

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

Лабораторна робота №1
Дослідження кількості інформації при різних варіантах кодування

Виконав студент
2 курсу СА-КІ
Глушко Гліб
[github](#)

Київ 2019

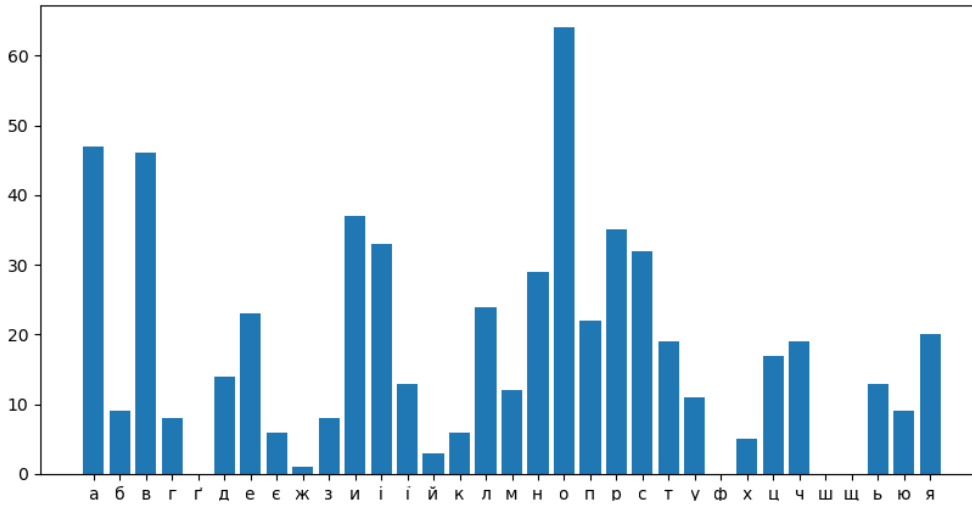
1. Дослідження кількості інформації в тексті

1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування

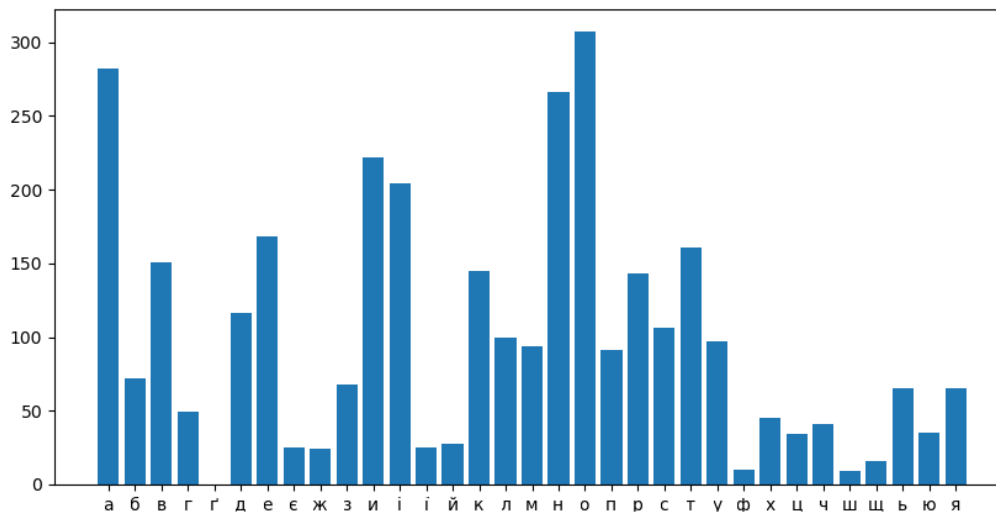
Я вибрав тексти:

- 1) Пісня Олега Винника – Вовчиця
 - 2) Стаття про петицію щодо легалізації медичного канабісу, та плюси і мінуси канабісу
 - 3) Докер, основні можливості і процеси
2. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
 - a. обраховує частоти (імовірності) появи символів в тексті
 - b. обраховує середню ентропію алфавіту для даного тексту
 - c. виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
 - d. виводить на екран значення частот, ентропії та кількості інформації

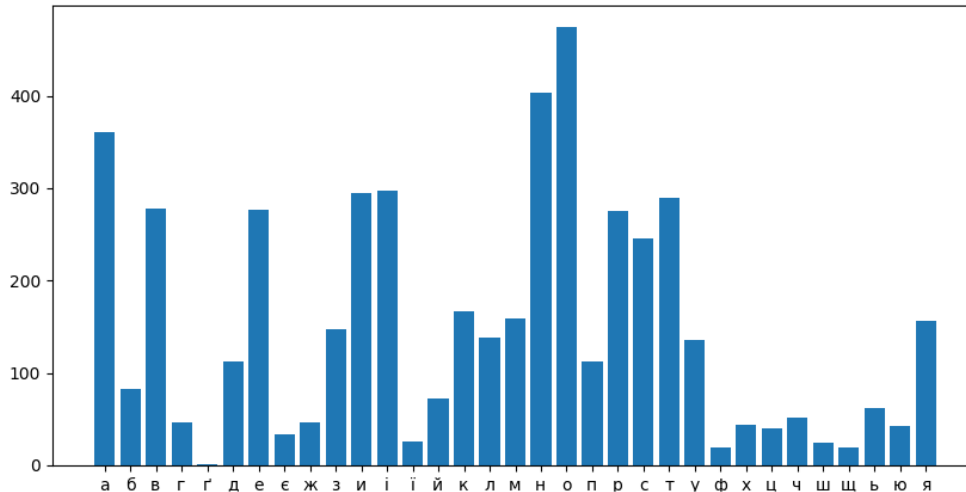
```
vinnyk.txt
entropy = 4.4891538663370945,
letters_count = 585,
total_symbols_count = 792,
file size = 1464,
predicted_file_size = 1313.0775059036002,
```



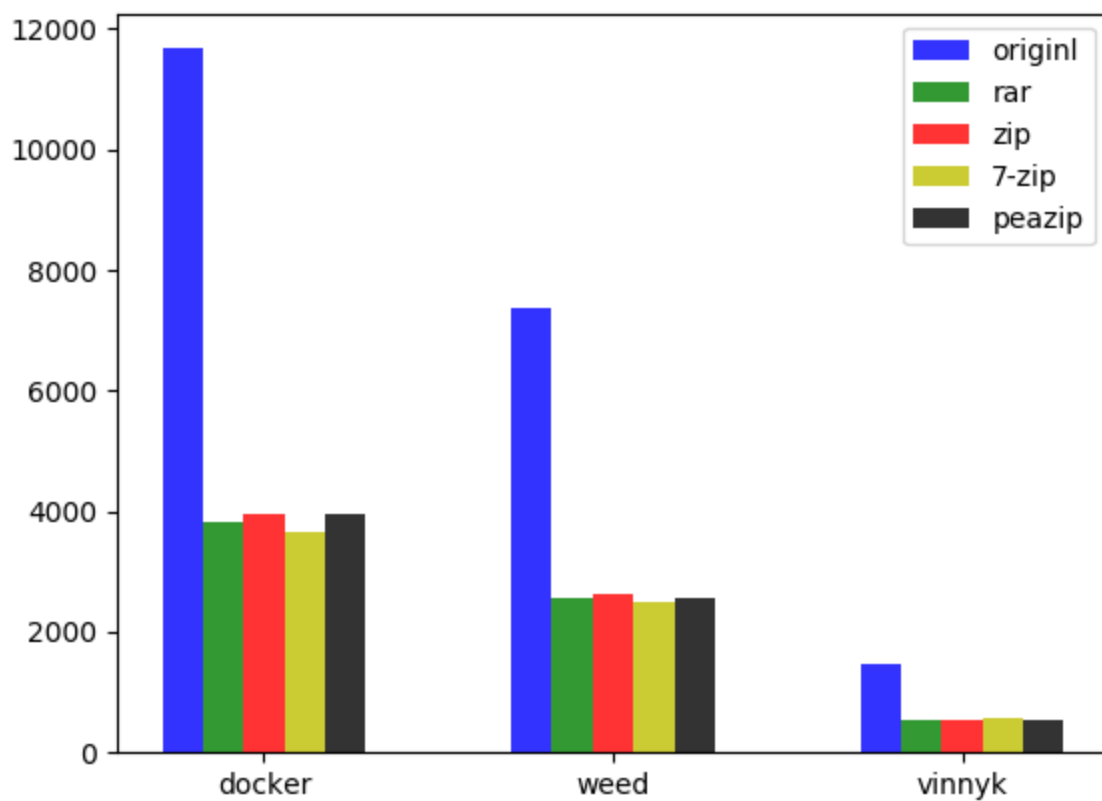
```
weed.txt
entropy = 4.5615542393162976,
letters_count = 3264,
total_symbols_count = 4021,
file size = 7381,
predicted_file_size = 7444.456518564198,
```



```
docker.txt
entropy = 4.539268288356095,
letters_count = 4938,
total_symbols_count = 6631,
file size = 11675,
predicted_file_size = 11207.4534039512,
```



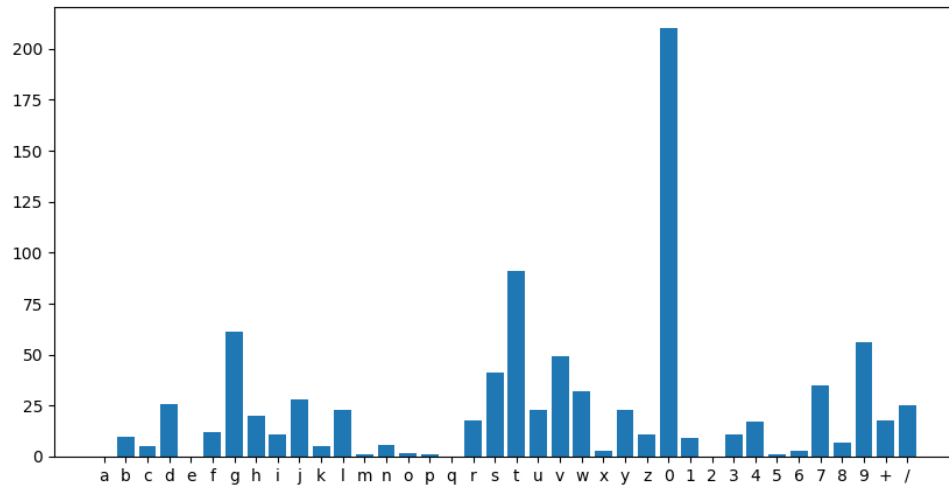
3. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення). Порівняйте результуючі обсяги архівів з обчисленою кількістю інформації та **наведіть у звіті висновки** щодо кореляції цих величин для обраних вами файлів (яка відмінність, що вийшло більше і чому)



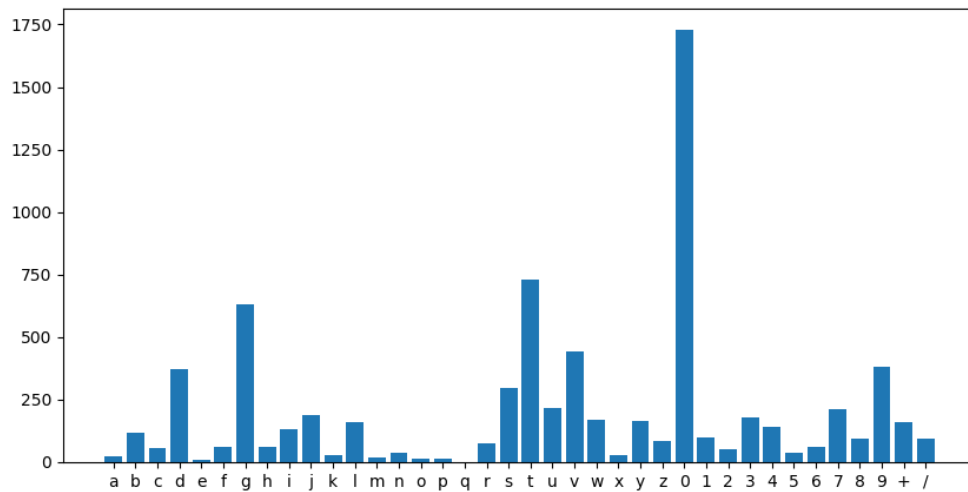
[illegible]

2. Закодуйте в Base64 обрані вами текстові файли

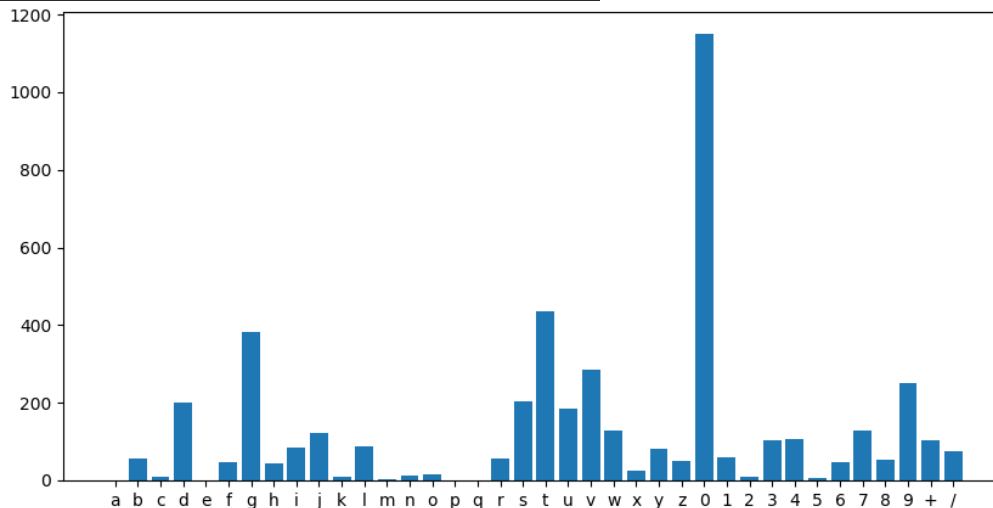
```
vinnyk.txt
True
entropy = 4.196306592113332,
letters_count = 894,
total_symbols_count = 1896,
file size = 1945,
predicted_file_size = 1875.7490466746592,
```



```
docker.txt
True
entropy = 4.246146498023844,
letters_count = 7372,
total_symbols_count = 15516,
file size = 15565,
predicted_file_size = 15651.295991715888,
```



```
weed.txt
True
entropy = 4.13324380032741,
letters_count = 4616,
total_symbols_count = 9808,
file size = 9857,
predicted_file_size = 9539.526691155663,
```



По-перше, я змінив алфавіт на англійський, по-друге, додав до алфавіту символи та цифри, які я використовув у кодуванні в **base64**, тому що, якщо рахувати лише символи, які є літерами, то виходить неправильне значення ентропії і неправильний результат ентропії, тому що текст складається з великої кількості цифр і символів, ними не можна знехтувати, як у текстах на оригінальній мові. Можна побачити, що цифра 0 дуже часто використовується

- a. Порівняйте отримане значення з кількістю інформації вихідного файлу
- b. Зробіть висновки з отриманого результату

У закодованому вигляді, розміри файлів збільшились у 1.33 рази, тому що тепер для 1 байт замінюється на символи, розміри яких теж 1 байт, але 1 символ кодує лише 6 біт. Після кодування, я обрахував ентропію і кількість інформації, і кількість інформації теж пропорційно збільшилась

Висновок: Під час лабораторної роботи, було досліджено і розроблено алгоритм кодування у base64, було досліджено, що українські літери займають два байти у кодуванні UTF-8, було порівняно різні архіватори і результати архівування. Досліджено явище ентропії і вираховано кількість інформації, яка приблизно співпадає з реальними розмірами. Неточності можна пояснити, тим, що ми не брали до уваги деякі, символи, а лише алфавіт.