

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  
KI - СА  
Понзель Максим

**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 репку" Леся Подерв'янського та специфікацію інтерфейсу PCI)
  - army.txt
  - song.txt
  - film.txt
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
  - Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gz, bz2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення).

## Analyze: army.txt

```
Symbol а probability: 0,07329843 Entropy for symbol: 0,27634050
Symbol б probability: 0,01570681 Entropy for symbol: 0,09412251
Symbol в probability: 0,05497382 Entropy for symbol: 0,23007157
Symbol г probability: 0,00523560 Entropy for symbol: 0,03967240
Symbol ґ probability: 0,00000000 Entropy for symbol: NaN
Symbol д probability: 0,02879581 Entropy for symbol: 0,14737688
Symbol е probability: 0,02617801 Entropy for symbol: 0,13757855
Symbol є probability: 0,00261780 Entropy for symbol: 0,02245400
Symbol ж probability: 0,01570681 Entropy for symbol: 0,09412251
Symbol з probability: 0,02356021 Entropy for symbol: 0,12740192
Symbol и probability: 0,07591623 Entropy for symbol: 0,28236646
Symbol і probability: 0,05497382 Entropy for symbol: 0,23007157
Symbol ї probability: 0,01832461 Entropy for symbol: 0,10573434
Symbol й probability: 0,01570681 Entropy for symbol: 0,09412251
Symbol к probability: 0,04450262 Entropy for symbol: 0,19981524
Symbol л probability: 0,01832461 Entropy for symbol: 0,10573434
Symbol м probability: 0,02094241 Entropy for symbol: 0,11680479
Symbol н probability: 0,07853403 Entropy for symbol: 0,28826216
Symbol о probability: 0,07853403 Entropy for symbol: 0,28826216
Symbol п probability: 0,01308901 Entropy for symbol: 0,08187828
Symbol р probability: 0,04450262 Entropy for symbol: 0,19981524
Symbol с probability: 0,04712042 Entropy for symbol: 0,20768343
Symbol т probability: 0,04973822 Entropy for symbol: 0,21534169
Symbol у probability: 0,05497382 Entropy for symbol: 0,23007157
Symbol ф probability: 0,00000000 Entropy for symbol: NaN
Symbol х probability: 0,00785340 Entropy for symbol: 0,05491466
Symbol ц probability: 0,00523560 Entropy for symbol: 0,03967240
Symbol ч probability: 0,00523560 Entropy for symbol: 0,03967240
Symbol ш probability: 0,00000000 Entropy for symbol: NaN
Symbol щ probability: 0,00523560 Entropy for symbol: 0,03967240
Symbol ь probability: 0,02356021 Entropy for symbol: 0,12740192
Symbol ю probability: 0,00785340 Entropy for symbol: 0,05491466
Symbol я probability: 0,01570681 Entropy for symbol: 0,09412251
Total entropy: 4,265476
The amount of information in the text: 203,676 bytes
```

Size: 802 bytes

Analyze: song.txt

```
Symbol а probability: 0,09067358 Entropy for symbol: 0,31401837
Symbol б probability: 0,02331606 Entropy for symbol: 0,12643209
Symbol в probability: 0,03367876 Entropy for symbol: 0,16475706
Symbol г probability: 0,00518135 Entropy for symbol: 0,03933916
Symbol ґ probability: 0,00000000 Entropy for symbol: NaN
Symbol д probability: 0,02072539 Entropy for symbol: 0,11590585
Symbol е probability: 0,01295337 Entropy for symbol: 0,08122447
Symbol є probability: 0,00000000 Entropy for symbol: NaN
Symbol ж probability: 0,00518135 Entropy for symbol: 0,03933916
Symbol з probability: 0,00000000 Entropy for symbol: NaN
Symbol и probability: 0,04663212 Entropy for symbol: 0,20623206
Symbol і probability: 0,04663212 Entropy for symbol: 0,20623206
Symbol ї probability: 0,01554404 Entropy for symbol: 0,09338074
Symbol й probability: 0,03108808 Entropy for symbol: 0,15567341
Symbol к probability: 0,00777202 Entropy for symbol: 0,05446239
Symbol л probability: 0,00777202 Entropy for symbol: 0,05446239
Symbol м probability: 0,01813472 Entropy for symbol: 0,10491118
Symbol н probability: 0,04145078 Entropy for symbol: 0,19036091
Symbol о probability: 0,04922280 Entropy for symbol: 0,21384990
Symbol п probability: 0,02590674 Entropy for symbol: 0,13654220
Symbol р probability: 0,02849741 Entropy for symbol: 0,14627793
Symbol с probability: 0,04663212 Entropy for symbol: 0,20623206
Symbol т probability: 0,04145078 Entropy for symbol: 0,19036091
Symbol у probability: 0,04663212 Entropy for symbol: 0,20623206
Symbol ф probability: 0,00000000 Entropy for symbol: NaN
Symbol х probability: 0,04145078 Entropy for symbol: 0,19036091
Symbol ц probability: 0,00777202 Entropy for symbol: 0,05446239
Symbol ч probability: 0,00259067 Entropy for symbol: 0,02226025
Symbol ш probability: 0,00000000 Entropy for symbol: NaN
Symbol щ probability: 0,00000000 Entropy for symbol: NaN
Symbol ь probability: 0,00000000 Entropy for symbol: NaN
Symbol ю probability: 0,04663212 Entropy for symbol: 0,20623206
Symbol я probability: 0,01554404 Entropy for symbol: 0,09338074
Total entropy: 3,612923
The amount of information in the text: 174,324 bytes
```

Size: 768 bytes

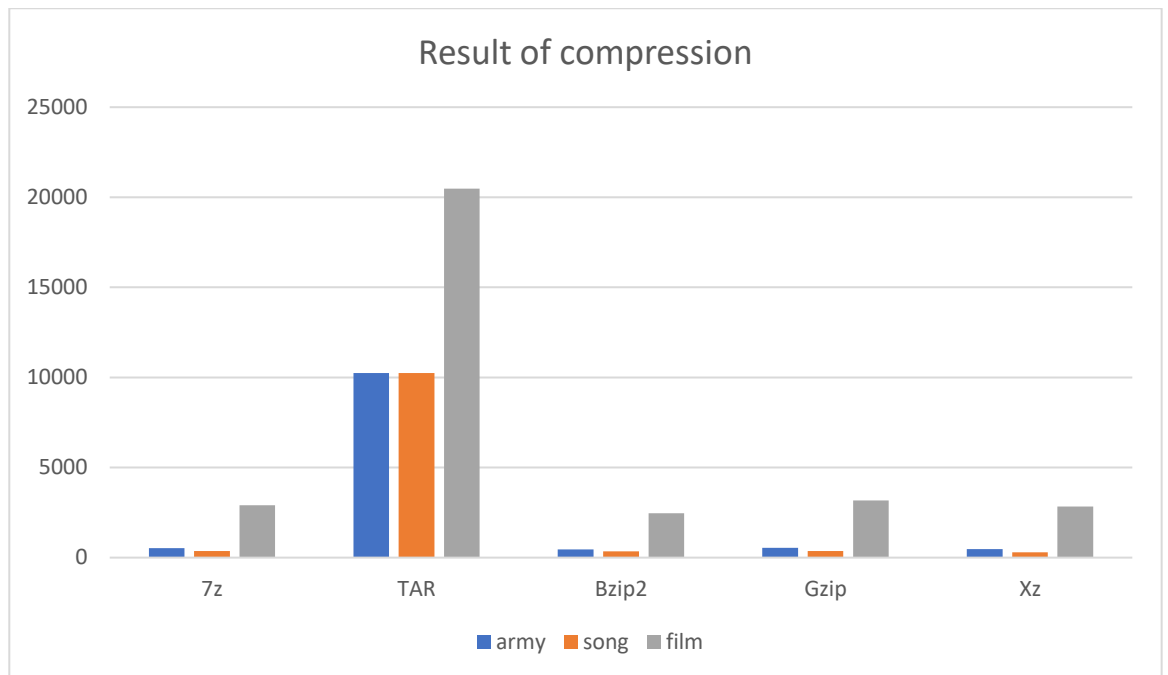
Analyze: film.txt

```
Symbol а probability: 0,07061393 Entropy for symbol: 0,27002084
Symbol б probability: 0,01625819 Entropy for symbol: 0,09661737
Symbol в probability: 0,04416404 Entropy for symbol: 0,19878163
Symbol г probability: 0,01674351 Entropy for symbol: 0,09879096
Symbol ґ probability: 0,00000000 Entropy for symbol: NaN
Symbol д probability: 0,03858287 Entropy for symbol: 0,18118113
Symbol е probability: 0,04246542 Entropy for symbol: 0,19353903
Symbol є probability: 0,01504489 Entropy for symbol: 0,09109054
Symbol ж probability: 0,01189032 Entropy for symbol: 0,07602751
Symbol з probability: 0,02863383 Entropy for symbol: 0,14678088
Symbol и probability: 0,04974521 Entropy for symbol: 0,21536186
Symbol і probability: 0,06139287 Entropy for symbol: 0,24715449
Symbol ї probability: 0,00461053 Entropy for symbol: 0,03578165
Symbol й probability: 0,01383159 Entropy for symbol: 0,08542239
Symbol к probability: 0,03033244 Entropy for symbol: 0,15296635
Symbol л probability: 0,04173744 Entropy for symbol: 0,19126241
Symbol м probability: 0,02208202 Entropy for symbol: 0,12147284
Symbol н probability: 0,05459840 Entropy for symbol: 0,22904015
Symbol о probability: 0,08638680 Entropy for symbol: 0,30520848
Symbol п probability: 0,02305266 Entropy for symbol: 0,12538163
Symbol р probability: 0,04149478 Entropy for symbol: 0,19049949
Symbol с probability: 0,03033244 Entropy for symbol: 0,15296635
Symbol т probability: 0,05362776 Entropy for symbol: 0,22635614
Symbol у probability: 0,03567095 Entropy for symbol: 0,17154541
Symbol ф probability: 0,00048532 Entropy for symbol: 0,00534277
Symbol х probability: 0,00970638 Entropy for symbol: 0,06490513
Symbol ц probability: 0,00412521 Entropy for symbol: 0,03267711
Symbol ч probability: 0,01019170 Entropy for symbol: 0,06743300
Symbol ш probability: 0,00558117 Entropy for symbol: 0,04177626
Symbol щ probability: 0,00412521 Entropy for symbol: 0,03267711
Symbol ь probability: 0,02281000 Entropy for symbol: 0,12441006
Symbol ю probability: 0,01310362 Entropy for symbol: 0,08194859
Symbol я probability: 0,02353798 Entropy for symbol: 0,12731376
Total entropy: 4,381733
The amount of information in the text: 2257,140 bytes
```

Size: 8 806 bytes

### Result of compression

Name	7z	TAR	Bzip2	Gzip	Xz	Amount of information
army.txt	532	10 240	458	537	464	204
song.txt	360	10 240	353	360	292	174
film.txt	2 907	20 480	2 470	3 172	2 840	2257



## Дослідження способів кодування інформації на прикладі Base64

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)**

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**


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**

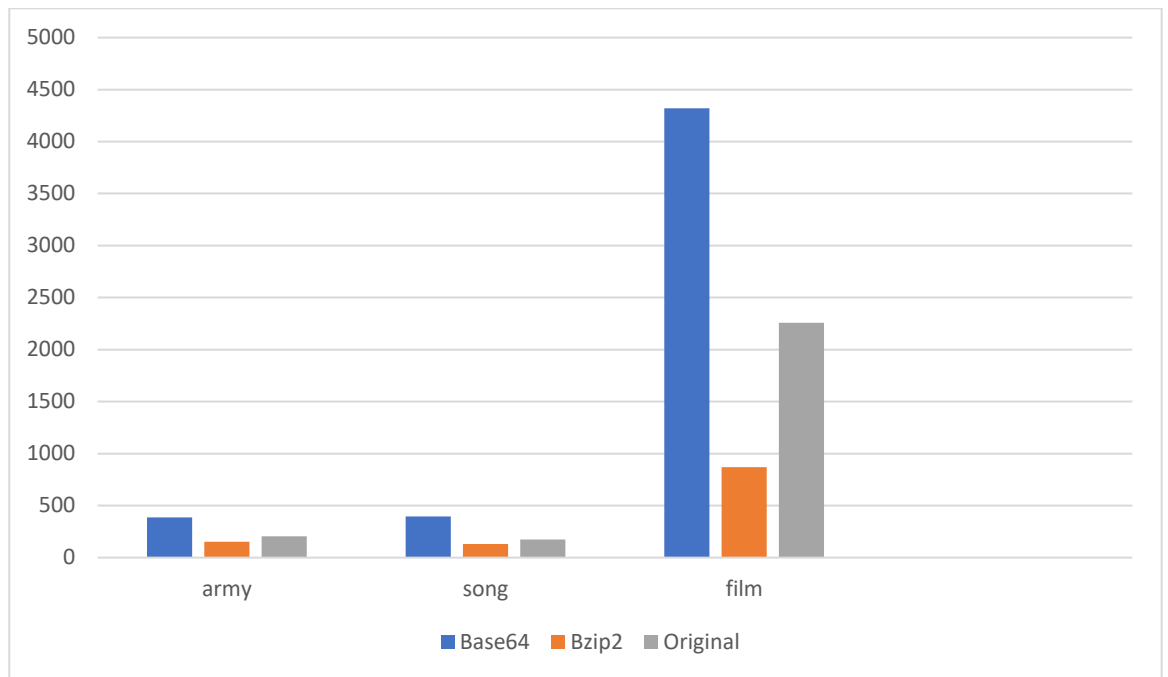
**Зробіть висновки з отриманого результату**

 [armyBase64.txt](#)

 [songBase64.txt](#)

 [filmBase64.txt](#)

Name	BZip	Base 64	Orig
armyBase64.txt	152	388	204
songBase64.txt	131	395	174
filmBase64.txt	870	4 321	2257



#### Висновок:

В даній лабораторній роботі проведено аналіз текстових файлів.

Виявлено, що обсяг інформації є значно меншим аніж обсяг стиснутого файлу на диску.

Встановлено що найкращим з перевірених алгоритмом стиснення є BZip2.

GitHub link: <https://github.com/Ponzel0106>