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
Бондаренко Владислав

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)

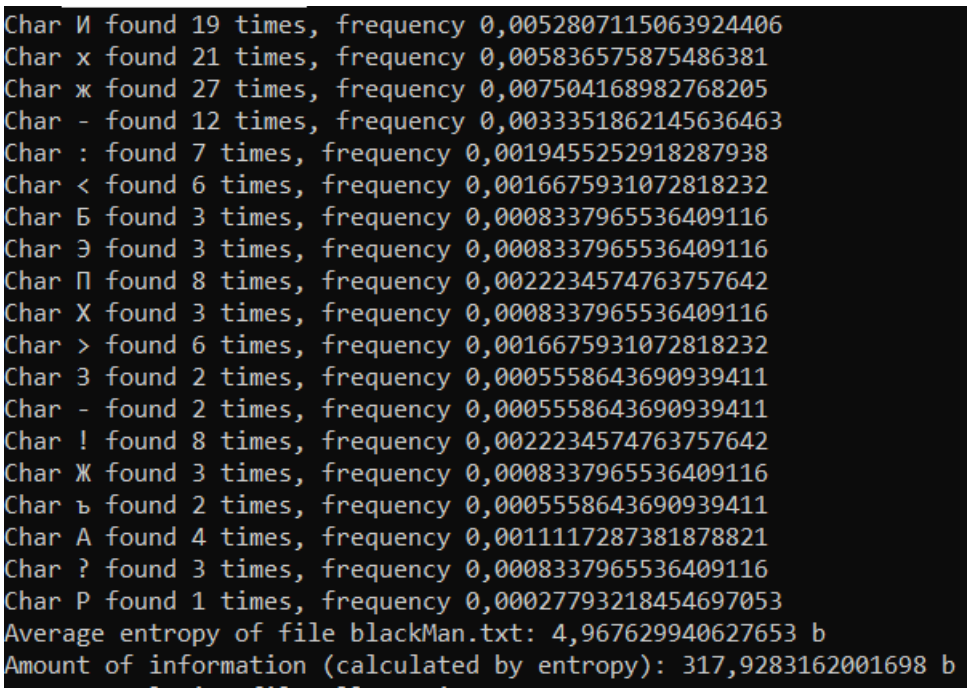
    1. **blackMan.txt** – Вірш. Автор: Сергій Есенін. Название: “Черный человек”.
    2. **allCreations.txt** – Уривок з книги. Автор: Д. Херриот. Назва: “О всех созданиях”.
    3. **trash.txt –** Стаття. Автор: Жемжуров Михаил Леонидович. Назва: “Радиотоксичность облученного ядерного топлива реактора ВВЭР-1200 в зависимости от выгорания и времени выдержки”.

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

blackMan.txt

```
-------Analyzing file blackMan.txt
Symbols count 64
Char Д found 4 times, frequency 0,0011117287381878821
Char р found 108 times, frequency 0,03001667593107282
Char у found 74 times, frequency 0,02056698165647582
Char г found 41 times, frequency 0,011395219566425792
Char   found 434 times, frequency 0,12062256809338522
Char м found 84 times, frequency 0,023346303501945526
Char о found 273 times, frequency 0,07587548638132295
Char й found 64 times, frequency 0,017787659811006114
Char , found 83 times, frequency 0,023068371317398556
Char д found 60 times, frequency 0,016675931072818232
 found 181 times, frequency 0,05030572540300167
Char
 found 181 times, frequency 0,05030572540300167
Char Я found 7 times, frequency 0,0019455252918287938
Char ч found 40 times, frequency 0,011117287381878822
Char е found 207 times, frequency 0,0575319622012229
Char н found 158 times, frequency 0,04391328515842134
Char ь found 54 times, frequency 0,01500833796553641
Char и found 139 times, frequency 0,038632573652028906
Char б found 35 times, frequency 0,009727626459143969
Char л found 120 times, frequency 0,033351862145636464
Char . found 93 times, frequency 0,02584769316286826
Char С found 21 times, frequency 0,005836575875486381
Char а found 163 times, frequency 0,0453029460811562
Char з found 38 times, frequency 0,010561423012784881
Char ю found 29 times, frequency 0,008060033351862146
Char т found 154 times, frequency 0,042801556420233464
Char к found 97 times, frequency 0,026959421901056144
Char в found 103 times, frequency 0,028627015008337964
Char я found 48 times, frequency 0,013340744858254585
Char с found 121 times, frequency 0,033629794330183434
Char э found 8 times, frequency 0,0022234574763757642
Char Т found 11 times, frequency 0,003057254030016676
Char Н found 16 times, frequency 0,0044469149527515284
Char п found 42 times, frequency 0,011673151750972763
Char ы found 65 times, frequency 0,018065591995553083
Char щ found 8 times, frequency 0,0022234574763757642
Char О found 3 times, frequency 0,0008337965536409116
Char Г found 6 times, frequency 0,0016675931072818232
Char ш found 28 times, frequency 0,007782101167315175
Char К found 8 times, frequency 0,0022234574763757642
Char ц found 11 times, frequency 0,003057254030016676
Char Е found 2 times, frequency 0,0005558643690939411
Char М found 5 times, frequency 0,0013896609227348527
Char Ч found 17 times, frequency 0,004724847137298499
Char В found 14 times, frequency 0,0038910505836575876
```



allCreations.txt

```
-------Analyzing file allCreations.txt
Symbols count 79
Char б found 257 times, frequency 0,012220637184973846
Char ъ found 5 times, frequency 0,00023775558725630053
Char е found 1403 times, frequency 0,06671421778411793
Char з found 317 times, frequency 0,015073704232049453
Char д found 506 times, frequency 0,024060865430337613
Char а found 1229 times, frequency 0,05844032334759867
Char   found 3247 times, frequency 0,15439847836424156
Char с found 802 times, frequency 0,038135996195910606
Char л found 820 times, frequency 0,038991916310033285
Char о found 1813 times, frequency 0,08621017593913458
Char в found 685 times, frequency 0,032572515454113174
Char т found 908 times, frequency 0,04317641464574418
Char р found 749 times, frequency 0,03561578697099382
Char п found 419 times, frequency 0,019923918212077982
Char и found 1032 times, frequency 0,04907275320970043
Char ь found 277 times, frequency 0,013171659533999049
Char я found 361 times, frequency 0,0171659533999049
Char м found 520 times, frequency 0,024726581074655255
Char н found 1073 times, frequency 0,05102234902520209
Char у found 494 times, frequency 0,023490252020922493
Char ю found 84 times, frequency 0,003994293865905849
Char , found 367 times, frequency 0,017451260104612457
Char к found 561 times, frequency 0,02667617689015692
Char ш found 123 times, frequency 0,005848787446504993
Char ч found 244 times, frequency 0,011602472658107465
Char щ found 71 times, frequency 0,0033761293390394674
Char ж found 162 times, frequency 0,007703281027104137
Char ы found 310 times, frequency 0,014740846409890632
Char ! found 40 times, frequency 0,0019020446980504042
Char И found 26 times, frequency 0,0012363290537327628
Char г found 290 times, frequency 0,01378982406086543
Char й found 179 times, frequency 0,008511650023775559
Char . found 239 times, frequency 0,011364717070851165
 found 250 times, frequency 0,011887779362815026
Char
 found 250 times, frequency 0,011887779362815026
Char О found 21 times, frequency 0,0009985734664764623
Char х found 125 times, frequency 0,005943889681407513
Char Д found 32 times, frequency 0,0015216357584403233
Char В found 36 times, frequency 0,0017118402282453639
Char ц found 58 times, frequency 0,002757964812173086
Char - found 150 times, frequency 0,007132667617689016
Char Й found 2 times, frequency 9,51022349025202E-05
```

```
Char Я found 15 times, frequency 0,0007132667617689016
Char э found 35 times, frequency 0,0016642891107941037
Char : found 19 times, frequency 0,000903471231573942
Char Т found 26 times, frequency 0,0012363290537327628
Char Л found 2 times, frequency 9,51022349025202E-05
Char - found 34 times, frequency 0,0016167379933428436
Char Х found 5 times, frequency 0,00023775558725630053
Char ; found 2 times, frequency 9,51022349025202E-05
Char С found 8 times, frequency 0,0003804089396100808
Char М found 24 times, frequency 0,0011412268188302425
Char З found 37 times, frequency 0,0017593913456966238
Char ф found 55 times, frequency 0,002615311459819306
Char ? found 30 times, frequency 0,0014265335235378032
Char Е found 9 times, frequency 0,00042796005706134097
Char А found 19 times, frequency 0,000903471231573942
Char К found 30 times, frequency 0,0014265335235378032
Char Х found 1 times, frequency 4,75511174512601E-05
Char V found 1 times, frequency 4,75511174512601E-05
Char І found 3 times, frequency 0,0001426533523537803
Char Б found 7 times, frequency 0,0003328578221588207
Char Н found 46 times, frequency 0,0021873514027579647
Char П found 17 times, frequency 0,0008083689966714218
Char : found 17 times, frequency 0,0008083689966714218
Char Р found 3 times, frequency 0,0001426533523537803
Char У found 7 times, frequency 0,0003328578221588207
Char Ф found 5 times, frequency 0,00023775558725630053
Char ' found 16 times, frequency 0,0007608178792201616
Char ( found 1 times, frequency 4,75511174512601E-05
Char 1 found 1 times, frequency 4,75511174512601E-05
Char 9 found 1 times, frequency 4,75511174512601E-05
Char 4 found 1 times, frequency 4,75511174512601E-05
Char 8 found 1 times, frequency 4,75511174512601E-05
Char ) found 1 times, frequency 4,75511174512601E-05
Char Г found 4 times, frequency 0,0001902044698050404
Char Ч found 5 times, frequency 0,00023775558725630053
Char Э found 4 times, frequency 0,0001902044698050404
Char 2 found 1 times, frequency 4,75511174512601E-05
Average entropy of file allCreations.txt: 4,750810283577105 b
Amount of information (calculated by entropy): 375,31401240259135 b
```

trash.txt

```
-------Analyzing file trash.txt
Symbols count 118
Char A found 4 times, frequency 0,00015124016938898973
Char н found 618 times, frequency 0,02336660617059891
Char о found 970 times, frequency 0,036675741076830005
Char т found 651 times, frequency 0,024614337568058076
Char а found 564 times, frequency 0,02132486388384755
Char ц found 68 times, frequency 0,002571082879612825
Char и found 828 times, frequency 0,03130671506352087
Char я found 212 times, frequency 0,008015728977616455
Char : found 7 times, frequency 0,000264670296430732
Char   found 1518 times, frequency 0,0573956442831216
Char Р found 32 times, frequency 0,0012099213551119178
Char с found 343 times, frequency 0,012968844525105869
Char ч found 109 times, frequency 0,00412129461584997
Char р found 413 times, frequency 0,015615547489413188
Char д found 345 times, frequency 0,013044464609800363
Char к found 264 times, frequency 0,009981851179673321
Char ь found 87 times, frequency 0,003289473684210526
Char б found 99 times, frequency 0,0037431941923774955
Char л found 338 times, frequency 0,012779794313369631
Char у found 181 times, frequency 0,0068436176648517845
Char е found 587 times, frequency 0,02219449485783424
Char г found 115 times, frequency 0,004348154869933454
Char п found 213 times, frequency 0,008053539019963703
Char в found 356 times, frequency 0,013460375075620085
Char В found 79 times, frequency 0,002986993345432547
Char Э found 21 times, frequency 0,000794010889292196
Char - found 124 times, frequency 0,004688445251058681
Char 1 found 838 times, frequency 0,03168481548699335
Char 2 found 596 times, frequency 0,02253478523895947
Char 0 found 1784 times, frequency 0,06745311554748941
Char ы found 248 times, frequency 0,009376890502117362
Char х found 101 times, frequency 0,00381881427707199
Char й found 79 times, frequency 0,002986993345432547
Char м found 202 times, frequency 0,00763762855414398
Char . found 1132 times, frequency 0,042800967937084086
Char ю found 38 times, frequency 0,0014367816091954023
Char ж found 54 times, frequency 0,002041742286751361
Char щ found 32 times, frequency 0,0012099213551119178
 found 2971 times, frequency 0,11233363581367212
Char
 found 2971 times, frequency 0,11233363581367212
Char A found 43 times, frequency 0,0016258318209316394
Char b found 5 times, frequency 0,00018905021173623715
```

```
Char s found 15 times, frequency 0,0005671506352087115
Char t found 32 times, frequency 0,0012099213551119178
Char r found 26 times, frequency 0,0009830611010284331
Char a found 34 times, frequency 0,0012855414398064125
Char c found 14 times, frequency 0,000529340592861464
Char T found 1 times, frequency 3,781004234724743E-05
Char h found 10 times, frequency 0,0003781004234724743
Char e found 40 times, frequency 0,0015124016938898972
Char d found 21 times, frequency 0,000794010889292196
Char i found 32 times, frequency 0,0012099213551119178
Char o found 27 times, frequency 0,0010208711433756805
Char x found 8 times, frequency 0,00030248033877797946
Char y found 11 times, frequency 0,00041591046581972174
Char f found 10 times, frequency 0,0003781004234724743
Char n found 20 times, frequency 0,0007562008469449486
Char u found 67 times, frequency 0,0025332728372655777
Char l found 10 times, frequency 0,0003781004234724743
Char V found 2 times, frequency 7,562008469449486E-05
Char E found 923 times, frequency 0,034898669086509376
Char R found 2 times, frequency 7,562008469449486E-05
Char w found 3 times, frequency 0,00011343012704174228
Char g found 6 times, frequency 0,00022686025408348456
Char p found 27 times, frequency 0,0010208711433756805
Char m found 93 times, frequency 0,003516333938294011
Char , found 104 times, frequency 0,003932244404113733
Char z found 1 times, frequency 3,781004234724743E-05
Char v found 2 times, frequency 7,562008469449486E-05
Char k found 1 times, frequency 3,781004234724743E-05
Char K found 8 times, frequency 0,00030248033877797946
Char э found 139 times, frequency 0,005255595886267393
Char ; found 12 times, frequency 0,0004537205081669691З
Char К found 1 times, frequency 3,781004234724743E-05
Char У found 2 times, frequency 7,562008469449486E-05
Char Д found 8 times, frequency 0,00030248033877797946
Char 5 found 486 times, frequency 0,01837568058076225
Char 9 found 307 times, frequency 0,011607683000604961
Char 6 found 484 times, frequency 0,018300060496067756
Char 3 found 439 times, frequency 0,01659860859044162
Char 7 found 463 times, frequency 0,01750604960677556
Char ш found 21 times, frequency 0,000794010889292196
Char ( found 32 times, frequency 0,0012099213551119178
Char ) found 32 times, frequency 0,0012099213551119178
Char э found 20 times, frequency 0,0007562008469449486
Char T found 38 times, frequency 0,0014367816091954023
Char ф found 22 times, frequency 0,0008318209316394435
```

```
Char П found 11 times, frequency 0,00041591046581972174
Char М found 3 times, frequency 0,00011343012704174228
Char И found 4 times, frequency 0,00015124016938898973
Char С found 3 times, frequency 0,00011343012704174228
Char Н found 12 times, frequency 0,00045372050816696913
Char М found 1 times, frequency 3,781004234724743E-05
Char С found 38 times, frequency 0,0014367816091954023
Char У found 24 times, frequency 0,00090744101633393830
Char Р found 55 times, frequency 0,00207955232909860870
Char D found 3 times, frequency 0,00011343012704174228
Char [ found 3 times, frequency 0,00011343012704174228
Char ] found 3 times, frequency 0,00011343012704174228
Char Б found 11 times, frequency 0,00041591046581972174
Char / found 30 times, frequency 0,001134301270417423
Char 8 found 433 times, frequency 0,016371748336358138
Char 4 found 504 times, frequency 0,019056261343012703
Char N found 18 times, frequency 0,00068058076225045370
Char О found 31 times, frequency 0,00117211131276467030
Char Я found 23 times, frequency 0,00086963097398669080
Char Г found 21 times, frequency 0,00079401088929219600
Char · found 19 times, frequency 0,00071839080459770110
Char + found 855 times, frequency 0,032327586206896554
Char З found 12 times, frequency 0,00045372050816696913
Char = found 2 times, frequency 7,562008469449486E-05
Char - found 12 times, frequency 0,00045372050816696913
Char ь found 3 times, frequency 0,00011343012704174228
Char L found 1 times, frequency 3,781004234724743E-05
Char Z found 2 times, frequency 7,562008469449486E-05
Char j found 3 times, frequency 0,00011343012704174228
Char % found 10 times, frequency 0,00037810042347247430
Char · found 2 times, frequency 7,562008469449486E-05
Average entropy of file trash.txt: 5,124651574492334 b
Amount of information (calculated by entropy): 604,7088857900953 b
```

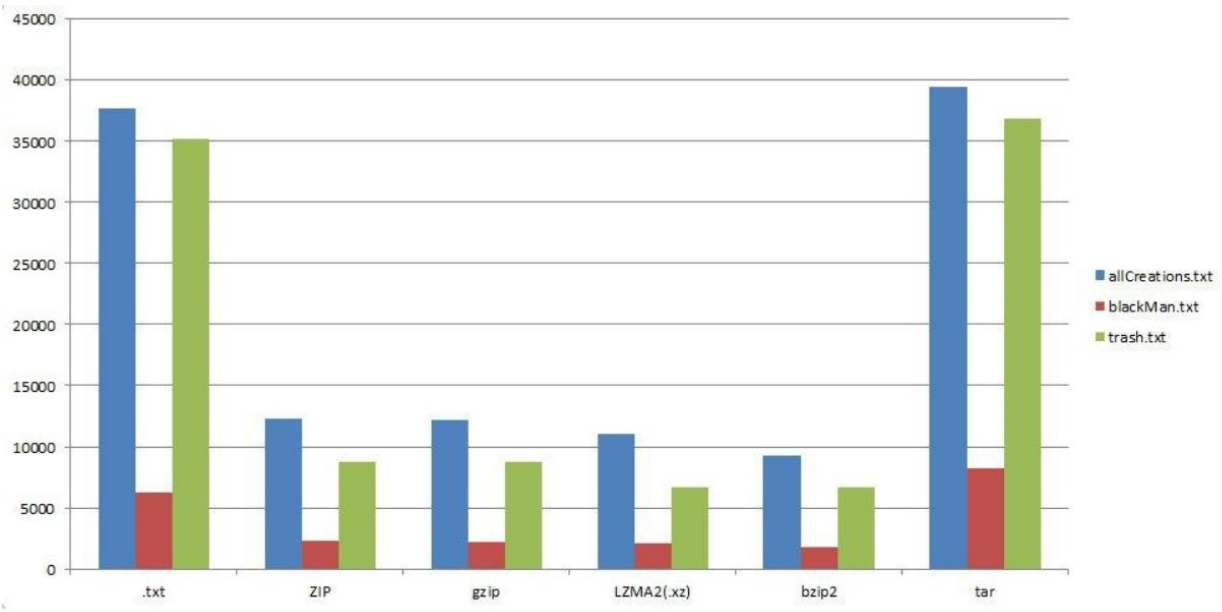
```
The directory assets contains the following files:
The size of allCreations.txt is 37721 bytes.
The size of blackMan.txt is 6210 bytes.
The size of trash.txt is 35155 bytes.
```

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 assets contains the following files:
The size of allCreations.tar is 39424 bytes.
The size of allCreations.txt is 37721 bytes.
The size of allCreations.txt.bz2 is 9225 bytes.
The size of allCreations.txt.gz is 12180 bytes.
The size of allCreations.txt.xz is 11056 bytes.
The size of allCreations.txt.zip is 12292 bytes.
The size of blackMan.tar is 8192 bytes.
The size of blackMan.txt is 6210 bytes.
The size of blackMan.txt.bz2 is 1829 bytes.
The size of blackMan.txt.gz is 2240 bytes.
The size of blackMan.txt.xz is 2144 bytes.
The size of blackMan.txt.zip is 2344 bytes.
The size of trash.tar is 36864 bytes.
The size of trash.txt is 35155 bytes.
The size of trash.txt.bz2 is 6635 bytes.
The size of trash.txt.gz is 8704 bytes.
The size of trash.txt.xz is 6904 bytes.
The size of trash.txt.zip is 8802 bytes.
```

Розміри файлів

Файл	.txt	ZIP	gzip	LZMA2(.xz)	bzip2	tar	К-кість інформації (за доп. ентропії)
allCreations.txt	37721	12292	12180	11056	9225	39424	375.3
blackMan.txt	6210	2344	2240	2144	1829	8192	317.9
trash.txt	35155	8802	8704	6635	6635	36864	604.7



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

```
The directory assets contains the following files:
The size of allCreations.txt is 37721 bytes.
The size of allCreations.txt.b64 is 36404 bytes.
The size of blackMan.txt is 6210 bytes.
The size of blackMan.txt.b64 is 6216 bytes.
The size of trash.txt is 35155 bytes.
The size of trash.txt.b64 is 39734 bytes.
```

blackMan.txt.b64

```
Average entropy of file blackMan.txt.b64: 5,504442879518631 b
Amount of information (calculated by entropy): 368,7976729277483 b
```

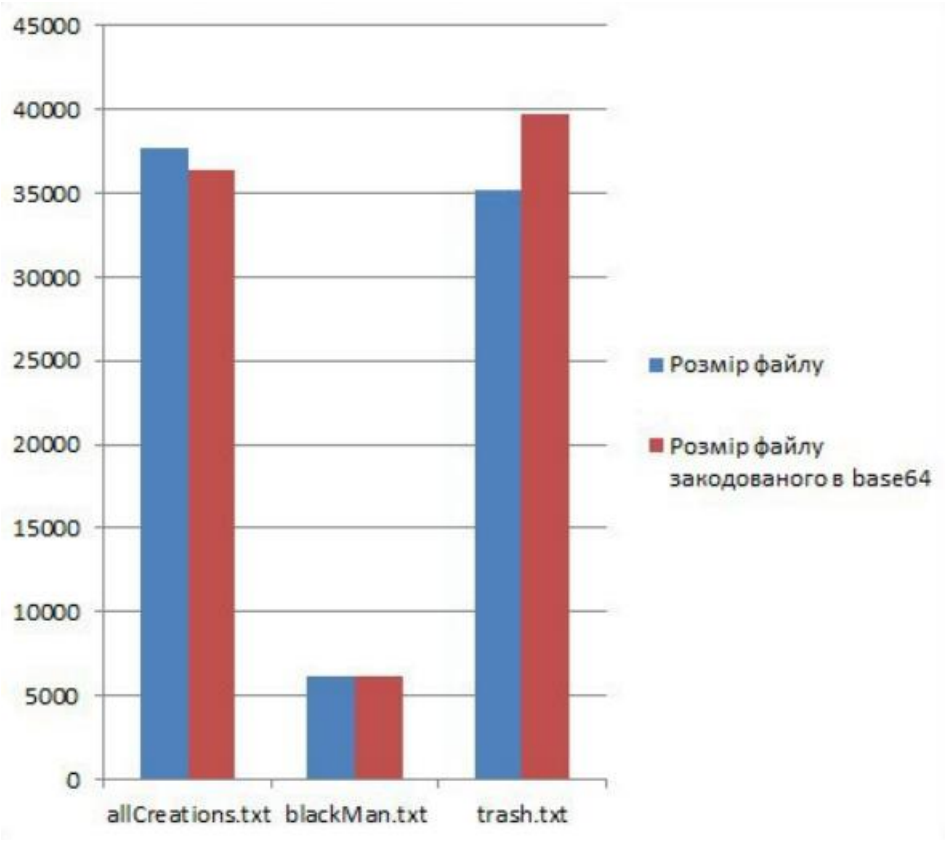
allCreations.txt.b64

```
Average entropy of file allCreations.txt.b64: 5,43430540188516 b
Amount of information (calculated by entropy): 364,09846192630573 b
```

trash.txt.b64

```
Average entropy of file trash.txt.b64: 5,3716287145789945 b
Amount of information (calculated by entropy): 359,89912387679266 b
```

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
allCreations.txt	37721	36404	4.75	5.43	375.3	364
blackMan.txt	6210	6216	4.97	5.5	318	369
trash.txt	35155	39734	5.12	5.37	604.7	360



Висновок: В даній лабораторній роботі було навчено обрахуванню ентропії, написанню алгоритму для обрахування ентропії, написання алгоритму для кодування бінарного коду в текстовий формат base64, який широко використовується у Всесвітній паутині, наприклад, для вставки зображень і інших двійкових ресурсів в HTML і CSS.

Github: https://github.com/Lrazerz/CompSystemsLabs