Київський національний університет імені Тараса Шевченка
радіофізичний факультет

Звіт до
лабораторної роботи № 1
з предмету «Комп’ютерні системи»
Тема: «Дослідження кількості інформації при різних варіантах кодування»

Роботу виконав
студент 3 курсу
Комп’ютерна інженерія
Качмарський Олекса

Київ 2019

Мета: Дослідити імовірнісні параметри української мови для оцінки кількості інформації текстів. Дослідити вплив різних методів кодування інформації на її кількість.

Хід роботи

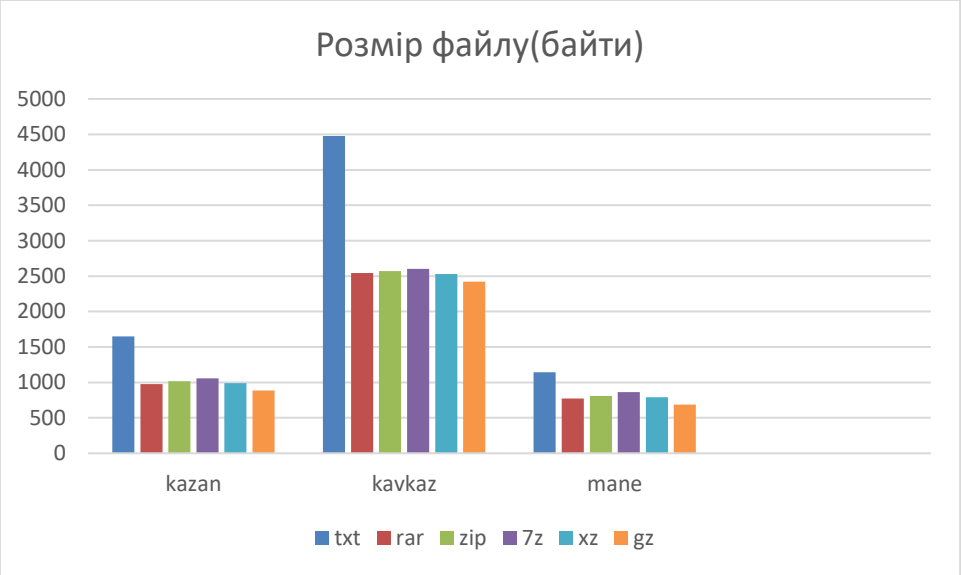
1. Дослідження кількості інформації в тексті

1) Аналіз файлів

Назва	Ентропія	Кількість інформації(байт)	Розмір файлу(байт)
kazan.txt	4.697	958.114	1648
kavkaz.txt	4.930	2529.720	4475
mane.txt	4.680	662.744	1145

2) Розміри файлів після стиснення

Назва	rar(байт)	zip(байт)	7z(байт)	xz(байт)	gz(байт)
kazan.txt	974	1017	1057	988	886
kavkaz.txt	2544	2571	2604	2528	2421
mane.txt	772	809	862	792	687



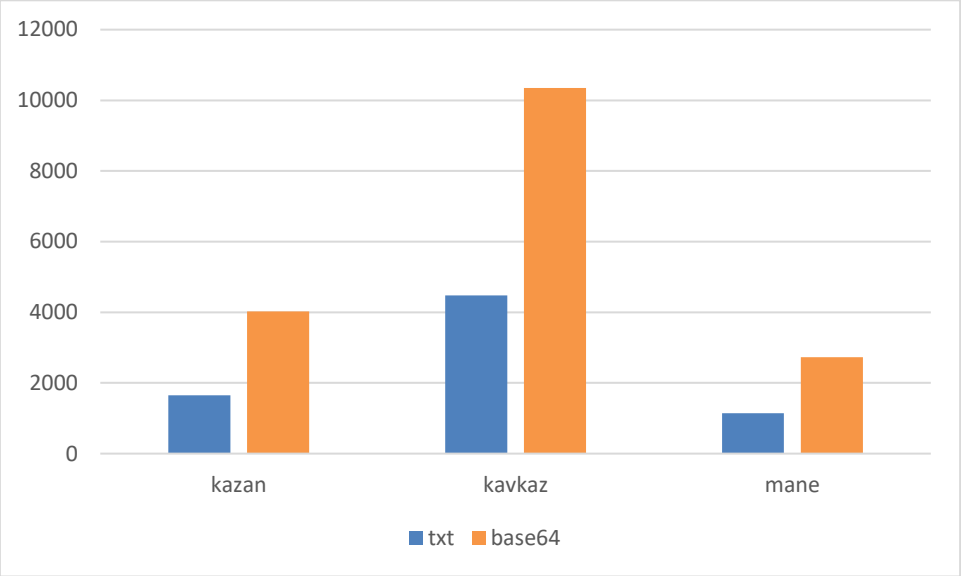
У ідеальному варіанті розмір файлу після стиснення повинен дорівнювати кількості інформації, але у моєму випадку розміри файлів після стиснення виявилися більше ніж у ідеальному варіанті. Це відбувається, тому що алгоритми архіваторів використовують повторювані частини тексту, тобто використана формула для підрахунку кількості інформації не досконала. Вона не враховує частоту появи символів та передбачення наступної частити тексту.

2. Дослідження способів кодування інформації на прикладі Base64

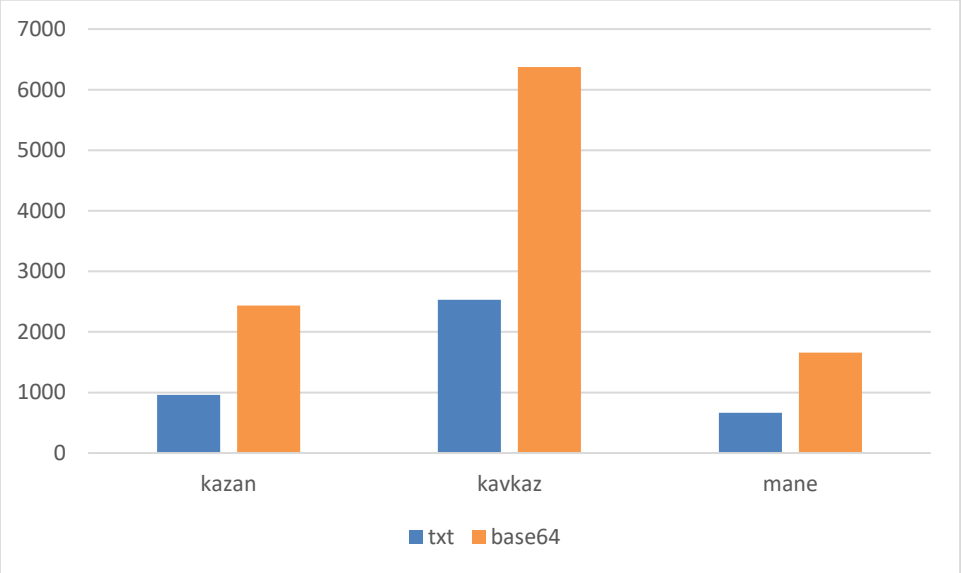
1) Аналіз закодованих файлів

Назва	Кількість символів	Ентропія	Кількість інформації	Розмір файлу(байт)
kazan.txt	4024	4.845	2437.132	4024
kavkaz.txt	10352	4.926	6373.607	10352
mane.txt	2728	4.870	1660.774	2728

Розмір файлів до і після кодування:



Кількість інформації у файлах до і після кодування:

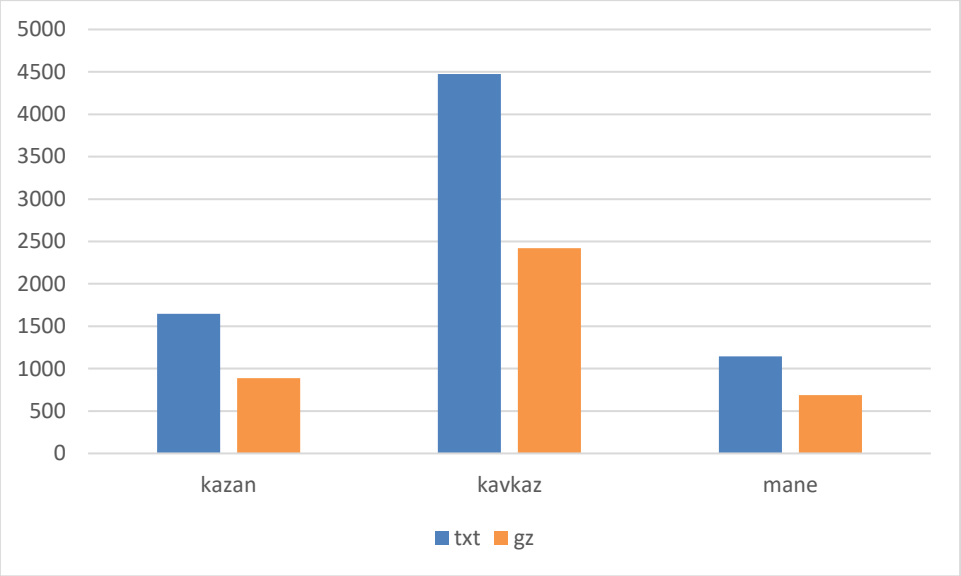


Кількість інформації у закодованих файлах збільшилася, це пов’язано з особливостями base64 кодування, яке збільшує кількість символів у тексті.

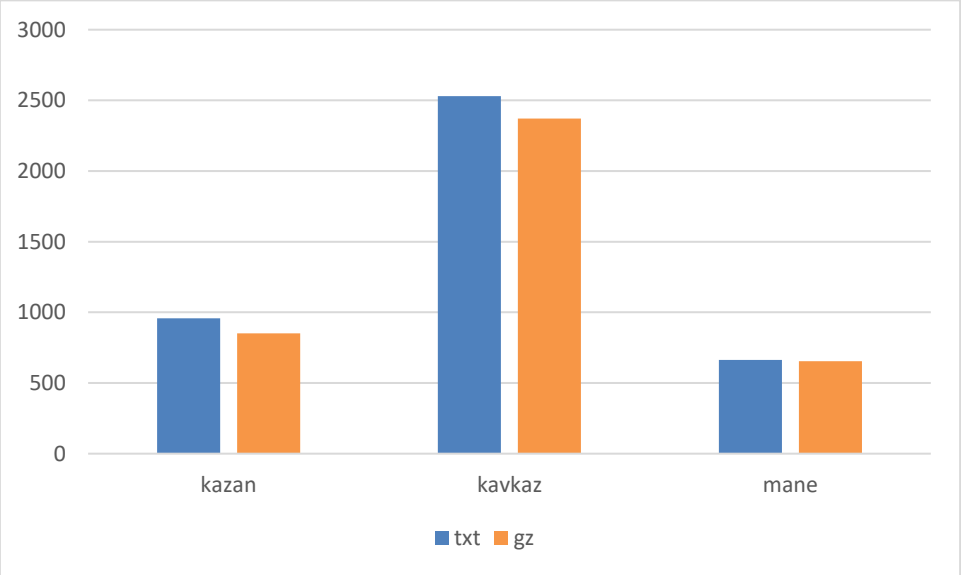
2) Порівняння файлів після стиснення

Найкращим інструментом для стиснення файлів, як можна побачити на першій діаграмі, виявився gz.

Розмір файлів до і після архівування:



Кількість інформації у файлах до і після архівування:



Приклади роботи створених програм:

```
--------1ST FILE----------

kazan.txt

General amount of symbols in the text  1632
Frequency of symbol(max 1) '?' = 0.001225
Frequency of symbol(max 1) '?' = 0.086397
Frequency of symbol(max 1) '?' = 0.022059
Frequency of symbol(max 1) '?' = 0.017770
Frequency of symbol(max 1) '?' = 0.006127
Frequency of symbol(max 1) '?' = 0.083946
Frequency of symbol(max 1) '?' = 0.038603
Frequency of symbol(max 1) '?' = 0.050245
Frequency of symbol(max 1) '?' = 0.013480
Frequency of symbol(max 1) ' ' = 0.123775
Frequency of symbol(max 1) '?' = 0.042279
Frequency of symbol(max 1) '?' = 0.045956
Frequency of symbol(max 1) '?' = 0.064951
Frequency of symbol(max 1) '?' = 0.017157
Frequency of symbol(max 1) '?' = 0.060049
Frequency of symbol(max 1) '?' = 0.025735
Frequency of symbol(max 1) '?' = 0.039216
Frequency of symbol(max 1) '1' = 0.003676
Frequency of symbol(max 1) '5' = 0.006127
Frequency of symbol(max 1) '2' = 0.000613
Frequency of symbol(max 1) '?' = 0.031863
Frequency of symbol(max 1) '?' = 0.024510
Frequency of symbol(max 1) '?' = 0.038603
Frequency of symbol(max 1) '?' = 0.011029
Frequency of symbol(max 1) '?' = 0.010417
Frequency of symbol(max 1) '?' = 0.028799
Frequency of symbol(max 1) '?' = 0.002451
Frequency of symbol(max 1) '.' = 0.012255
Frequency of symbol(max 1) '?' = 0.001838
Frequency of symbol(max 1) '?' = 0.010417
Frequency of symbol(max 1) '?' = 0.003064
Frequency of symbol(max 1) '?' = 0.007353
Frequency of symbol(max 1) '?' = 0.012255
Frequency of symbol(max 1) '?' = 0.006127
Frequency of symbol(max 1) '-' = 0.000613
Frequency of symbol(max 1) ',' = 0.009804
Frequency of symbol(max 1) '?' = 0.007966
Frequency of symbol(max 1) '?' = 0.003064
Frequency of symbol(max 1) '?' = 0.001838
Frequency of symbol(max 1) '?' = 0.004289
Frequency of symbol(max 1) '?' = 0.001225
Frequency of symbol(max 1) '?' = 0.001838
Frequency of symbol(max 1) '?' = 0.001225
Frequency of symbol(max 1) '?' = 0.003676
Frequency of symbol(max 1) '-' = 0.001838
Frequency of symbol(max 1) '0' = 0.000613
Frequency of symbol(max 1) '3' = 0.001225
Frequency of symbol(max 1) '?' = 0.001225
Frequency of symbol(max 1) '?' = 0.000613
Frequency of symbol(max 1) '?' = 0.000613
Frequency of symbol(max 1) '4' = 0.000613
Frequency of symbol(max 1) '?' = 0.001225
Frequency of symbol(max 1) '?' = 0.000613
Frequency of symbol(max 1) '?' = 0.001838
Frequency of symbol(max 1) '6' = 0.000613
Frequency of symbol(max 1) '?' = 0.001225
Frequency of symbol(max 1) '?' = 0.000613
Frequency of symbol(max 1) ''' = 0.000613
Frequency of symbol(max 1) '?' = 0.000613

Average entropy in this text -> 4.697 bit
Amount of information in this text -> 958.114 byte
```

```
--------ENCODED FILE----------

kazanBase64.txt

General amount of symbols in the text  4024
Frequency of symbol(max 1) '0' = 0.113072
Frequency of symbol(max 1) 'K' = 0.001988
Frequency of symbol(max 1) 'D' = 0.036531
Frequency of symbol(max 1) 'Q' = 0.080517
Frequency of symbol(max 1) 'v' = 0.029573
Frequency of symbol(max 1) 't' = 0.040010
Frequency of symbol(max 1) 'C' = 0.094930
Frequency of symbol(max 1) '3' = 0.011183
Frequency of symbol(max 1) 'L' = 0.089960
Frequency of symbol(max 1) '/' = 0.005219
Frequency of symbol(max 1) 'G' = 0.038022
Frequency of symbol(max 1) 'H' = 0.007704
Frequency of symbol(max 1) 'u' = 0.023857
Frequency of symbol(max 1) '9' = 0.028082
Frequency of symbol(max 1) 'w' = 0.014165
Frequency of symbol(max 1) 'Y' = 0.028579
Frequency of symbol(max 1) 'R' = 0.033300
Frequency of symbol(max 1) 'j' = 0.012922
Frequency of symbol(max 1) 'y' = 0.010934
Frequency of symbol(max 1) '6' = 0.005964
Frequency of symbol(max 1) 'I' = 0.020626
Frequency of symbol(max 1) 'N' = 0.044732
Frequency of symbol(max 1) '7' = 0.012922
Frequency of symbol(max 1) 'M' = 0.003976
Frequency of symbol(max 1) 'S' = 0.001491
Frequency of symbol(max 1) '+' = 0.012425
Frequency of symbol(max 1) 'g' = 0.037276
Frequency of symbol(max 1) 'd' = 0.023111
Frequency of symbol(max 1) '8' = 0.004722
Frequency of symbol(max 1) 'T' = 0.003728
Frequency of symbol(max 1) 'U' = 0.002485
Frequency of symbol(max 1) '1' = 0.004722
Frequency of symbol(max 1) 'i' = 0.008201
Frequency of symbol(max 1) 'r' = 0.004970
Frequency of symbol(max 1) 'W' = 0.005219
Frequency of symbol(max 1) 'P' = 0.004970
Frequency of symbol(max 1) 'x' = 0.001491
Frequency of symbol(max 1) 'z' = 0.007704
Frequency of symbol(max 1) 'B' = 0.007207
Frequency of symbol(max 1) 'm' = 0.000746
Frequency of symbol(max 1) 'f' = 0.004225
Frequency of symbol(max 1) 's' = 0.022863
Frequency of symbol(max 1) 'Z' = 0.005716
Frequency of symbol(max 1) '4' = 0.012922
Frequency of symbol(max 1) 'l' = 0.005964
Frequency of symbol(max 1) 'n' = 0.002485
Frequency of symbol(max 1) 'X' = 0.003479
Frequency of symbol(max 1) 'A' = 0.007455
Frequency of symbol(max 1) 'o' = 0.000994
Frequency of symbol(max 1) 'b' = 0.004225
Frequency of symbol(max 1) 'h' = 0.002485
Frequency of symbol(max 1) '2' = 0.000746
Frequency of symbol(max 1) 'c' = 0.000746
Frequency of symbol(max 1) 'p' = 0.000497
Frequency of symbol(max 1) 'J' = 0.001740
Frequency of symbol(max 1) '5' = 0.000994
Frequency of symbol(max 1) 'k' = 0.000994
Frequency of symbol(max 1) 'O' = 0.000746
Frequency of symbol(max 1) 'F' = 0.000994
Frequency of symbol(max 1) 'a' = 0.000249
Frequency of symbol(max 1) '=' = 0.000249

Average entropy in this text -> 4.845 bit
Amount of information in this text -> 2437.132 byte
```

Посилання на github:
https://github.com/skantorp/CompSystems
Посилання на файли:
https://github.com/skantorp/CompSystems/tree/master/Lab1/lab1/bin/Debug

Висновок

Під час лабораторної роботи я дослідив імовірнісні параметри української мови для оцінки кількості інформації текстів та вплив різних методів кодування інформації на її кількість.