

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

Комп'ютерні системи

Лабораторна робота № 1

Тема: «Дослідження кількості інформації при різних варіантах
кодування»

Роботу виконав
студент 3 курсу
спеціальності “КІ-СА”
Ситниченко Денис Вікторович

Київ 2021

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

- a) 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)
 - [І МЕРТВИМ, І ЖИВИМ, І НЕНАРОЖДЕННИМ](#)
 - [Жаба й Віл](#)
 - [Радіотехніка](#)
- b) 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

Код розташований за посиланням:

<https://gl.vlabs.knu.ua/frecs/ce/cs/2020-2021/DenysS/lab1>

Текст №1(І МЕРТВИМ, І ЖИВИМ, І НЕНАРОЖДЕНИМ)

Name of file:
text1

=====

І смеркає, і світає,
День божий минає,
І знову люд потомлений
І все спочиває.
Тільки я, мов окаянный,
І день і ніч плачу
На розпуттях велелюдних,
І ніхто не бачить,
І не бачить, і не знає –
Оглухли, не чують;
Кайданами міняються,
Правдою торгують.
І господа зневажать, –
Людей запрягають
В тяжкі ярма. Орють лихо,
Лихом засівають.;:
А що вродить? Побачите,
Які будуть жнива!
Схаменіться, недолюди,
Діти юродивії
Подивіться на рай тихий,
На свою країну!
Полюбіте щирим серцем
Велику руїну!
Розкуйтеся, братайтеся!
У чужому краю
Не шукайте, не питайте
Того, що немає
І на небі, а не тільки
На чужому полі.
В своїй хаті своя й правда,
І сила, і воля.

Number of letters in text: 649
Entropy (bits): 5,0058
Ammount of information (bits): 3248,7735
Ammount of information (bytes): 406,0967

Filesize: 1360 bytes

Letter	Frequency
І	0,0138674884437596
	0,121725731895223
с	0,024653312788906
м	0,0215716486902928
е	0,0446841294298921
р	0,0261941448382126
к	0,0169491525423729
а	0,0647149460708783
є	0,00924499229583975
,	0,0338983050847458
і	0,0323574730354391
в	0,0277349768875193
т	0,049306625577812
	0,049306625577812
д	0,00308166409861325
н	0,0446841294298921
ь	0,0261941448382126
б	0,012326656394453
о	0,0508474576271186
ж	0,00924499229583975
и	0,0431432973805855
й	0,0200308166409861
з	0,0107858243451464
у	0,0277349768875193
л	0,024653312788906
ю	0,024653312788906
д	0,0215716486902928
п	0,0138674884437596
ч	0,0138674884437596
.	0,00924499229583975
Т	0,00308166409861325
я	0,0215716486902928
Н	0,0061633281972265
х	0,0138674884437596
—	0,00308166409861325
О	0,00308166409861325
г	0,00770416024653313
;	0,00308166409861325
К	0,00154083204930663
П	0,0061633281972265
Л	0,00308166409861325
В	0,00462249614791988
:	0,00154083204930663
А	0,00154083204930663
щ	0,00462249614791988
?	0,00154083204930663
Я	0,00154083204930663
!	0,0061633281972265
С	0,00154083204930663
і	0,0061633281972265
ц	0,00154083204930663
р	0,00154083204930663

Текст №2(Жаба й Віл)

```
Name of file:
text2
=====
Раз Жаба вилізла на берег подивиться
Та й трошечки на сонечку погріється.
Побачила Вола
Та й каже подрузі тихенько
(Вигадлива була!):
— Який здоровий, моя ненько!
Ну що, сестрице, як надмусь,
То й я така зроблюсь?
От будуть жаби дивуваться!
— І де вже, сестро, нам рівняться... —
Казать їй друга почала;
А та не слуха... дметься... дметься...
— Що, сестро, як тобі здається,
Побільшала хоч трохи я?
— Та ні, голубонько моя!
— Ну, а теперечки? Дивися!
— Та годі, сестро, схаменися! —
Не слуха Жаба, дметься гірш,
Все думає, що стане більш.
Та й що, дурна, собі зробила?
З натуги луснула — та й одубіла!
Такі і в світі жаби є,
Прощайте, ніде правди діти;
А по мені — найлучче жити,
Як милосердний Бог дає.

Number of letters in text: 703
Entropy (bits): 4,9621
Amount of information (bits): 3488,3589
Amount of information (bytes): 436,0449
```

```
Filesize: 1454 bytes
Letter      Frequency
Р           0,00142247510668563
а           0,0753911806543385
з           0,0113798008534851
            0,145092460881935
Ж           0,00284495021337127
б           0,0227596017069701
в           0,0170697012802276
и           0,0384068278805121
л           0,0284495021337127
і           0,027027027027027
н           0,0298719772403983
е           0,042674253200569
р           0,0284495021337127
г           0,0128022759601707
п           0,00995732574679943
о           0,0540540540540541
д           0,0312944523470839
т           0,0412517780938834
ь           0,0241820768136558
с           0,042674253200569
я           0,0241820768136558

            0,0355618776671408
Т           0,00995732574679943
й           0,015647226173542
ш           0,00568990042674253
ч           0,0113798008534851
к           0,0184921763869132
у           0,0284495021337127
.           0,0213371266002845
П           0,0042674253200569
В           0,0042674253200569
ж           0,00711237553342817
х           0,0085348506401138
(           0,00142247510668563
!           0,00995732574679943
)           0,00142247510668563
:           0,00142247510668563
—           0,0128022759601707
Я           0,00284495021337127
,           0,0298719772403983
м           0,015647226173542
Н           0,0042674253200569
щ           0,00568990042674253
ц           0,00142247510668563
ю           0,00142247510668563
?           0,00568990042674253
О           0,00142247510668563
І           0,00142247510668563
К           0,00142247510668563
ї           0,00142247510668563
;           0,00284495021337127
А           0,00284495021337127
Щ           0,00142247510668563
є           0,00568990042674253
Д           0,00142247510668563
```

Текст №3(Радіотехніка)

Name of file:

text3

=====

Освоєнню ультракороткохвильової апаратури сприяла поява нових областей радіотехніки: дальнього радіорелейного зв'язку та радіолокації. Теоретичною базою для проектування радіорелейної приймальної і передавальної апаратури слугували наукові досягнення в області багатоканального телефонного ущільнення ліній провідного зв'язку, оскільки техніка ущільнених передач по проводах досягла до цього часу широкого розвитку.

Радіолокація (в тому числі і локаційний прийом) стала можливою з розвитком техніки надвисоких частот. Проте цього було недостатньо. Потрібна була глибока розробка ще однієї галузі науки, яка отримала назву «імпульсна техніка» і яка базується на вивченні перехідних процесів в колах радіоапаратури. Інтереси радіолокації зажадали розширення знань в області антен надвисоких частот.

У повоєнний час радіотехніка починає розвиватися прискореними темпами. Узагальнення наукових і практичних досягнень привело до того, що ці досягнення вже не охоплювалися старим поняттям «радіотехніка» – довелося говорити про надзвичайно обширну галузь науки – радіоелектроніку, в яку з кожним роком входили і входять нові галузі знань і застосувань. Окремі галузі радіоелектроніки, такі як радіозв'язок, радіомовлення, радіолокація, телебачення, радіонавігація, телеуправління, радіоастрономія та ін. пред'являють свої специфічні вимоги до радіоприймальних пристроїв. У кожній з цих областей використовується не один, а кілька діапазонів радіохвиль, що вносить ще більшу різноманітність на шляху вдосконалення техніки радіо.

Повоєнний розвиток радіоелектроніки дозволив поставити і вирішити задачу радіоприйому без пошуку кореспондента і без підстроювання на його частоту. Науковими передумовами для цього стали досягнення у справі стабілізації частоти автогенераторів та автоматичного регулювання. Висока стабільність частоти забезпечується коливаннями кварцових еталонів. Принцип автоматичного підстроювання діапазону генератора по гармоніках еталонної частоти дозволяє отримати широку сітку стабільних частот при наявності одного кварцового еталона. Проблема стабілізації частоти виявилася однією з тих проблем, які стимулювали розробку теорії автоматичного регулювання, що знайшла в даний час застосування в різноманітних галузях механіки, енергетики і радіоелектроніки.

Стабільна радіолінія дозволяє порівняно простими способами здійснити синхронний радіотелеграфний, радіотелефонний прийом, на одній бічній смузі частот. В останньому випадку передавач випромінює тільки корисну потужність

Number of letters in text: 3605

Entropy (bits): 4,6973

Amount of information (bits): 16933,8581

Amount of information (bytes): 2116,7323







Filesize: 7234 bytes







Letter	Frequency
О	0,000554785020804438
с	0,0305131761442441
в	0,0366158113730929
о	0,0940360610263523
є	0,00416088765603329
н	0,0690707350901526
ю	0,00527045769764216
	0,118169209431345
у	0,0238557558945908
л	0,0332871012482663
ь	0,0169209431345354
т	0,0477115117891817
р	0,0435506241331484
а	0,074618585298197
к	0,0260748959778086
х	0,011373092926491
и	0,0454923717059639
ї	0,00471567267683773
п	0,0249653259361997
я	0,0208044382801664
б	0,0124826629680999
е	0,0416088765603329
й	0,0124826629680999
д	0,0282940360610264
і	0,0529819694868239
:	0,000277392510402219
г	0,0141470180305132
з	0,0185852981969487
'	0,00166435506241331
ц	0,00610263522884882
.	0,0072122052704577
Т	0,000277392510402219
ч	0,0135922330097087
м	0,0199722607489598
ф	0,00221914008321775
щ	0,00249653259361997
,	0,00748959778085992
ш	0,00527045769764216







0,00360610263522885

Р	0,000832177531206657
(0,0013869625520111
)	0,0013869625520111
ж	0,00305131761442441
П	0,00221914008321775
«	0,000554785020804438
»	0,000554785020804438
І	0,000277392510402219
У	0,000832177531206657
—	0,000832177531206657
Н	0,000554785020804438
В	0,000832177531206657
С	0,00110957004160888
К	0,000832177531206657
1	0,000277392510402219
6	0,000277392510402219
Ц	0,000277392510402219
З	0,000277392510402219
Ш	0,000554785020804438
А	0,000277392510402219
М	0,000277392510402219

с) Текстові файли було стиснуто алгоритмами **zip, 7z, bzip2, xz, rar**

 text1.txt	08.02.2021 15:37	Файл TXT	2 КБ
 text1.txt.7z	08.02.2021 18:22	Архив WinRAR	1 КБ
 text1.txt.bz2	08.02.2021 18:17	Архив WinRAR	1 КБ
 text1.txt.rar	08.02.2021 18:15	Архив WinRAR	1 КБ
 text1.txt.xz	08.02.2021 18:16	Архив WinRAR	1 КБ
 text1.txt.zip	08.02.2021 18:11	Архив ZIP - WinR...	1 КБ

 text2.txt	08.02.2021 16:19	Файл TXT	2 КБ
 text2.txt.7z	08.02.2021 18:33	Архив WinRAR	1 КБ
 text2.txt.bz2	08.02.2021 18:34	Архив WinRAR	1 КБ
 text2.txt.rar	08.02.2021 18:34	Архив WinRAR	1 КБ
 text2.txt.xz	08.02.2021 18:34	Архив WinRAR	1 КБ
 text2.txt.zip	08.02.2021 18:34	Архив ZIP - WinR...	1 КБ

 text3.txt	08.02.2021 17:45	Файл TXT	8 КБ
 text3.txt.7z	08.02.2021 18:36	Архив WinRAR	2 КБ
 text3.txt.bz2	08.02.2021 18:36	Архив WinRAR	2 КБ
 text3.txt.rar	08.02.2021 18:36	Архив WinRAR	3 КБ
 text3.txt.xz	08.02.2021 18:36	Архив WinRAR	2 КБ
 text3.txt.zip	08.02.2021 18:37	Архив ZIP - WinR...	3 КБ

d) 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)

Текст №1([І МЕРТВИМ](#), [І ЖИВИМ](#), [І НЕНАРОЖДЕННИМ](#))

```
Number of letters in text: 649
Entropy (bits): 5,0058
Ammount of information (bits): 3248,7735
Ammount of information (bytes): 406,0967

Filesize: 1360 bytes
.rar: 648
.zip: 720
.7z: 678
.bz2: 505
.xz: 608
```

Текст №2([Жаба й Віл](#))

```
Number of letters in text: 703
Entropy (bits): 4,9621
Ammount of information (bits): 3488,3589
Ammount of information (bytes): 436,0449

Filesize: 1454 bytes
.rar: 676
.zip: 748
.7z: 698
.bz2: 533
.xz: 628
```

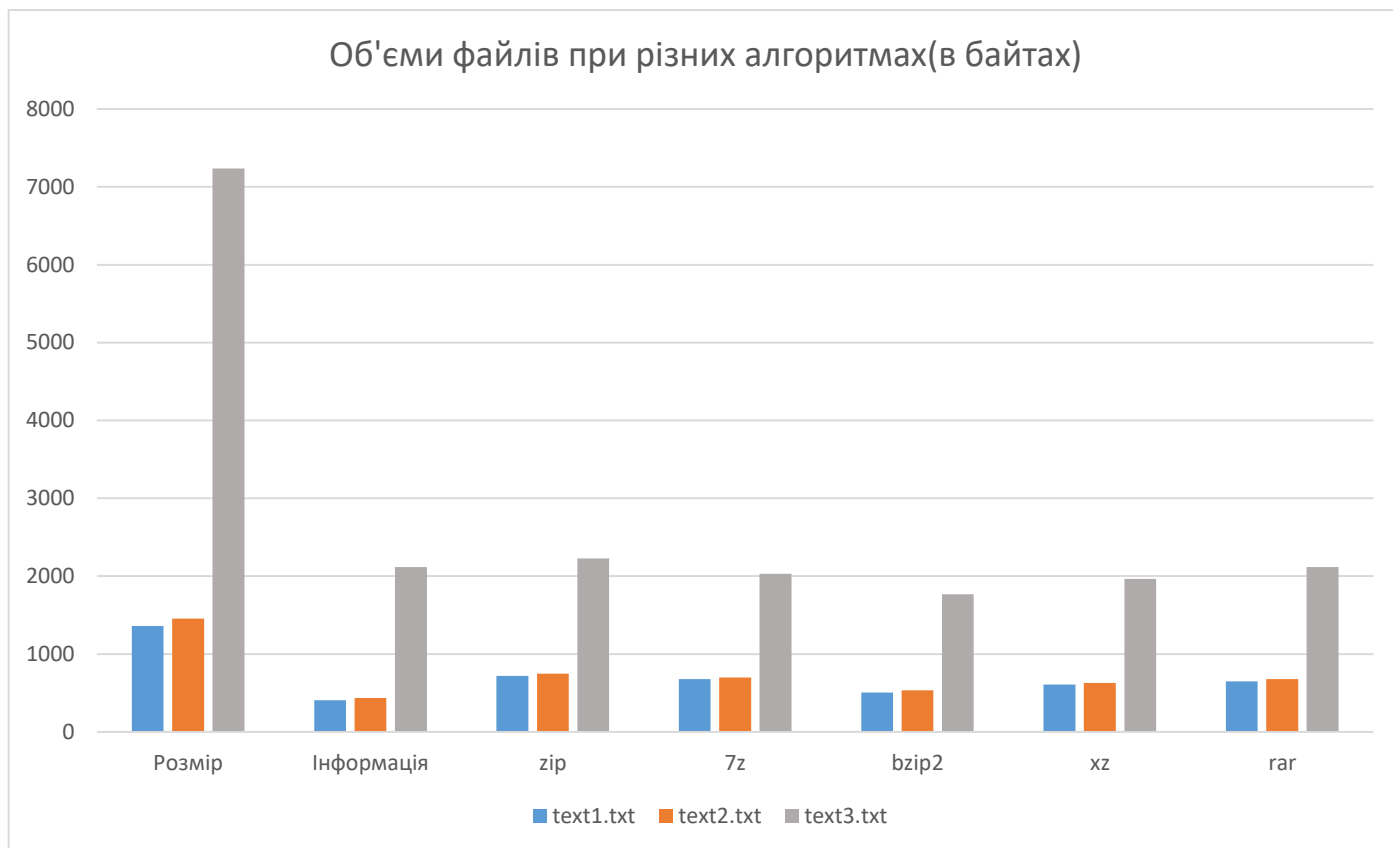
Текст №3([Радіотехніка](#))

```
Number of letters in text: 3605
Entropy (bits): 4,6973
Ammount of information (bits): 16933,8581
Ammount of information (bytes): 2116,7323

Filesize: 7234 bytes
.rar: 2116
.zip: 2227
.7z: 2032
.bz2: 1767
.xz: 1964
```

Таблиця розмірів файлів та об'ємів інформації в байтах

Файл	text1.txt	text2.txt	text3.txt
Розмір	1360	1454	7234
Ентропія	5,01 bits	4,96 bits	4,69 bits
Інформація	406,09	436,04	2116,73
zip	720	748	2227
7z	678	698	2032
bzip2	505	533	1767
xz	608	628	1964
rar	648	676	2116



Висновок: як видно з таблиці та діаграми реальний розмір архівів виявився досить близьким до апроксимованого значення інформації для текстового файлу великого розміру(text3.txt), і в середньому був навіть дещо менше, що демонструє ефективність сучасних алгоритмів стиснення. Для менших файлів(text1.txt та text2.txt) реальний розмір архівів виявився більшим ніж апроксимоване значення інформації, це пояснюється тим, що розмір текстових файлів менший, тому й службова інформація, що додається при стисненні є більш вагомою в порівнянні з великим текстовим файлом, також так як файл малий - то менше повторюваної інформації, яку можна замінити при стисненні. Серед всіх алгоритмів використаних для порівняння **bzip2** показав себе найкраще в стисненні файлів, але на роботу цього методу стиснення йшло більше часу ніж на інші алгоритми.

Як можна побачити, все було правильно декодовано.

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

```
Name of file:
text1
=====
Number of letters in text: 1581
Entropy (bits): 4,9503
Ammount of information (bits): 7826,3977
Ammount of information (bytes): 978,2997

Filesize: 1585 bytes
```

```
Name of file:
text2
Number of letters in text: 1681
Entropy (bits): 4,9492
Ammount of information (bits): 8319,5811
Ammount of information (bytes): 1039,9476

Filesize: 1685 bytes
```

```
Name of file:
text3
Number of letters in text: 8953
Entropy (bits): 4,8725
Ammount of information (bits): 43623,2273
Ammount of information (bytes): 5452,9034

Filesize: 8957 bytes
```

Таблиця об'ємів інформації в байтах оригінального і закодованого файлів

Файл	text1.txt	text2.txt	text3.txt
Інформація	406,09	436,04	2116,73
Інформація(Base64)	978,29	1039,94	5452,90



Висновок: згідно таблиці та діаграми після кодування Base64 к-сть інформації зросла в середньому в 2.66 рази. Це пояснюється тим, що кожен 3 байти оригінального файлу кодуються 4 –ма символами (збільшення на $\frac{1}{3}$). Також могло також вплинути на збільшення кінцевої інформації те, що після кодування кожен байт рахується як окремий символ - це і дало таке велике збільшення інформації.

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

Текстові файли були закодовані в **bzip2**, так як він показав себе найкраще

```
Name of file:
text1
Number of letters in text: 733
Entropy (bits): 7,6922
Ammount of information (bits): 5638,3837
Ammount of information (bytes): 704,7980

Filesize: 732 bytes
```

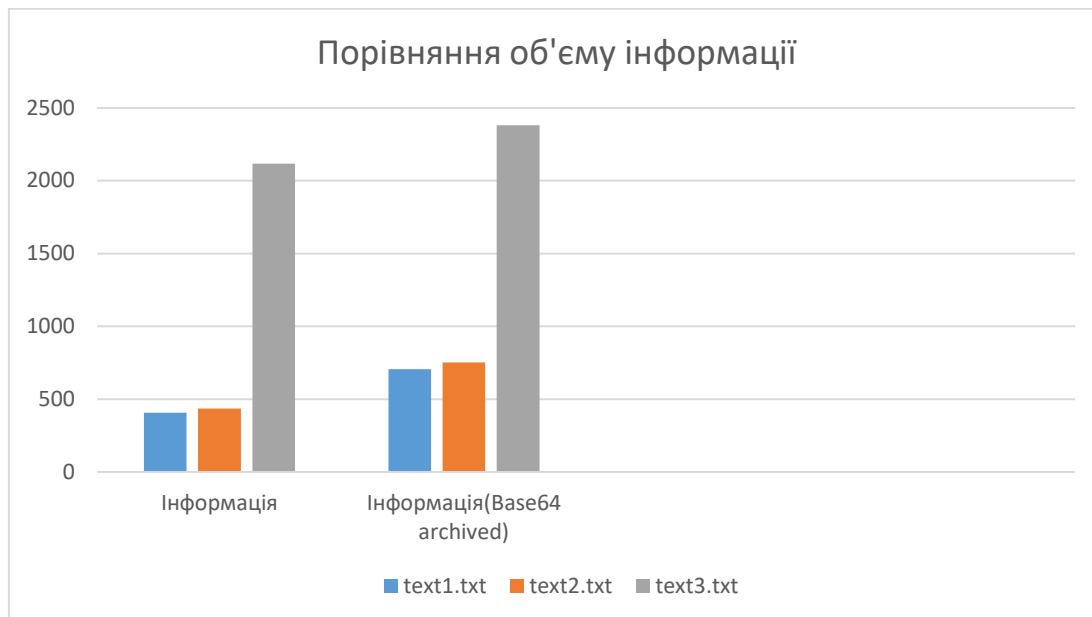
```
Name of file:
text2
Number of letters in text: 780
Entropy (bits): 7,7171
Ammount of information (bits): 6019,3266
Ammount of information (bytes): 752,4158

Filesize: 779 bytes
```

```
Name of file:
text3
Number of letters in text: 2414
Entropy (bits): 7,8860
Ammount of information (bits): 19036,9245
Ammount of information (bytes): 2379,6156

Filesize: 2413 bytes
```

Файл	text1.txt	text2.txt	text3.txt
Інформація	406,09	436,04	2116,73
Інформація(Base64 archived)	704,79	752,41	2379,61



Як видно з таблиці і графіку overhead у файлів малого розміру складає більшу частку розміру ніж у великих файлів.

Висновок: було досліджено що таке ентропія і як її підраховувати, також було розглянуто як працюють алгоритми стиснення і як вони впливають на кінцевий розмір файлу. Згідно отриманих результатів можна зробити висновок, що в архівах інформація зберігається майже без надлишкових даних і в максимально стиснутому виді.