Київський національний університет імені Тараса Шевченка
Факультет радіофізики, електроніки та компютерних систем

Лабораторна робота №1

Дослідження кількості інформації при різних варіантах кодування

Виконав студент

3 курсу СА-КІ

Герасименко Артем

1. Дослідження кількості інформації в тексті

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 РСІ) Я обрав:

- 1)Текст пісні Beliver
- 2)Інформацію про Skrillex з сайту wikipedia
- 3)1 розділ з книги Гра престолів

- 1. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
 - 1. обраховує частоти (імовірності) появи символів в тексті
 - 2. обраховує середню ентропію алфавіту для даного тексту
 - виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
 - 4. виводить на екран значення частот, ентропії та кількості інформації

Наведено результат роботи програми для 3 файлів відповідно Beliver Skrillex_wiki A GAME OF THRONES

Відносна частога появи символу
s -> 0.049179323860852525
0 -> 0.06167319941205292
n -> 0.05573248407643312
y -> 0.095615384615384616
-> 0.16137922586967174
j -> 0.0026335129838314553
h -> 0.02792748652621264
m -> 0.05635129838314553
h -> 0.02792748652621264
c -> 0.05643008329250368
e -> 0.09125428711415973
(-> 0.00141156225933366
b -> 0.014453699167074964
a -> 0.00141156225933366
b -> 0.014453699167074964
a -> 0.00251690347868692
1 -> 0.0014718275355218034
5 -> 0.0016536011758941694
, -> 0.00201690347868692
1 -> 0.00141156225933366
b -> 0.0016536011758941694
, -> 0.00329250367466928
w -> 0.00148623223909848
) -> 0.00164135277805
b -> 0.00164135277805
c -> 0.001549485546300833
c -> 0.0565899069037824
c -> 0.002231749142577167
d -> 0.01549485546300833
c -> 0.009799118079372856
c -> 0.009999118079372856
c -> 0.0009999118079372856
c -> 0.000999918079372856
c -> 0.000999918079372856
c -> 0.000999918079372856
c -> 0.00012484896593992
q -> 0.000542889759921607
c -> 0.000124889759921607
c -> 0.00018373346398824106
entproniff
H +> 4.533731447617764
Kijnkictb inhopmaniff B rekctii
H*k > 74026.76707670285

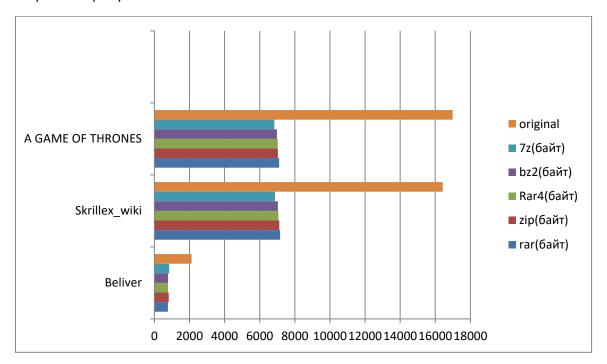
t -> 0.06328502415458938 h -> 0.05628019323671497 e -> 0.09136473429951691 -> 0.1820048309178744 m -> 0.01618357487922705 0 -> 0.05748792270531401 r -> 0.04818840579710145 n -> 0.04921497584541063 i -> 0.04432367149758454 g -> 0.015036231884057972 a -> 0.059782608695652176 d -> 0.04142512077294686 w -> 0.020169082125603865 c -> 0.01105072463768116 1 -> 0.03109903381642512 , -> 0.013345410628019324 s -> 0.04806763285024154 p -> 0.00857487922705314 f -> 0.017693236714975847 u -> 0.01920289855072464 . -> 0.0161231884057971 y -> 0.015821256038647345 b -> 0.015942028985507246 k -> 0.008816425120772946 v -> 0.004891304347826087 x -> 0.00042270531400966186 -> 0.003140096618357488 j -> 0.0038043478260869567 -> 0.021859903381642513 - -> 0.0007246376811594203 : -> 6.038647342995169e-05 q -> 0.0005434782608695652 " -> 0.0125 ? -> 0.0008454106280193237 ! -> 0.00036231884057971015 z -> 0.00036231884057971015 ентропія Н -> 4.3116643577174845 Кількість інформації в тексті H*k > 71401.16176380155

Назва	Кількість символів	Ентропія	Кількість інформації	Розмір файлу(байт)
Beliver	2060	4.286	8829	2119
Skrillex	16328	4.533	74,026	16,420
_wiki				
Α	16560	4.311	72,401	16,992
GAME				·
OF				
THRO				
NES				

Розміри файлів після стиснення

Назва	rar(байт)	zip(байт)	Rar4(байт)	bz2(байт)	7z(байт)
Beliver	773	818	770	784	844
Skrillex_wiki	7160	7,119	7,064	7,038	6,874
A GAME OF	7110	7,032	7,021	6,976	6,834
THRONES		·	·	·	

Результати архівування

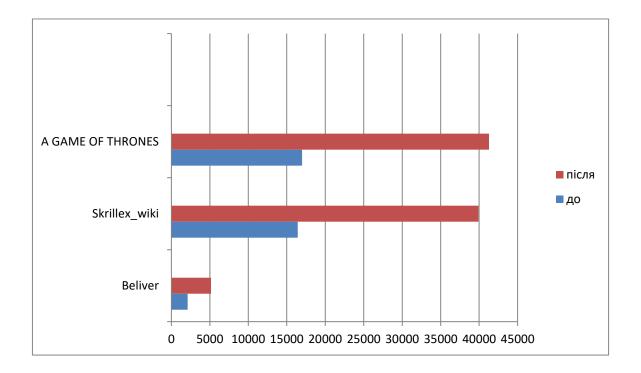


Дослідження способів кодування інформації на прикладі Base64

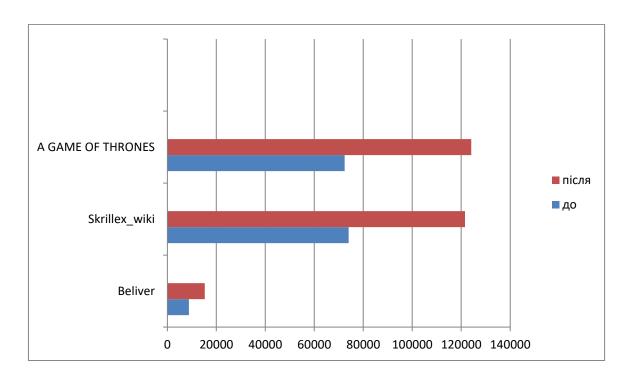
1) Аналіз закодованих файлів

Назва	Кількість символів	Ентропія	Кількість інформації	Розмір файлу(байт)
Beliver	2836	5.378	15252	5,159
Skrillex	21952	5.535	121526	39906
_wiki				
Α	22800	5.442	124082	41293
GAME				
OF				
THRO				
NES				

Розмір файлів до і після кодування:



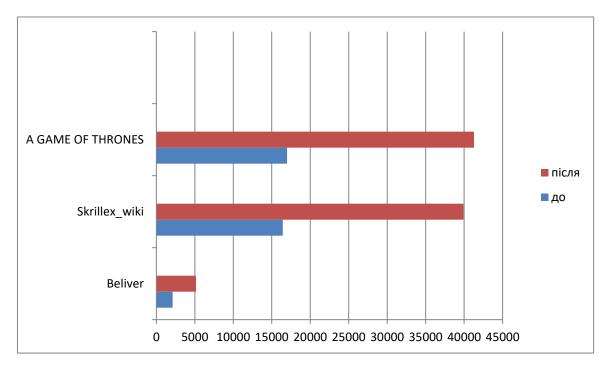
Кількість інформації у файлах до і після кодування:



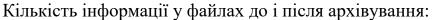
2) Порівняння файлів після стиснення

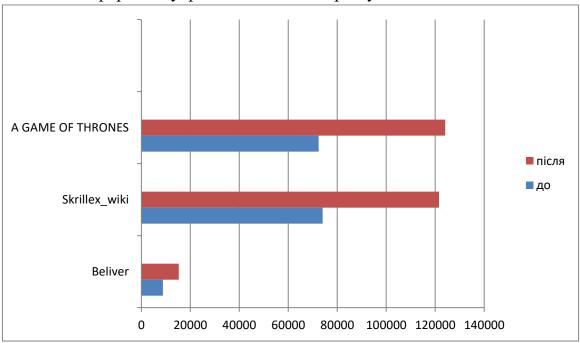
Найкращим інструментом для стиснення файлів, як можна побачити на першій діаграмі, виявився rar.

Розмір файлів до і після архівування:



Після кодування розмір файлів зростає приблизно у 1.5 рази





Після архівування кількість інформації зростає приблизно у 1/3 раз.

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.