

**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 репку” Леся Подерв’янського та специфікацію інтерфейсу 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

```
-----Analyzing file falldance.txt
Symbols count 53
Char B found 3 times, frequency 0,00224215246636771
Char o found 103 times, frequency 0,0769805680119581
Char л found 50 times, frequency 0,0373692077727952
Char н found 62 times, frequency 0,0463378176382661
Char в found 51 times, frequency 0,0381165919282511
Char а found 94 times, frequency 0,070254110612855
Char т found 64 times, frequency 0,0478325859491779
Char ь found 19 times, frequency 0,0142002989536622
Char  found 161 times, frequency 0,120328849028401
Char м found 26 times, frequency 0,0194319880418535
Char е found 79 times, frequency 0,0590433482810164
Char я found 27 times, frequency 0,0201793721973094
Char с found 43 times, frequency 0,0321375186846039
Char и found 46 times, frequency 0,0343796711509716
Char – found 5 times, frequency 0,00373692077727952
Char  found 54 times, frequency 0,0403587443946188
Char  found 54 times, frequency 0,0403587443946188
Char э found 1 times, frequency 0,000747384155455904
Char й found 16 times, frequency 0,0119581464872945
Char , found 32 times, frequency 0,0239162929745889
```

```

Char , found 32 times, frequency 0,0239162929745889
Char P 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
1. Char p found 41 times, frequency 0,0306427503736921
2. Char г found 22 times, frequency 0,0164424514200299
3. Char и found 9 times, frequency 0,00672645739910314
4. Char у found 37 times, frequency 0,0276532137518685
5. Char з found 25 times, frequency 0,0186846038863976
6. 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 3 found 4 times, frequency 0,00298953662182362
Char n found 24 times, frequency 0,0179372197309417
Char T 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 H found 7 times, frequency 0,00523168908819133
Char щ found 4 times, frequency 0,00298953662182362
Char C found 3 times, frequency 0,00224215246636771
Char x found 8 times, frequency 0,00597907324364723
Char O found 3 times, frequency 0,00224215246636771
Char ц found 3 times, frequency 0,00224215246636771

```

```

Char X 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

```
Amount of information (calculated by entropy): 256,1566
-----Analyzing file constitution.txt
Symbols count 76
Char C found 41 times, frequency 0,00202679321765782
Char T found 1048 times, frequency 0,0518068120025706
Char A found 1323 times, frequency 0,0654011567551535
Char Ъ found 327 times, frequency 0,0161649117603441
Char Я found 362 times, frequency 0,017895101092491
Char  found 2167 times, frequency 0,107123436650353
Char 2 found 22 times, frequency 0,00108754758020663
Char 1 found 9 times, frequency 0,00044490582826635
Char . found 175 times, frequency 0,00865094666073459
Char B found 9 times, frequency 0,00044490582826635
Char C found 924 times, frequency 0,0456769983686786
Char E found 1461 times, frequency 0,0722230461219042
Char Л found 550 times, frequency 0,0271886895051659
Char Ю found 122 times, frequency 0,00603094567205497
Char Д found 524 times, frequency 0,0259034060012853
Char И found 1427 times, frequency 0,0705422907706758
Char В found 743 times, frequency 0,0367294478224331
Char О found 1788 times, frequency 0,0883879578822483
Char 6 found 282 times, frequency 0,0139403826190123
Char Н found 1374 times, frequency 0,0679222897819961
Char Ы found 314 times, frequency 0,0155222700084038
Char Р found 822 times, frequency 0,0406347323149933
Char М found 451 times, frequency 0,022294725394236
Char П found 445 times, frequency 0,0219981215087251
Char X found 201 times, frequency 0,00993623016461516
Char П found 16 times, frequency 0,000790943694695734
Char  found 306 times, frequency 0,0151267981610559
Char
Char  found 306 times, frequency 0,0151267981610559
Char Ч found 182 times, frequency 0,00899698452716397
Char К found 336 times, frequency 0,0166098175886104
Char У found 398 times, frequency 0,0196747244055564
Char Ж found 245 times, frequency 0,0121113253250284
Char Ш found 40 times, frequency 0,00197735923673933
Char Г found 253 times, frequency 0,0125067971723763
Char , found 197 times, frequency 0,00973849424094122
Char З found 283 times, frequency 0,0139898165999308
Char Щ found 130 times, frequency 0,00642641751940284
Char Й found 179 times, frequency 0,00884868258440852
Char К found 39 times, frequency 0,00192792525582085
Char Ц found 98 times, frequency 0,00484453013001137
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
```



```

Char 4 found 19 times, frequency 0,000939245637451184
Char Г found 24 times, frequency 0,0011864155420436
Char H found 17 times, frequency 0,000840377675614217
Char P found 7 times, frequency 0,000346037866429384
Char : found 1 times, frequency 4,94339809184834E-05
Char - found 13 times, frequency 0,000642641751940284
Char ф found 41 times, frequency 0,00202679321765782
Char ; found 7 times, frequency 0,000346037866429384
Char 5 found 10 times, frequency 0,000494339809184834
Char у found 19 times, frequency 0,000939245637451184
Char 6 found 4 times, frequency 0,000197735923673933
Char И found 6 times, frequency 0,0002966038855109
Char - found 5 times, frequency 0,000247169904592417
Char 7 found 5 times, frequency 0,000247169904592417
Char 0 found 12 times, frequency 0,0005932077710218
Char 8 found 4 times, frequency 0,000197735923673933
Char 9 found 12 times, frequency 0,0005932077710218
Char 3 found 3 times, frequency 0,00014830194275545
Char 0 found 22 times, frequency 0,00108754758020663
Char Э found 3 times, frequency 0,00014830194275545
Char Ц found 1 times, frequency 4,94339809184834E-05
Char ( found 2 times, frequency 9,88679618369667E-05
Char ) found 2 times, frequency 9,88679618369667E-05
Char Ч found 1 times, frequency 4,94339809184834E-05
Char { found 4 times, frequency 0,000197735923673933
Char № found 3 times, frequency 0,00014830194275545
Char / found 4 times, frequency 0,000197735923673933
Char } found 4 times, frequency 0,000197735923673933
Char N found 1 times, frequency 4,94339809184834E-05
Char M found 1 times, frequency 4,94339809184834E-05
Char T found 1 times, frequency 4,94339809184834E-05
Char Ъ found 1 times, frequency 4,94339809184834E-05
Char Д found 1 times, frequency 4,94339809184834E-05
Average entropy of file constitution.txt: 4,72634766015639 b
Amount of information (calculated by entropy): 359,202422171885 b

```

briefHistoryOfTime:

```

-----Analyzing file briefHistoryOfTime.txt
Symbols count 82
Char М found 3 times, frequency 0,000435540069686411
Char o found 618 times, frequency 0,0897212543554007
Char д found 155 times, frequency 0,0225029036004646
Char e found 546 times, frequency 0,0792682926829268
Char л found 291 times, frequency 0,0422473867595819
Char ь found 109 times, frequency 0,0158246225319396
Char  found 978 times, frequency 0,14198606271777
Char П found 10 times, frequency 0,00145180023228804
Char т found 399 times, frequency 0,0579268292682927
Char м found 168 times, frequency 0,024390243902439
Char я found 140 times, frequency 0,0203252032520325
Char п found 155 times, frequency 0,0225029036004646
Char з found 101 times, frequency 0,0146631823461092
Char в found 200 times, frequency 0,0290360046457607
Char а found 389 times, frequency 0,0564750290360046
Char н found 392 times, frequency 0,0569105691056911
Char x found 30 times, frequency 0,00435540069686411
Char р found 229 times, frequency 0,033246225319396

```

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 6 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 3 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 ь 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
Char found 12 times, frequency 0,00174216027874564
Char found 12 times, frequency 0,00174216027874564
Char ч found 1 times, frequency 0,000145180023228804
Char а 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 продовження натисніть будь-яку клавішу . . .

```

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 продовження натисніть будь-яку клавішу . . .

```

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

```
Amount of information (calculated by entropy): 382,7
The directory files contains the following files:
The size of briefHistoryOfTime.txt is 12574 bytes.
The size of briefHistoryOfTime.txt.b64 is 12022 bytes.
The size of constitution.txt is 37152 bytes.
The size of constitution.txt.b64 is 35434 bytes.
The size of falldance.txt is 2369 bytes.
The size of falldance.txt.b64 is 2300 bytes.
Для продовження натисніть будь-яку клавішу . . .
```

falldance.txt.b64

```
Average entropy of file falldance.txt.b64: 5,41198056489947 b  
Amount of information (calculated by entropy): 346,366756153566 b
```

constitution.txt.b64

```
Average entropy of file constitution.txt.b64: 5,38551240721123 b  
Amount of information (calculated by entropy): 344,672794061519 b
```

briefHistoryOfTime.txt.b64

```
Average entropy of file briefHistoryOfTime.txt.b64: 5,38115583535709 b  
Amount of information (calculated by entropy): 344,393973462854 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
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 середньої ентропії та кількості інформації, алгоритму кодування бінарного коду в текстовий формат BASE64.

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