**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
2. 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 |

1. 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.

1. Дослідження способів кодування інформації на прикладі Base64
2. 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.

1. Порівняння файлів після стиснення

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

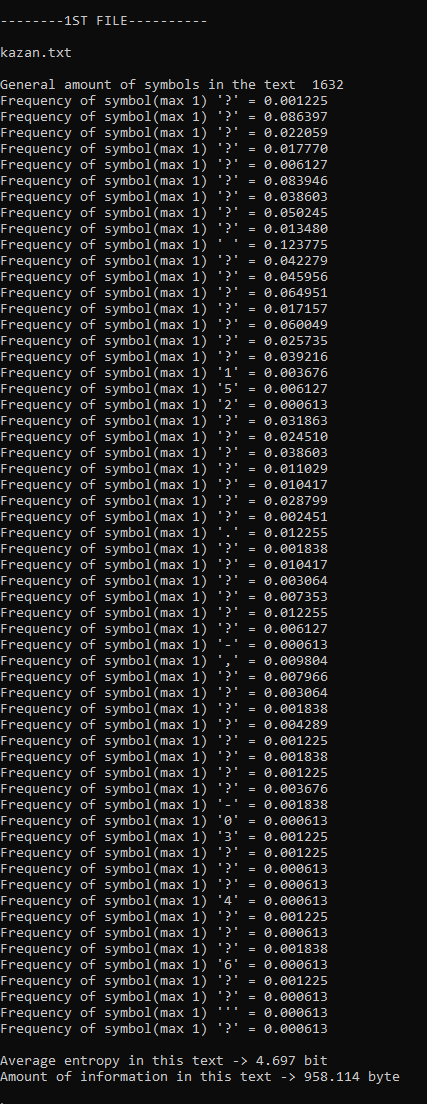
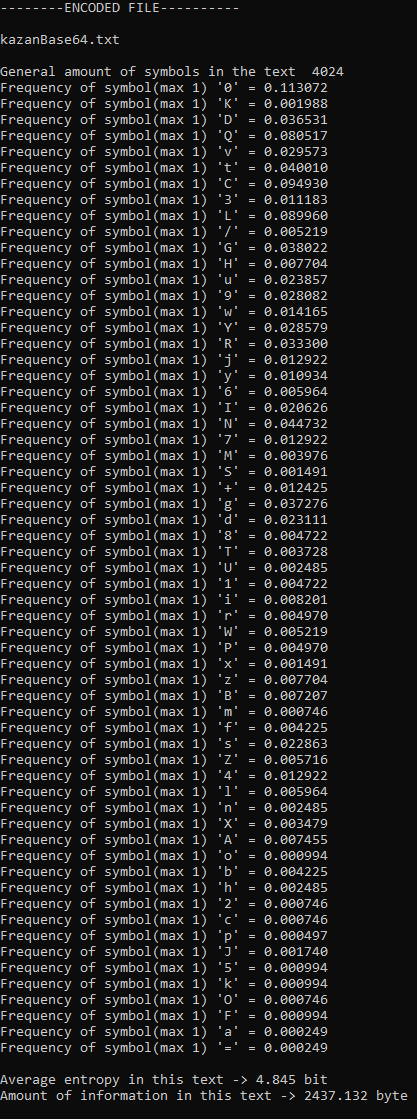
Розмір файлів до і після архівування:

Після кодування розмір файлів зростає приблизно у 1.5 рази.

Кількість інформації у файлах до і після архівування:

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

Приклади роботи створених програм:

Посилання на github:

[https://github.com/skantorp/CompSystems](https://github.com/andrew36986/LabsCS2019)

Посилання на файли:

<https://github.com/andrew36986/LabsCS2019/tree/master/Lab1CompSyst/Lab1CompSyst/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.