Київський національний університет імені Тараса Шевченка

Факультет радіофізики, електроніки та комп’ютерних систем

Навчальна дисципліна «Комп’ютерні системи»

**Звіт з лабораторної роботи №1**

на тему «**Дослідження кількості інформації**

**при різних варіантах кодування**»

Роботу виконав

Студент 3 курсу

КІ, група СА

Кравченко В’ячеслав

Васильович

Київ 2019

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

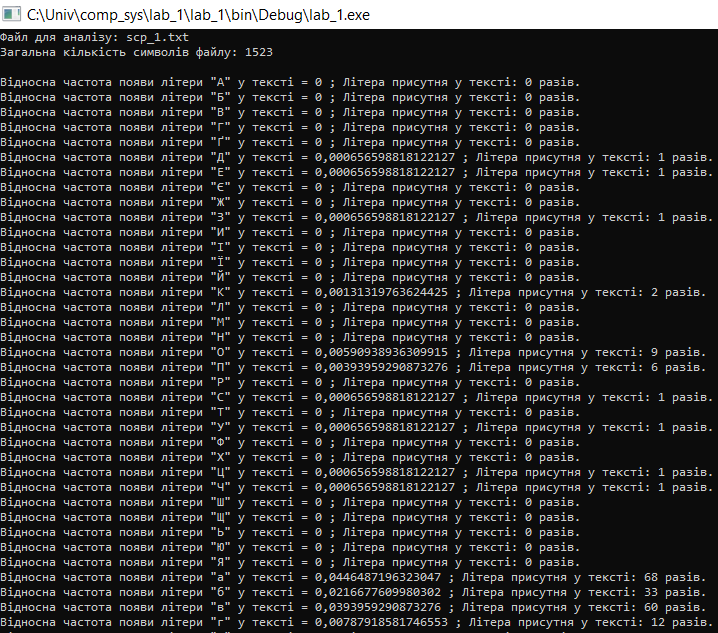
**Хід роботи**

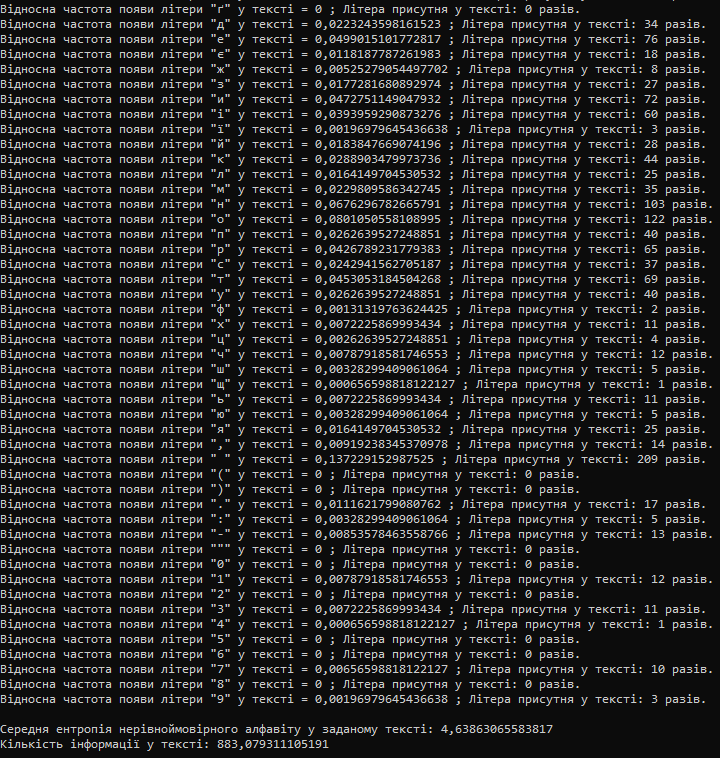
***Дослідження кількості інформації в тексті***

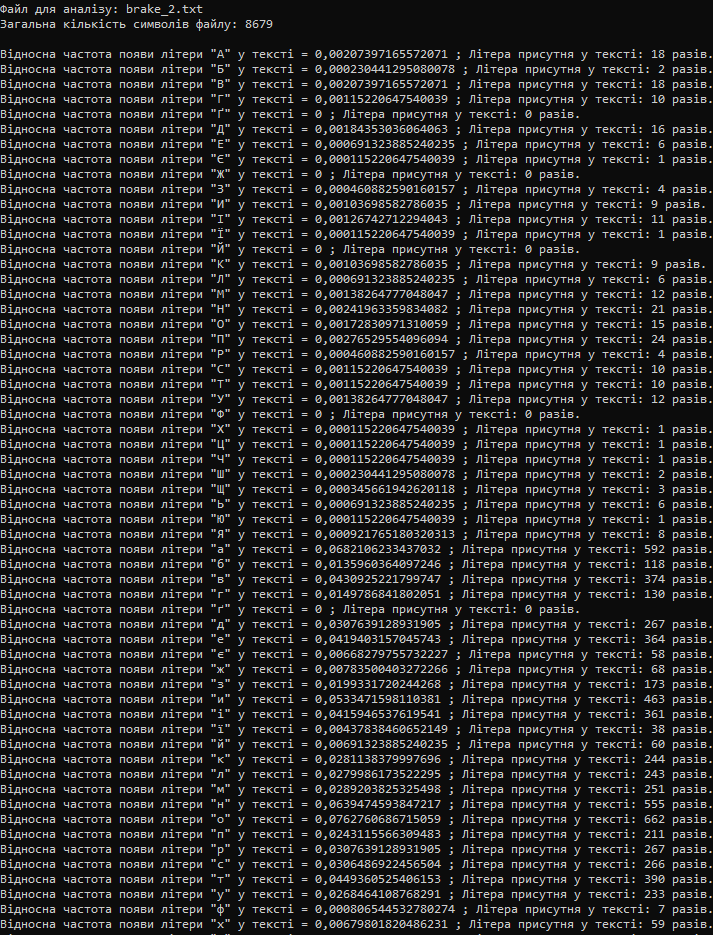
1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (файли також є у репозиторії):
   * [Об’єкт SCP-173](http://scp-ukrainian.wikidot.com/scp-173)
   * [Гальмування двигуном](http://vseznayko.com.ua/jak-galmuvati-dvigunom.html)
   * [Перекладений текст пісні Rick Astley - Never Gonna Give You Up (натисніть, щоб пійматися на рікролл українською)](https://www.youtube.com/watch?v=G510jeWiaV0)
2. 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!

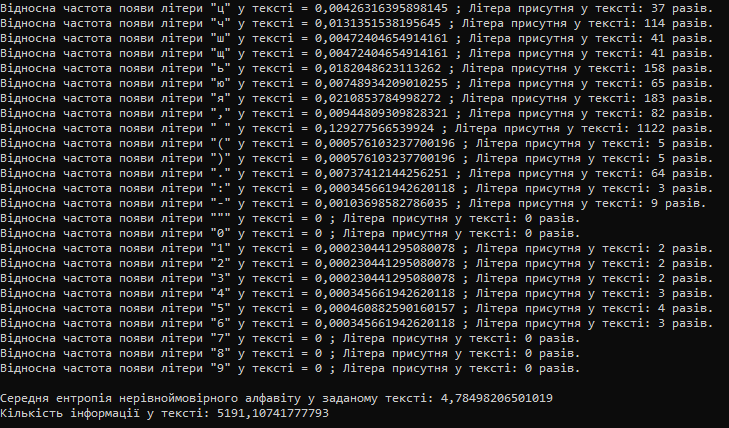
* Заради цього створено мною [цей документ](https://docs.google.com/spreadsheets/d/1lv_-aLfkekkrWQPtk8MoTJPQMLCKwFAy083K3bsuEKc/edit#gid=0)

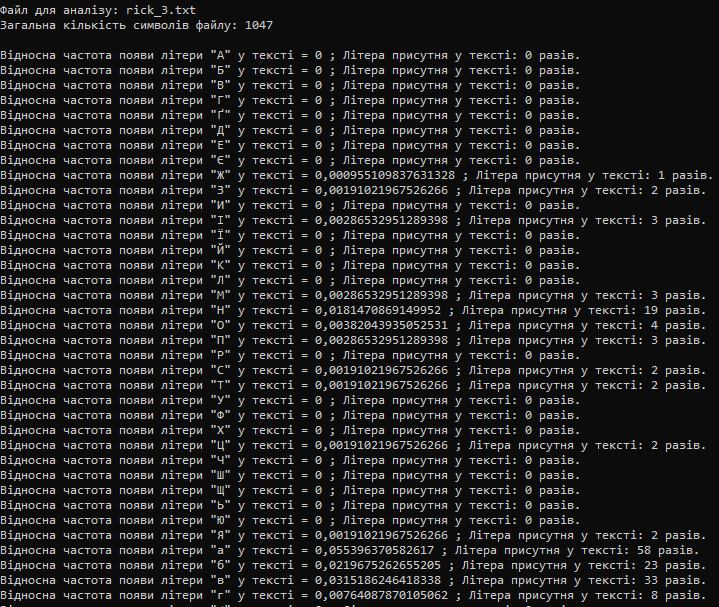
1. Код створеної програми міститься у репозиторії, посилання буде вкінці. Результат роботи програми:

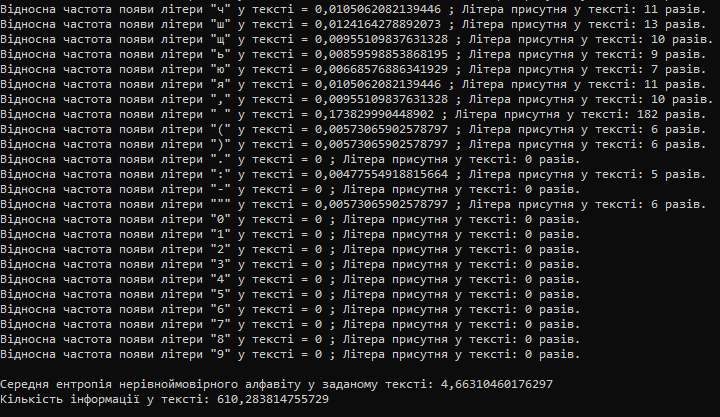












1. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення.

Для цього завдання я використовував WinRAR, WinZip та 7-Zip. Для кожного алгоритму я використовував рівень стиснення Normal, щоб усі були в рівних умовах.

|  |  |  |  |
| --- | --- | --- | --- |
| Файл | scp\_1.txt, Байт | brake\_2.txt, Байт | rick\_3.txt, Байт |
| rar | 1181 | 4083 | 544 |
| zip | 1216 | 4680 | 600 |
| gzip | 1092 | 4554 | 475 |
| xz | 1112 | 4436 | 516 |
| bz2 | 932 | 3685 | 466 |
| Розмір файлу ориг. | 2736 | 16174 | 1977 |
| Кількість інформації | 883 | 5191 | 610 |

1. У результаті ідеального стиснення розмір файлу повинен бути рівним кількості інформації. Але у реальності розміри архівованих файлів у більшості випадків дещо більші за кількість інформації, окрім випадку з великим текстом.

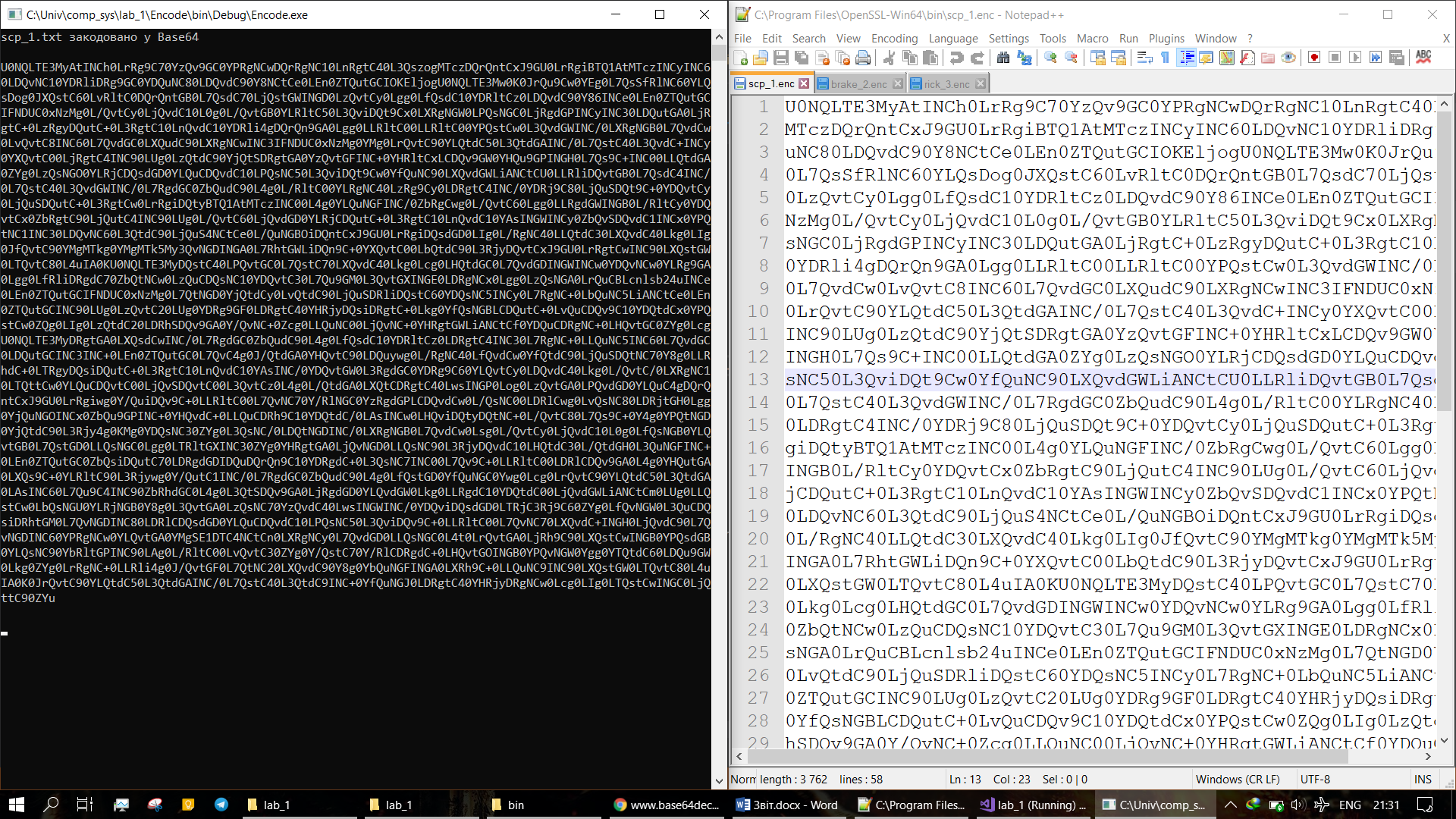
Це відбувається тому, що алгоритми архіваторів побудовані таким чином аби використати повторювані частини тексту. Виходячи з цього, формула розрахунку кількості інформації, використана для програми, не є досконалою, бо вона не враховує передбачення наступного шматочку текста.

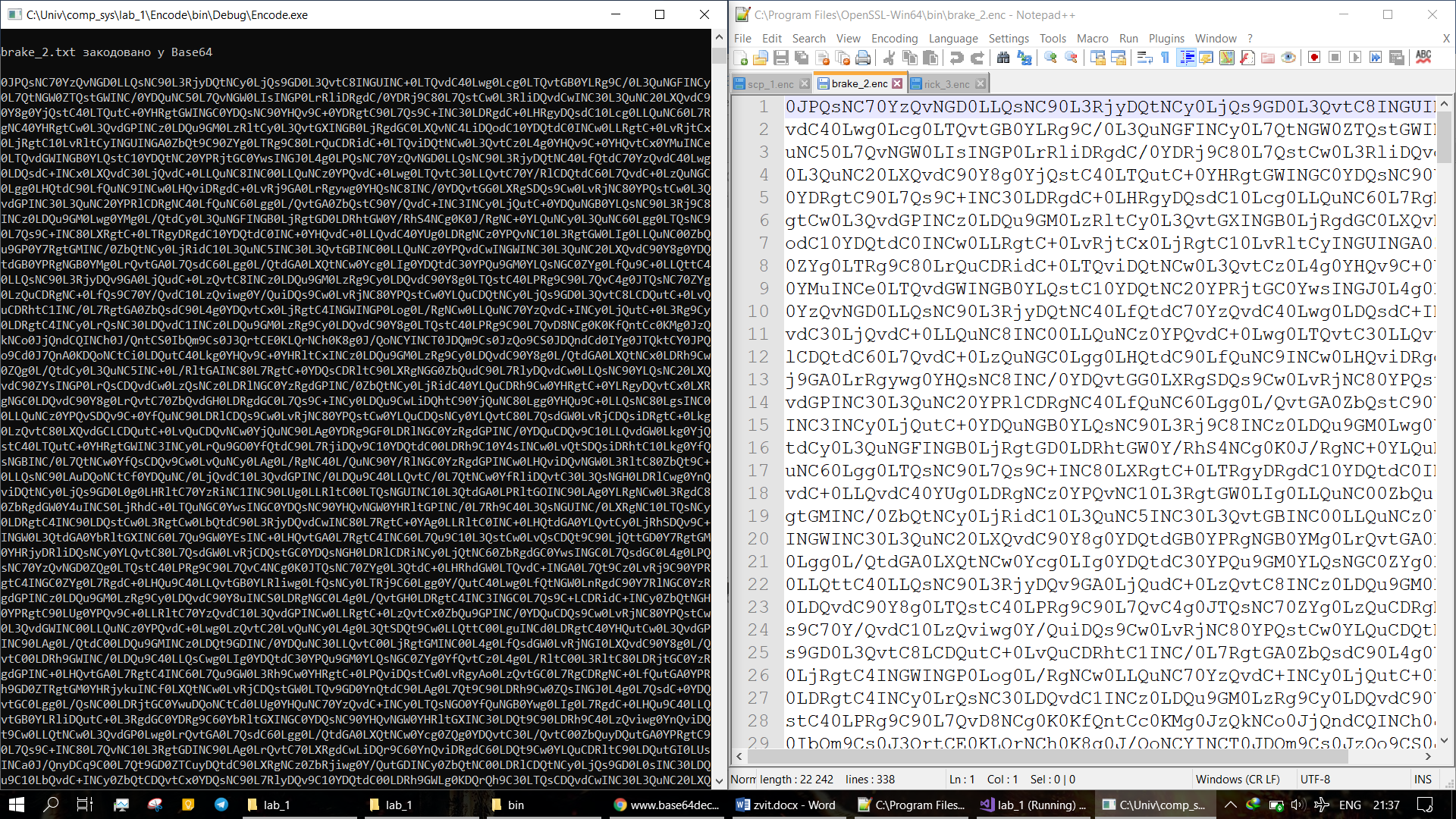
* Хочу звернути уваги на алгоритм bzip2, який виявився найефективнішим у всіх випадках. А також навіть упорався «ідеальним стисненням» у випадку великого файлу, тобто стиснений архів має розмір навіть менший, ніж кі-сть інформації. У випадку з піснею, хоч файл і меншого розміру, ніж текст про SCP-173, але є багато частин повторюваного тексту (як-от приспів)

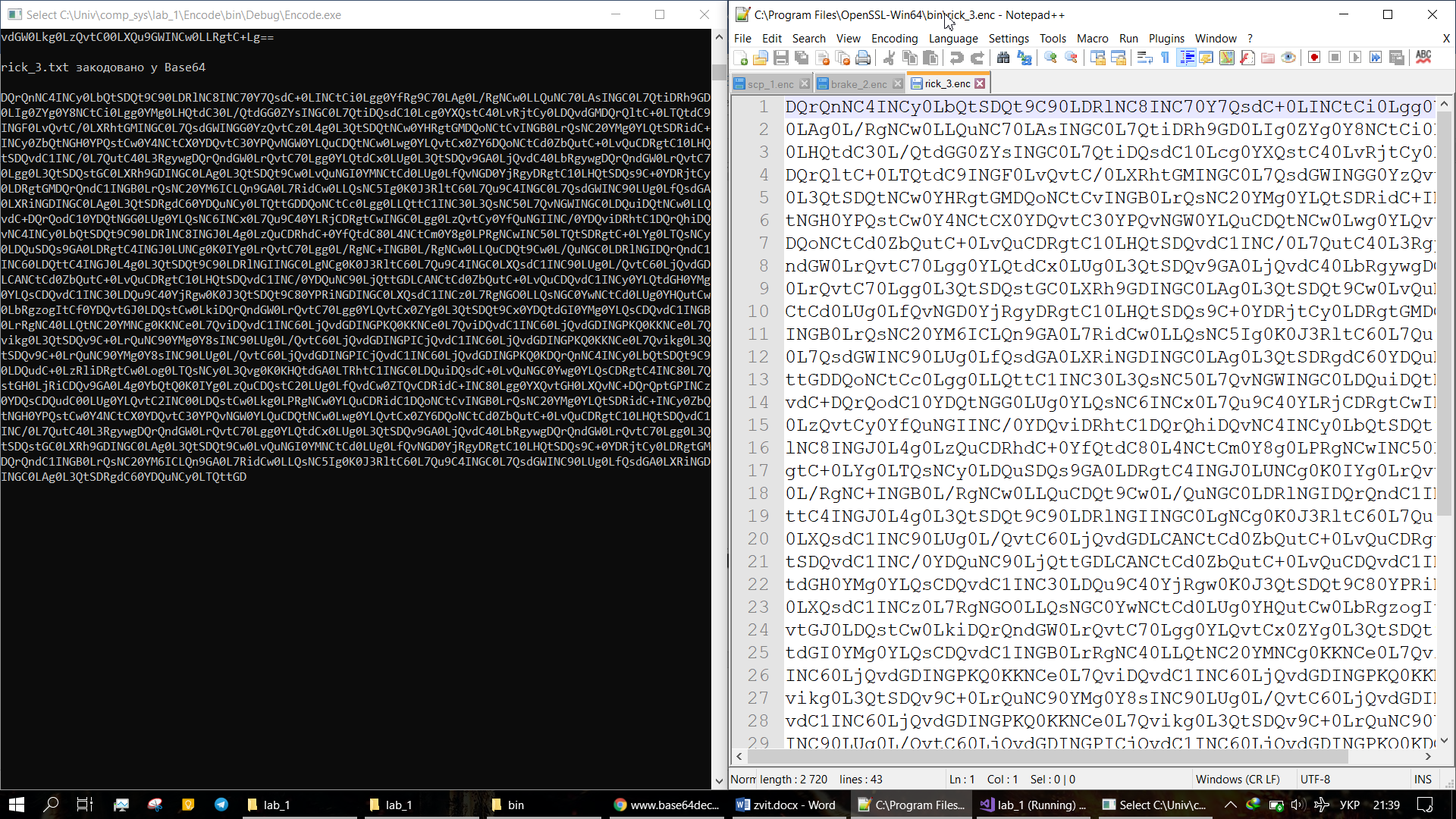
***Дослідження способів кодування інформації на прикладі Base64***

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)

Зліва – закодовано моєю програмою. Справа – використовуючи OpenSSL для Windows







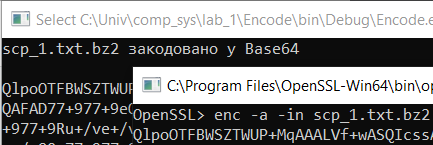
Коректність мого алгоритму також перевірена онлайн кодерами/декодерами.

1. Порівняння текстових файлів та їх версій, закодованих у base64:

Можна легко помітити, що розмір закодованих файлах зріс. Це пов'язано з алгоритмом кодування base64 – перетворення, наприклад 3 октетів (по 8 біт) у 4 секстети (по 6 біт), що збільшує розмір на третину.

1. Найефективнішим алгоритмом стиснення виявився bzip2, тому тепер порівняння його з кодування у base64:

Перевірка коректності:



Отже, ситуація аналогічна з попереднім пунктом.

**Висновок**

У ході виконання лабораторної роботи ознайомився з поняттям ентропії інформації та пов’язаних понять. Теоретичні знання закріпив практично. Також порівняв алгоритми стиснення – обрав кращий з них для випадків, коли треба буде зекономити місце на носії. Теоретично та практично ознайомився з алгоритмом кодування Base64, його перевагами та недоліками.

Код програм, звіт та текстові файли, використані у роботі містяться у репозиторії за [цим посиланням (натисніть мене).](https://github.com/Edocksil/Computer_Systems/tree/master/lab_1)