, frequency 0,08838795788224 83
Char б found 282 times, frequency 0,01394038261901 23
Char н found 1374 times, frequency 0,06792228978199 61
Char ы found 314 times, frequency 0,01552227000840 38
Char р found 822 times, frequency 0,04063473231499 33
Char м found 451 times, frequency 0,02229472539423 6
Char п found 445 times, frequency 0,02199812150872 51
Char х found 201 times, frequency 0,009936230164615 16
Char П found 16 times, frequency 0,000790943694695 734
 found 306 times, frequency 0,01512679816105 59
Char
 found 306 times, frequency 0,01512679816105 59
Char ч found 182 times, frequency 0,008996984527163 97
Char к found 336 times, frequency 0,01660981758861 04
Char у found 398 times, frequency 0,01967472440555 64
Char ж found 245 times, frequency 0,01211132532502 84
Char ш found 40 times, frequency 0,001977359236739 33
Char г found 253 times, frequency 0,01250679717237 63
Char , found 197 times, frequency 0,009738494240941 22
Char з found 283 times, frequency 0,01398981659993 08
Char щ found 130 times, frequency 0,006426417519402 84
Char й found 179 times, frequency 0,008848682584408 52
Char К found 39 times, frequency 0,001927925255820 85
Char ц found 98 times, frequency 0,004844530130011 37
Char ъ found 13 times, frequency 0,000642641751940284
Char 3 found 18 times, frequency 0,000889811656532701
Char э found 19 times, frequency 0,000939245637451184
Char 4 found 19 times, frequency 0,000939245637451184
```



briefHistoryOfTime:

```
Char с found 300 times, frequency 0,0435540069686411
Char к found 167 times, frequency 0,0242450638792102
Char ы found 94 times, frequency 0,0136469221835076
Char ж found 57 times, frequency 0,00827526132404181
Char и found 383 times, frequency 0,0556039488966318
Char б found 101 times, frequency 0,0146631823461092
Char , found 111 times, frequency 0,0161149825783972
Char ч found 96 times, frequency 0,0139372822299652
Char г found 88 times, frequency 0,0127758420441347
Char у found 128 times, frequency 0,0185830429732869
Char ш found 26 times, frequency 0,0037746806039489
Char Л found 4 times, frequency 0,000580720092915215
Char 2 found 2 times, frequency 0,000290360046457607
Char З found 8 times, frequency 0,00116144018583043
Char . found 51 times, frequency 0,00740418118466899
Char Э found 5 times, frequency 0,000725900116144019
Char й found 57 times, frequency 0,00827526132404181
Char ! found 1 times, frequency 0,000145180023228804
Char э found 17 times, frequency 0,00246806039488966
Char Х found 3 times, frequency 0,000435540069686411
Char Ц found 2 times, frequency 0,000290360046457607
Char В found 19 times, frequency 0,00275842044134727
Char щ found 20 times, frequency 0,00290360046457607
Char ю found 53 times, frequency 0,0076945412311266
Char Б found 2 times, frequency 0,000290360046457607
Char : found 2 times, frequency 0,000290360046457607
Char ф found 8 times, frequency 0,00116144018583043
Char О found 5 times, frequency 0,000725900116144019
Char 1 found 6 times, frequency 0,000871080139372822
Char 5 found 1 times, frequency 0,000145180023228804
Char 4 found 1 times, frequency 0,000145180023228804
Char Н found 16 times, frequency 0,00232288037166086
Char К found 13 times, frequency 0,00188734030197445
Char ( found 5 times, frequency 0,000725900116144019
Char ъ found 4 times, frequency 0,000580720092915215
Char ) found 5 times, frequency 0,000725900116144019
Char Е found 2 times, frequency 0,000290360046457607
Char С found 6 times, frequency 0,000871080139372822
Char ц found 21 times, frequency 0,00304878048780488
Char Д found 3 times, frequency 0,000435540069686411
Char - found 7 times, frequency 0,00101626016260163
Char И found 4 times, frequency 0,000580720092915215
Char Г found 4 times, frequency 0,000580720092915215
Char Т found 1 times, frequency 0,000145180023228804
Char А found 2 times, frequency 0,000290360046457607
Char 6 found 3 times, frequency 0,000435540069686411
```

```
Char 0 found 1 times, frequency 0,000145180023228804
Char 9 found 2 times, frequency 0,000290360046457607
Char Ю found 4 times, frequency 0,000580720092915215
Char P found 3 times, frequency 0,000435540069686411
Char - found 8 times, frequency 0,00116144018583043
 found 12 times, frequency 0,00174216027874564
Char
 found 12 times, frequency 0,00174216027874564
Char Ч found 1 times, frequency 0,000145180023228804
Char a found 1 times, frequency 0,000145180023228804
Char d found 1 times, frequency 0,000145180023228804
Char h found 1 times, frequency 0,000145180023228804
Char o found 1 times, frequency 0,000145180023228804
Char c found 1 times, frequency 0,000145180023228804
Char « found 5 times, frequency 0,000725900116144019
Char » found 5 times, frequency 0,000725900116144019
Char 8 found 1 times, frequency 0,000145180023228804
Char 7 found 1 times, frequency 0,000145180023228804
Char ? found 1 times, frequency 0,000145180023228804
Average entropy of file briefHistoryOfTime.txt: 4,66658766697502 b
Amount of information (calculated by entropy): 382,660188691951 b
```

```
The directory files contains the following files:
The size of briefHistoryOfTime.txt is 12574 bytes.
The size of constitution.txt is 37152 bytes.
The size of falldance.txt is 2369 bytes.
Для продолжения нажмите любую клавишу . . .
```

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).
5. 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)

```
The directory files contains the following files:
The size of briefHistoryOfTime.tar is 14336 bytes.
The size of briefHistoryOfTime.txt is 12574 bytes.
The size of briefHistoryOfTime.txt.bz2 is 3123 bytes.
The size of briefHistoryOfTime.txt.gz is 3848 bytes.
The size of briefHistoryOfTime.txt.xz is 3660 bytes.
The size of briefHistoryOfTime.txt.zip is 3972 bytes.
The size of constitution.tar is 38912 bytes.
The size of constitution.txt is 37152 bytes.
The size of constitution.txt.bz2 is 6487 bytes.
The size of constitution.txt.gz is 8574 bytes.
The size of constitution.txt.xz is 7764 bytes.
The size of constitution.txt.zip is 8686 bytes.
The size of falldance.tar is 4096 bytes.
The size of falldance.txt is 2369 bytes.
The size of falldance.txt.bz2 is 807 bytes.
The size of falldance.txt.gz is 1007 bytes.
The size of falldance.txt.xz is 1024 bytes.
The size of falldance.txt.zip is 1113 bytes.
Для продолжения нажмите любую клавишу . . .
```

Розміри файлів

Файл	.txt	ZIP	gzip	LZMA2(.xz)	bzip2	tar	К-кість інформації (за доп. ентропії)
briefHistoryOfTime.txt	12574	3972	3848	3660	3123	14336	382,7
constitution.txt	37152	8686	8574	7764	6487	38912	359,2
falldance.txt	2369	1113	1007	1024	807	4096	4,833

2. 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)

a. 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

a. Порівняйте отримане значення з кількістю інформації вихідного файлу

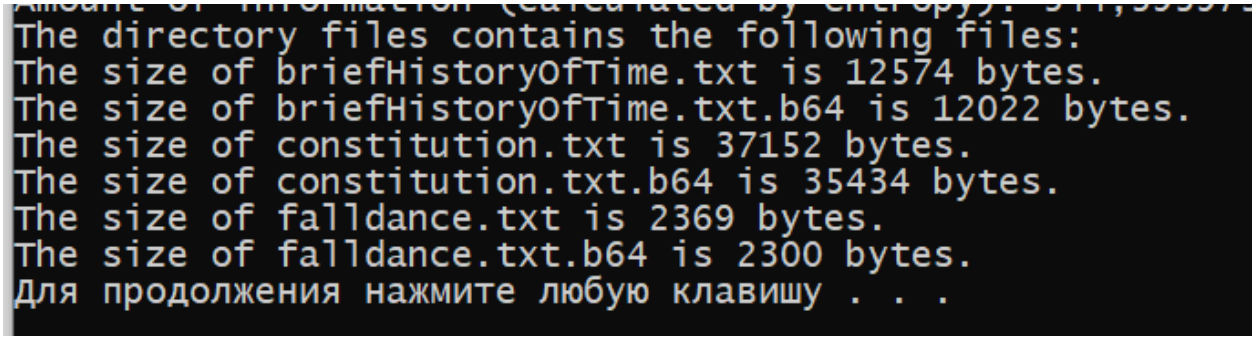
b. Зробіть висновки з отриманого результату

4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли

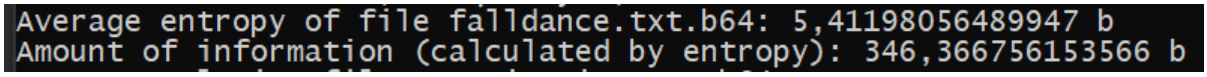
. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу

a. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу

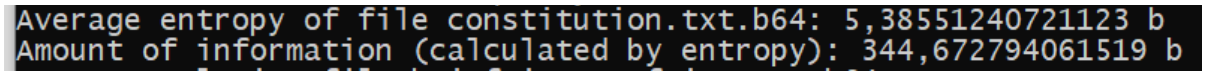
b. Зробіть висновки з отриманого результату



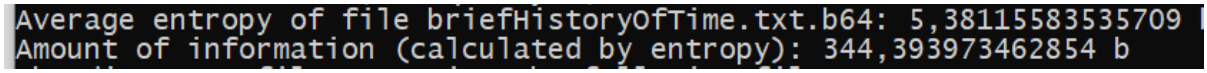
falldance.txt.b64



constitution.txt.b64



briefHistoryOfTime.txt.b64



Файл	Розмір файлу	Розмір файлу закодованого в base64	Середня ентропія файлу	Середня ентропія файлу закодованого в base64	К-кість інформації файлу	К-кість інформації файлу закодованого в base64
briefHistoryOfTime.txt	12574	12022	4.667	5.381	382.7	344.39
constitution.txt	37152	35434	4.726	5.385	359.2	344.67
falldance.txt	2639	2300	4.833	5.412	256.1	346.37

Висновок: в лабораторній роботі було вивчено алгоритми вираховування середної ентропії та кількості інформації, алгоритму кодування бінарного коду в текстовий формат BASE64.

Github: https://github.com/HleBASS/CS_labs