**Київський національний університет імені Тараса Шевченка**

**факультету радіофізики, електроніки та комп’ютерних систем**

## Лабораторна робота №1

**Тема:** «**Дослідження кількості інформації при різних варіантах кодування.**»

Роботу виконав

студент 3-го курсу  
КІ-СА

Куриний Дмитро Андрійович

Київ 2020

**Лабораторна робота № 1**

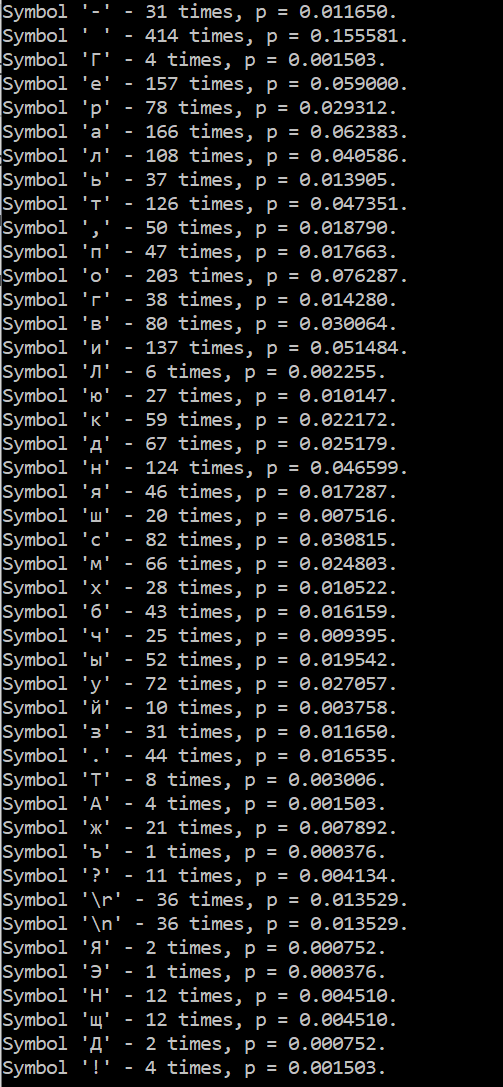
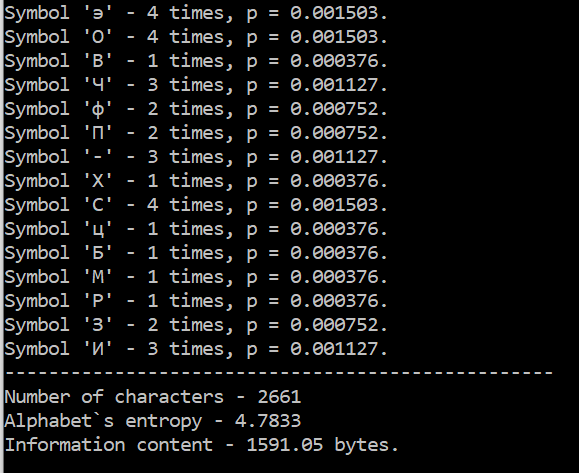
**Дослідження кількості інформації при різних варіантах кодування**

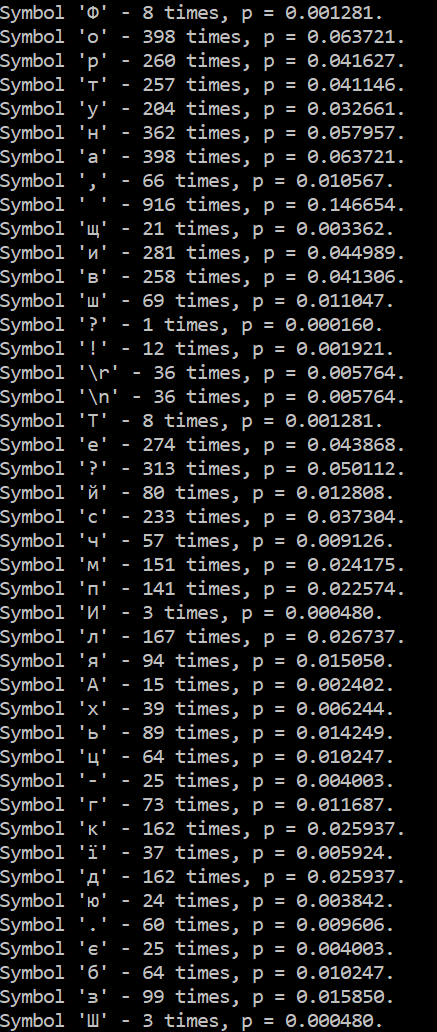
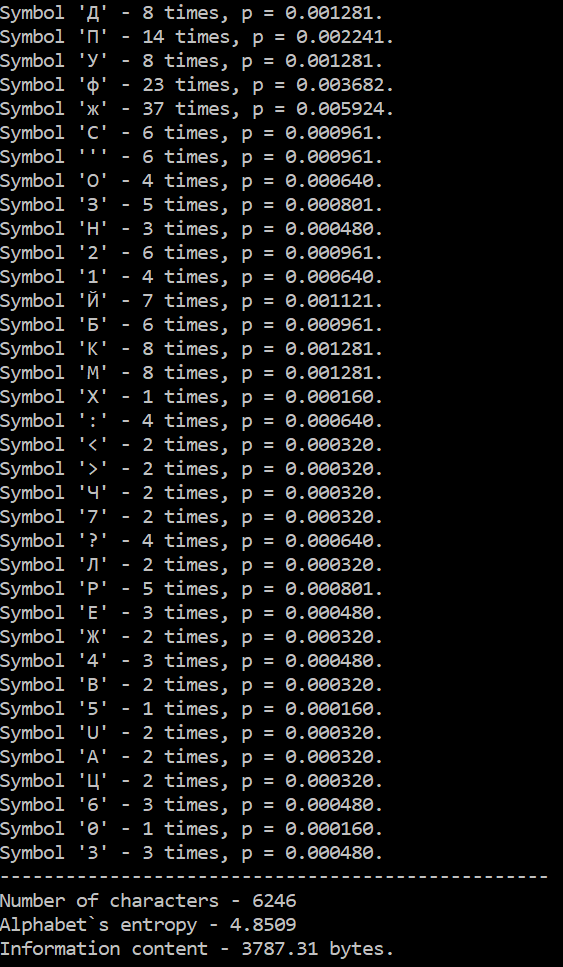
**Мета роботи:** дослідити імовірнісні параметри української мови для оцінки кількості інформації текстів. Дослідити вплив різних методів кодування інформації на її кількість.

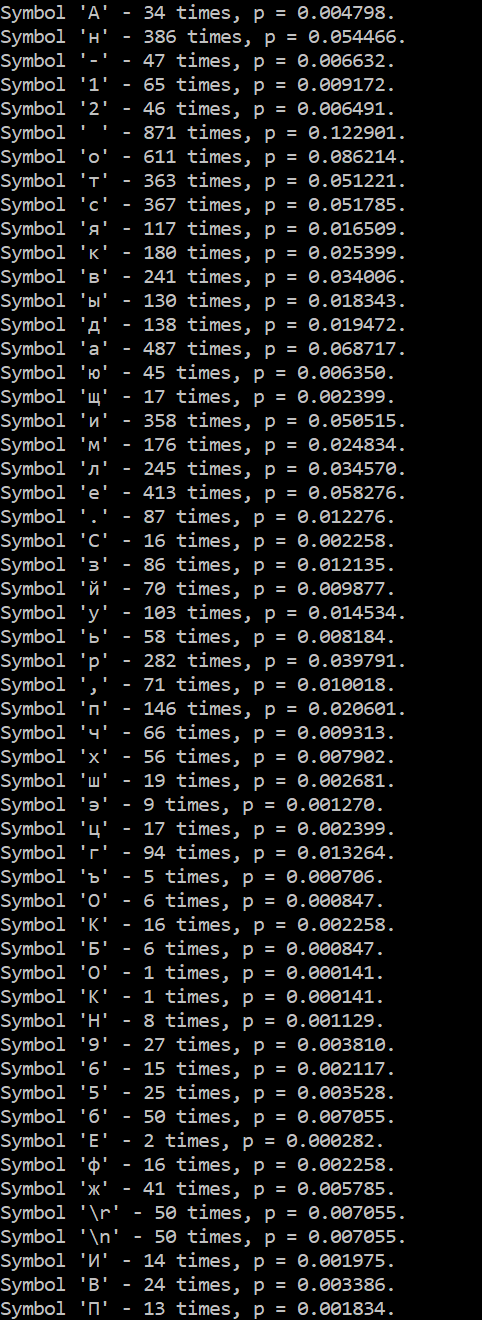
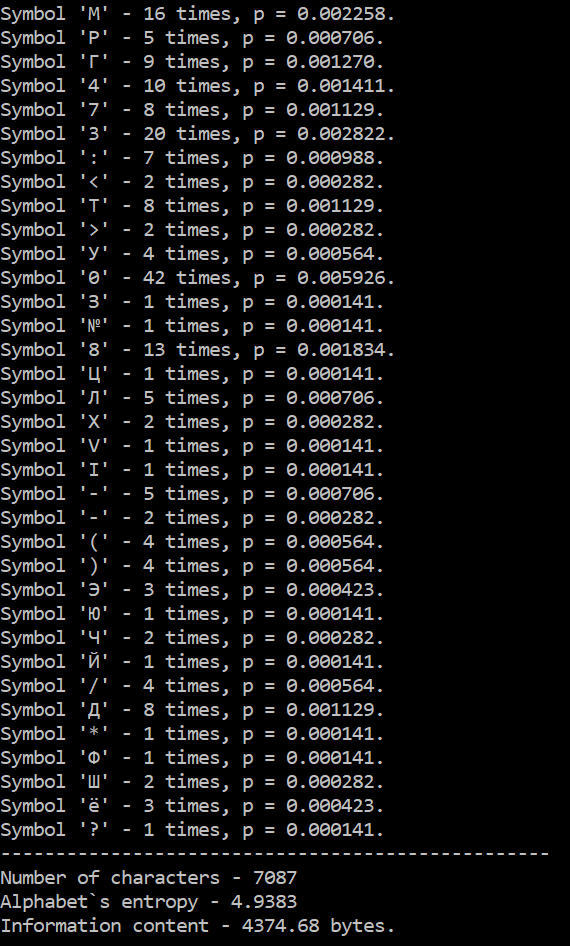
**Хід виконання лабораторної роботи:**

1. **Дослідження кількості інформації в тексті**
2. 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;
   3. виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів;
   4. виводить на екран значення частот, ентропії та кількості інформації.

**text1.txt**

**text2.txt** 

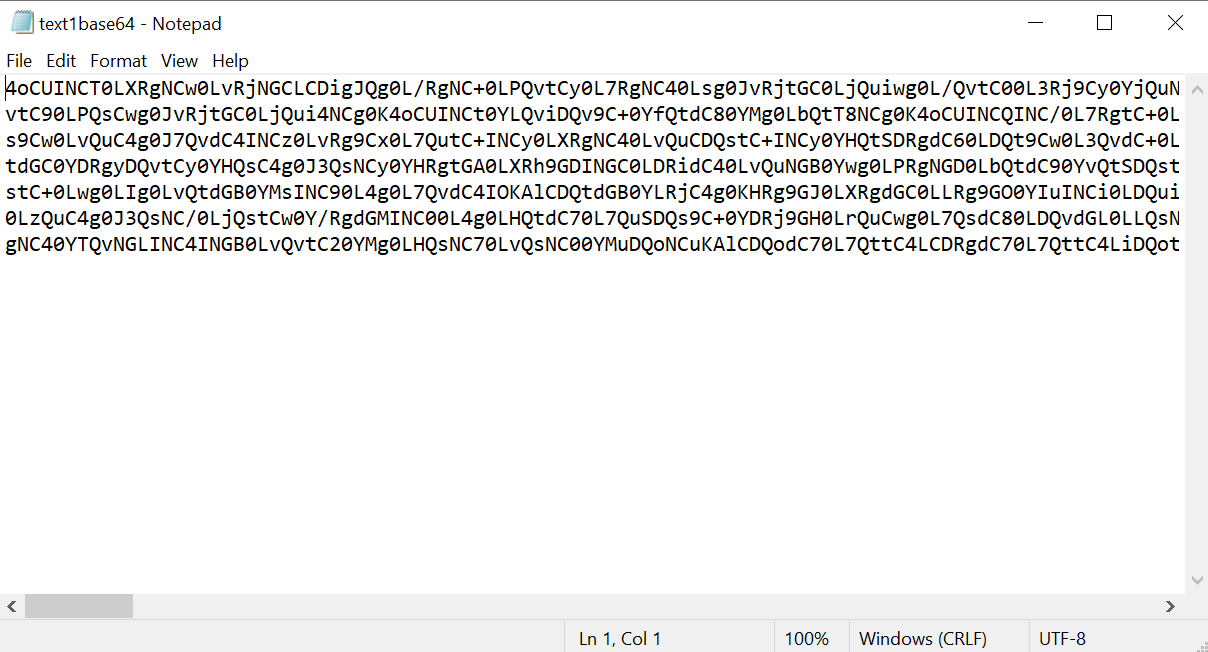
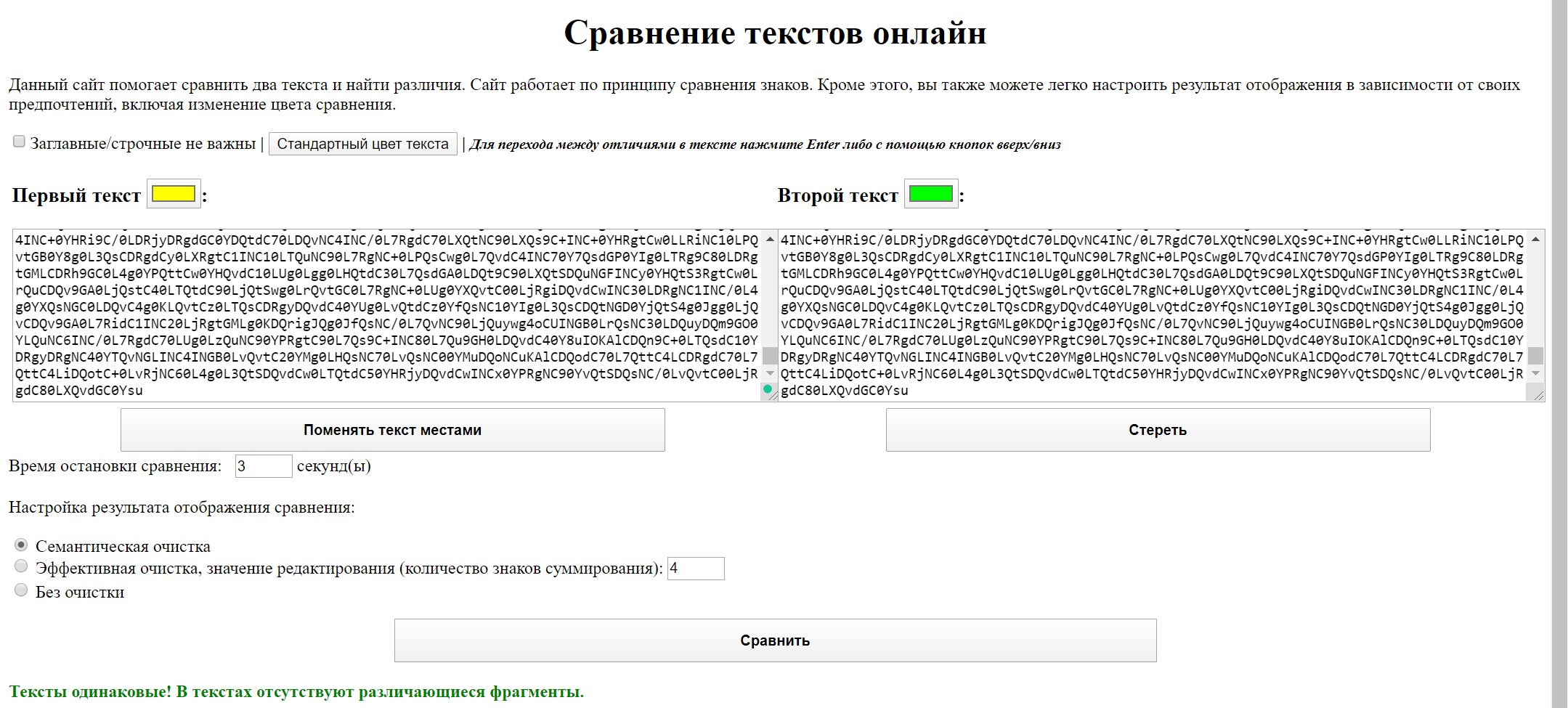
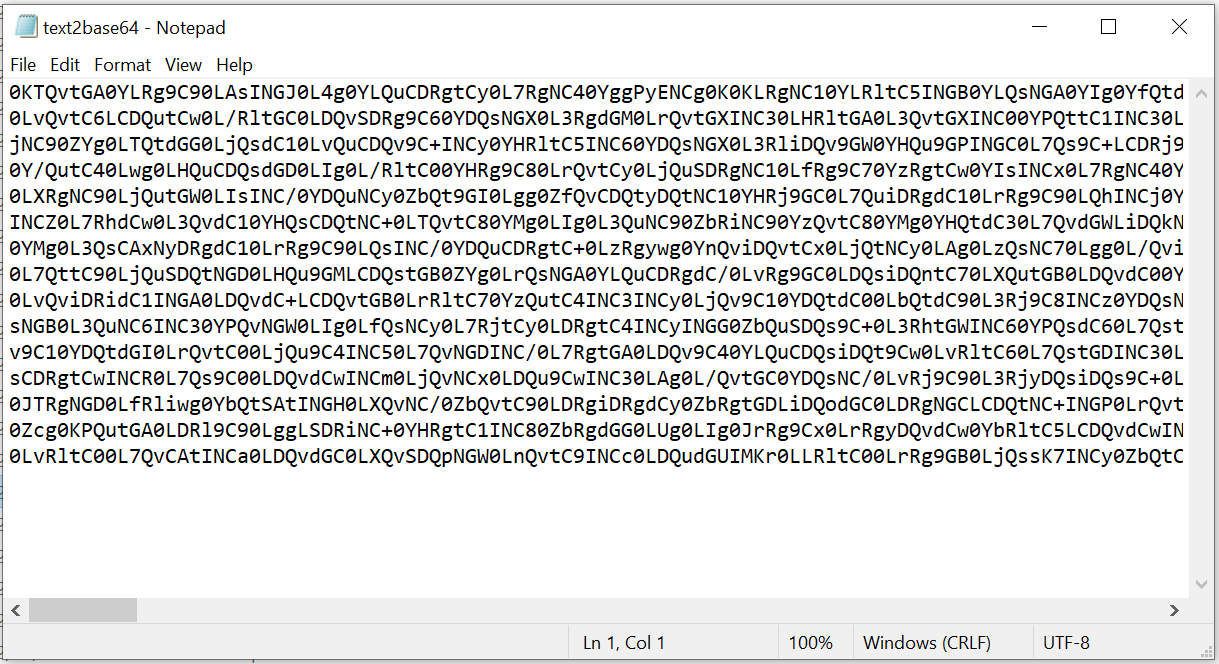
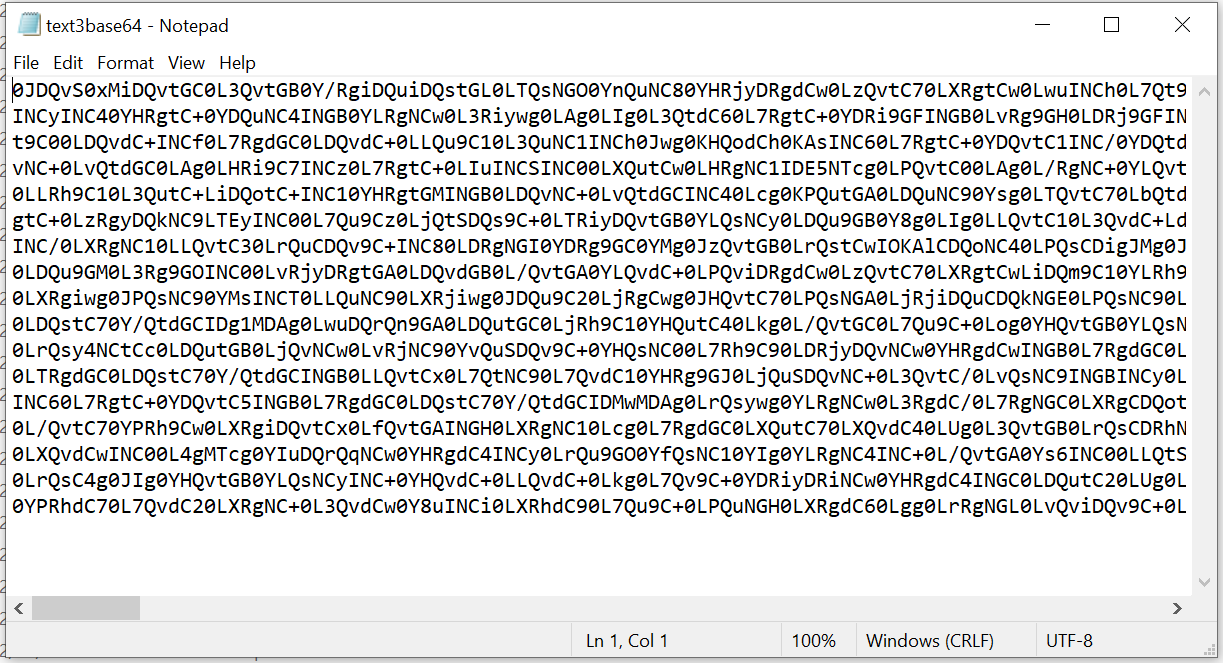
**text3.txt** 

1. 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).

|  |  |  |  |
| --- | --- | --- | --- |
| Файл | **text1.txt, byte** | **text2.txt, byte** | **text3.txt, byte** |
| **.zip** | 1951 | 3784 | 4103 |
| **.gzip** | 1890 | 3790 | 4115 |
| **.xz** | 1828 | 3584 | 3860 |
| **.7z** | 1895 | 3654 | 3929 |
| **.bz2** | 1464 | 3050 | 3385 |
| **Size** | 4755 | 11303 | 12709 |
| **Info. Content** | 1591 | 3787 | 4375 |

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).

Результуючі обсяги архівів можна бачити на діаграмах. Найкраще показав себе алгоритм .**bz2**, найгірше - **.zip**. Оскільки ми застосовували стиснення до текстового файлу, ми домоглися стиснення до 20-40% від початкового розміру (текстові файли мають чи не найбільший відсоток стиснення). В ідеалі повинно було бути так, аби розмір файлу дорівнював кількості інформації, адже алгоритми стиснення використовують ідею групування однакових частин, але в реальності ми бачимо різницю цих величин. Підсумовуючи, можна сказати, що формула розрахунку кількості інформації – не ідеальна. Вона не враховує передбачення наступної частини тексту.

1. **Дослідження способів кодування інформації на прикладі Base64.**
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)  
     
   Порівнювався з результатом з довільно обраного сайту для кодування в Base64 (<https://www.browserling.com/tools/file-to-base64>).  
   Результати порівняння (з сайту <http://text.num2word.ru/>):  
   
3. Закодуйте в Base64 обрані вами текстові файли  
     
   

Кількість інформації в закодованих файлах зросла, пов’язано це з особливостями алгоритму кодування base64.

1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли  
   В нашому випадку це архіви з алгоритмом стиснення **.bz2**.

Через те, що розмір архіву в нас менший, при цьому його ентропія більша, через те ми отримуємо приблизно рівну кількість інформації в файлах.

**Висновок:** під час лабораторної роботи було досліджено вплив різних методів кодувань на кількість інформації. Було порівняно різні алгоритми стиснення. Теоретично ознайомився з алгоритмом кодування Base64 та практично його реалізував, досліджено переваги та недоліки цього алгоритму кодування.

<https://github.com/fly4funDK/CS_Lab1>