**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

КІ-СА

Гуліцький Олександр Сергійович

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**

Посилання на репозиторій: [github](https://github.com/alexgul1/CE-Labs).

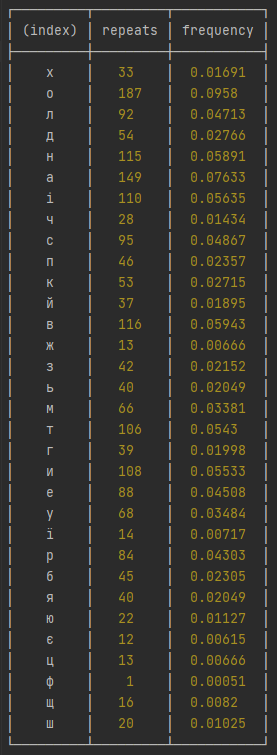
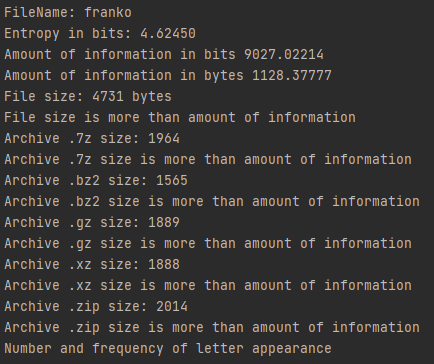
**Дослідження кількості інформації в тексті**

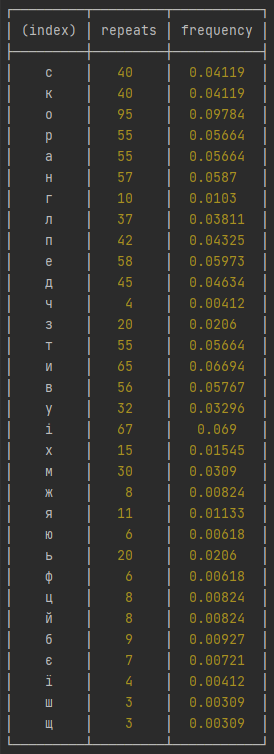
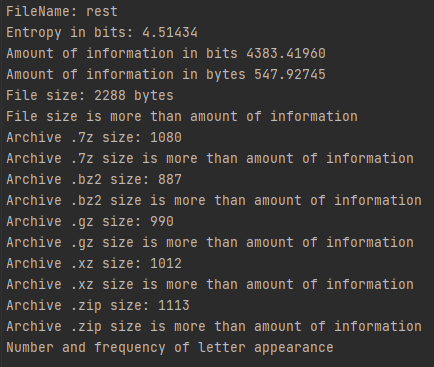
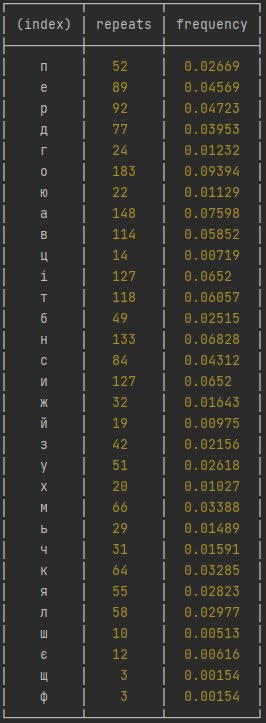
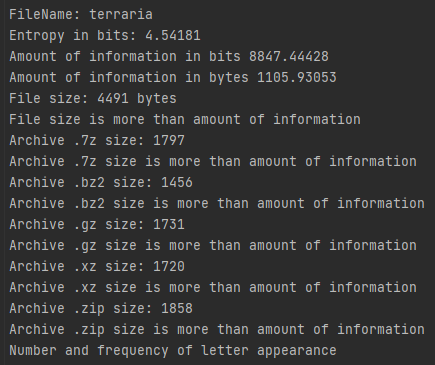
1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI).

* [REST](https://uk.wikipedia.org/wiki/REST)
* [Франко - Привид](https://www.i-franko.name/uk/Verses/ZivjaleLystja/1/20.html)

* [Terraria – ігровий процес](https://uk.wikipedia.org/wiki/Terraria)

1. Переконайтесь, що тексти, які ви використовуєте є унікальними і не повторюються у ваших колег! Використовуйте наявні електронні засоби зв’язку та документообігу, щоб уникнути дублювання! Вдруге аналіз того самого тексту не зараховується!  
     
   Виконано.
2. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
3. обраховує частоти (імовірності) появи символів в тексті
4. обраховує середню ентропію алфавіту для даного тексту
5. виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
6. виводить на екран значення частот, ентропії та кількості інформації
7. Франко – Привид



1. REST
2. Terraria – ігровий процес
3. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення).

За допомогою програми [7zip](https://www.7-zip.org/) файли було стиснено у 7z, bz2, gz, xz та zip.

1. 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).

|  |  |  |  |
| --- | --- | --- | --- |
| **Файл** | **Привид, байт** | **REST, байт** | **Terraria, байт** |
| Оригінал | 4731 | 2288 | 4491 |
| 7z | 1964 | 1080 | 1797 |
| bz2 | 1565 | 887 | 1456 |
| gz | 1889 | 990 | 1731 |
| xz | 1888 | 1012 | 1720 |
| zip | 2014 | 1113 | 1858 |
| Кількість інформації | 1128 | 547 | 1105 |

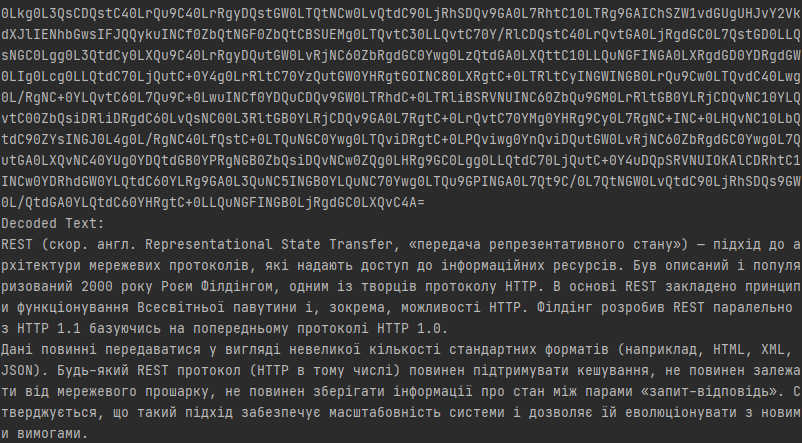
Для всіх файлів найкраще стискав алгоритм bzip2, але кількість інформації все одно більше чим обсяг файлу/архіву тому, що алгоритми стискання побудовані таким чином що вони використовують повторюванні частини тексту.

**Дослідження способів кодування інформації на прикладі Base64**

1. Ознайомтесь зі стандартом [RFC4648](https://tools.ietf.org/html/rfc4648)
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції).

a. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)

За допомогою існуючої функції в node.js декодив результат і отримав аналогічний текст, який використовувався для кодування.



1. Закодуйте в Base64 обрані вами текстові файли

a. Обрахуйте кількість інформації в base64-закодованому варіанті файлу

b. Порівняйте отримане значення з кількістю інформації вихідного файлу

c. Зробіть висновки з отриманого результату

|  |  |  |
| --- | --- | --- |
| **Файл** | **Кількість інформації у вихідному файлі** | **Кількість інформації у закодованому файлі** |
| Привид | 1128 | 3560 |
| REST | 547 | 1734 |
| Terraria | 1105 | 3378 |

Кількість інформації у закодованому файлі в середньому в 3.1 раза більше чим в вихідному.

1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли

a. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу

b. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу

c. Зробіть висновки з отриманого результату

|  |  |  |
| --- | --- | --- |
| **Файл** | **Кількість інформації у вихідному файлі** | **Кількість інформації у закодованому файлі bz2** |
| Привид | 1128 | 1336 |
| REST | 547 | 752 |
| Terraria | 1105 | 1242 |

Кількість інформації у закодованому bz2 файлі в середньому в 1.2 рази більше чим у вихідному файлі.

**Висновок:** під час виконання роботи було досліджено кількість інформації у тексті, розглянуто алгоритм кодування інформації Base64, і на основі алгоритму було створено програму кодування інформації в Base64. Правильність кодування власноруч написаної програми було перевірено за допомогою вже існуючих засобів декодування інформації з Base64.