**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 роботи**

Дослідження кількості інформації в тексті

1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (напр, вірш Тараса Шевченка «Мені тринадцятий минало», «Казка про ріпку» Леся Подерев’янського та специфікацію інтерфейсу PCI)

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!

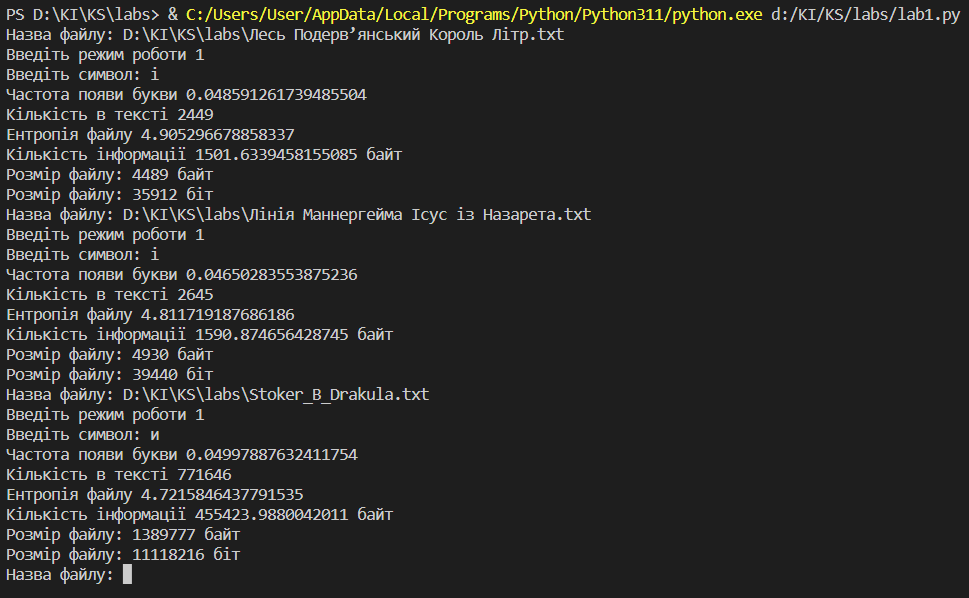
3. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:

a. Обраховує частоти (імовірності) появи символів в тексті

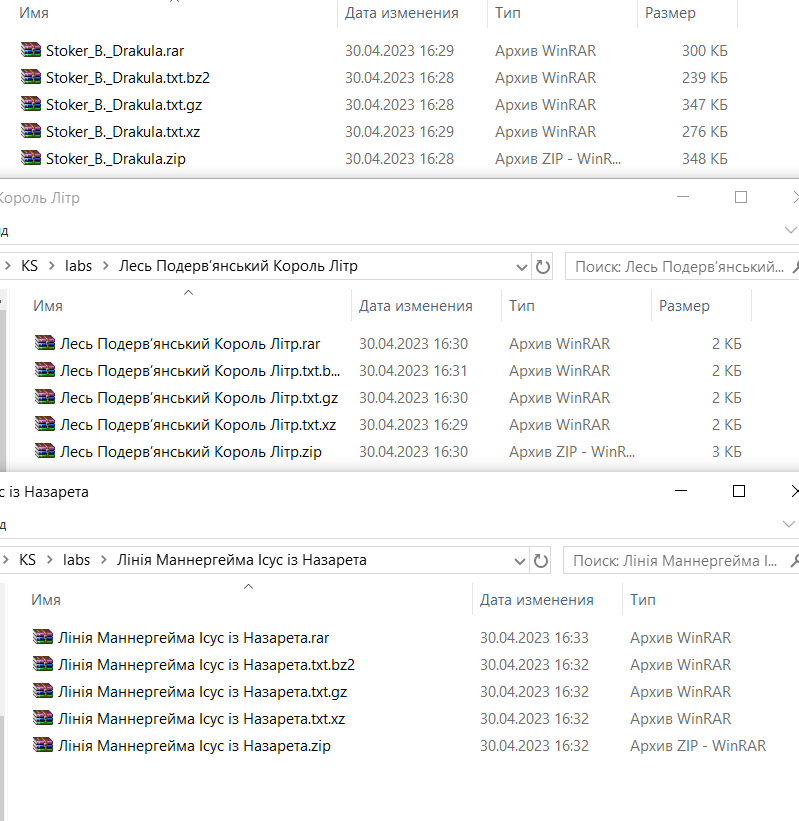
b. Обраховує середню ентропію алфавіту для даного тексту

c.      Виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів

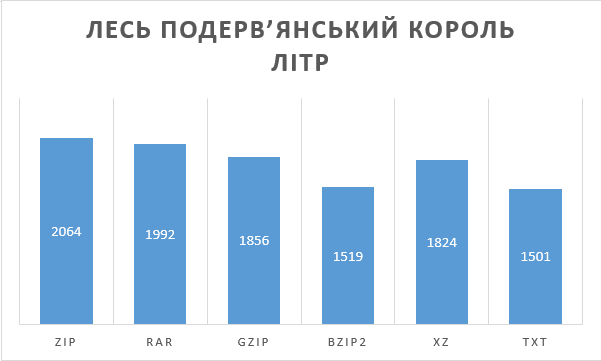
d. Виводить на екран значення частот, ентропії та кількості інформації

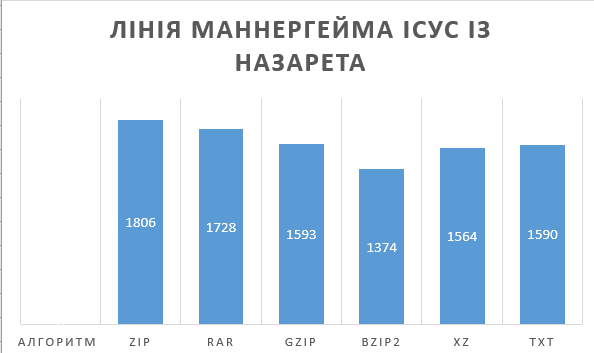


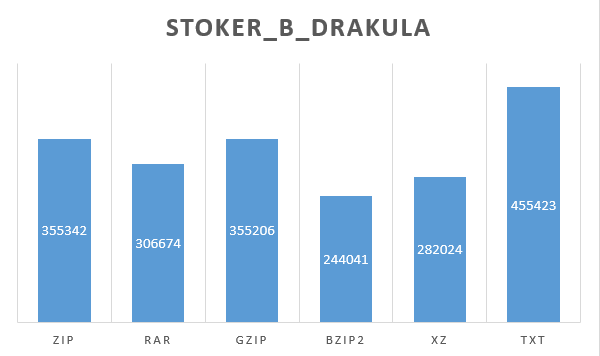
4. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби)



5. 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)





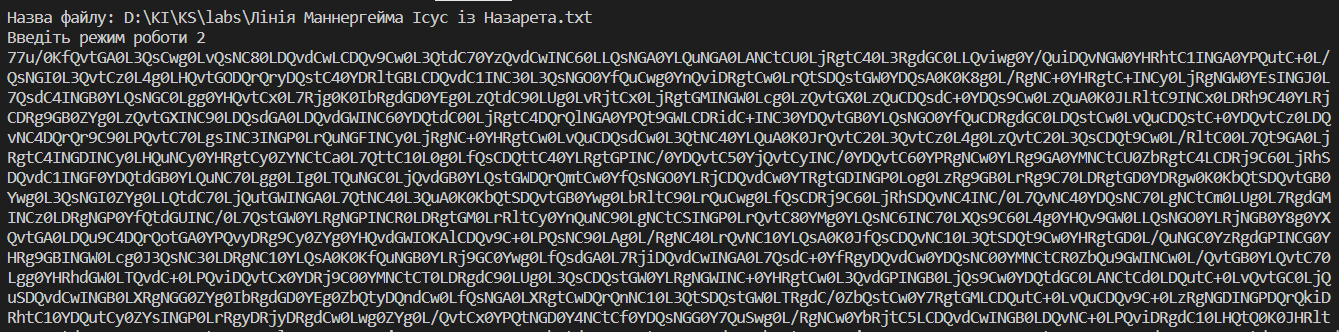


З наведених графіків видно, що найкращим алгоритмом архівації є bzip2, а найгіршим zip.

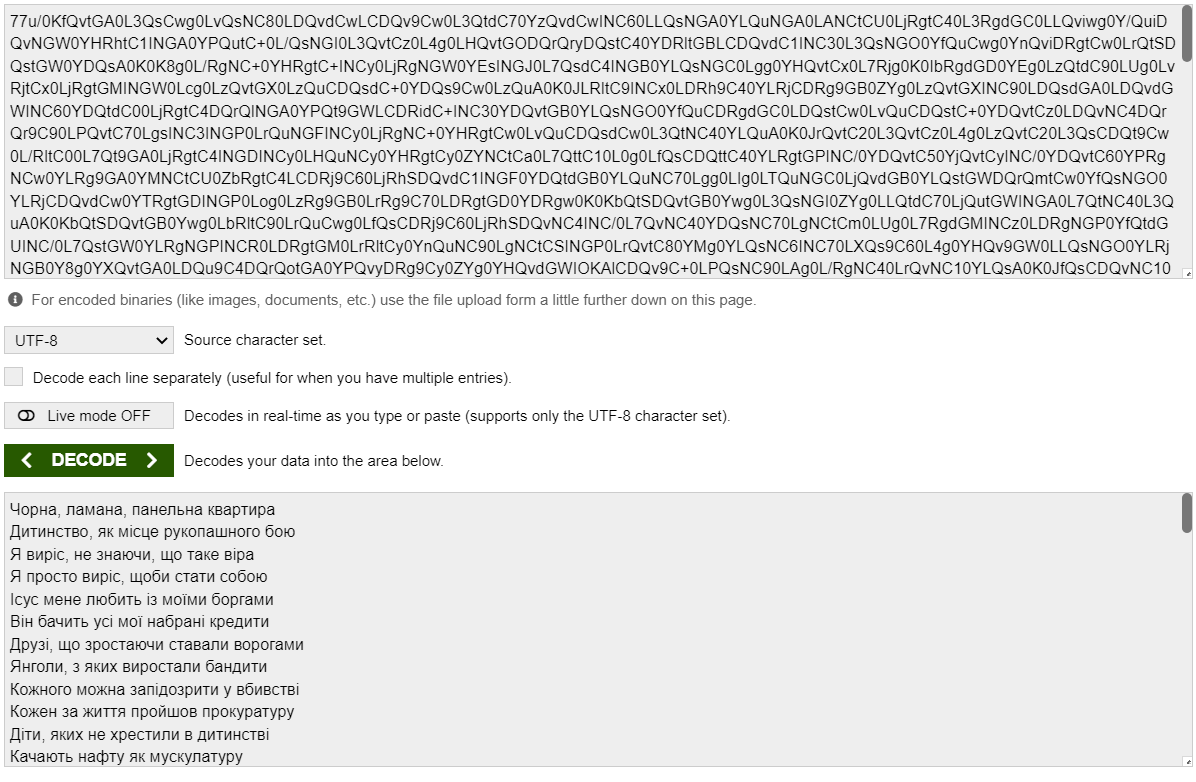
Дослідження способів кодування інформації на прикладі Base64

1. Ознайомтесь зі стандартом [RFC4648](https://datatracker.ietf.org/doc/html/rfc4648)

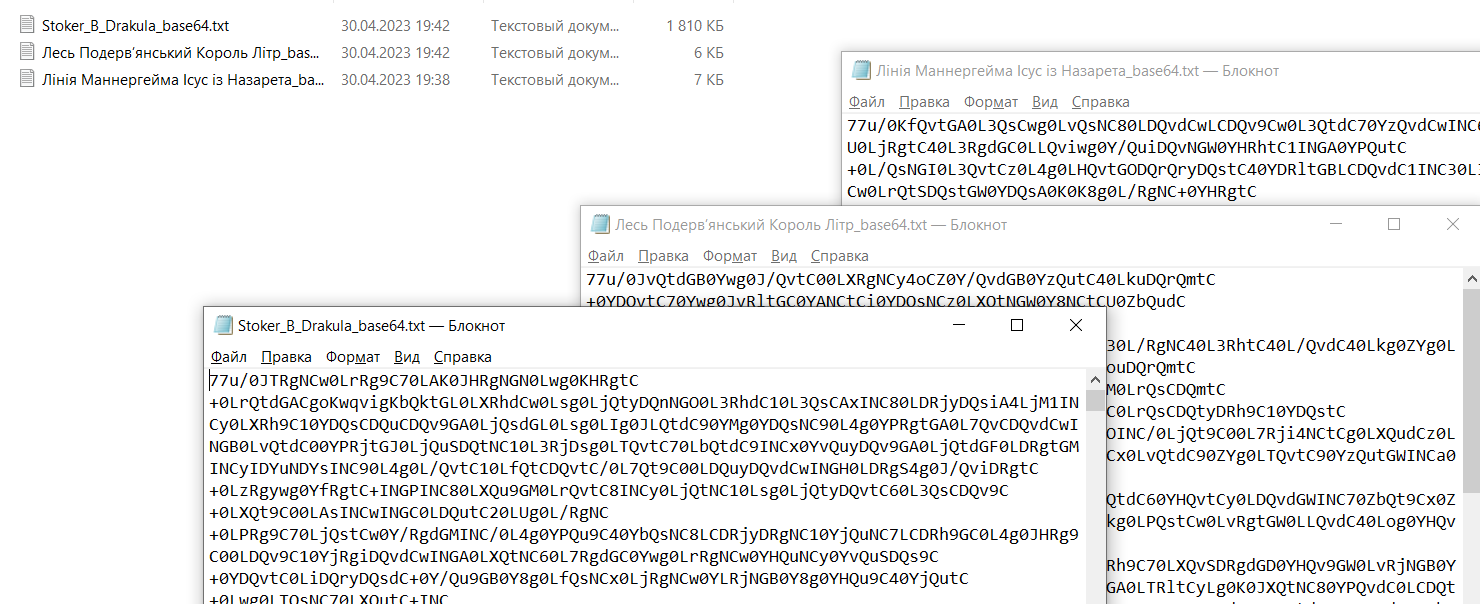
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)



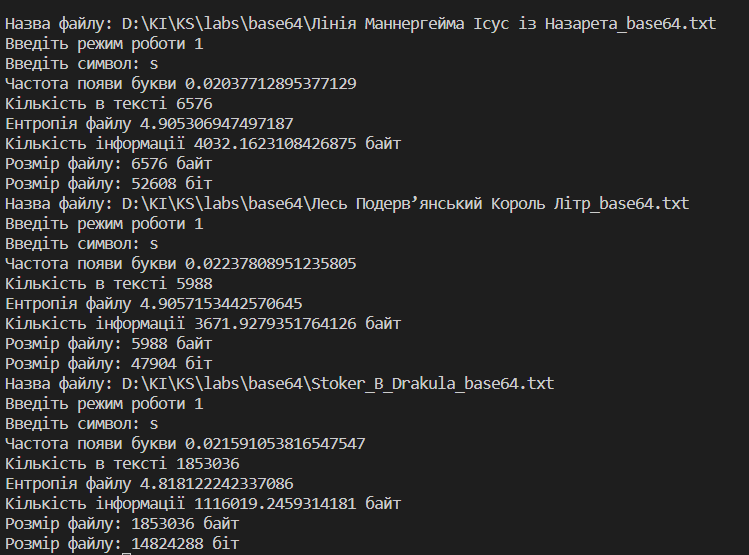
a. Перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (напр, openssl enc -base64)



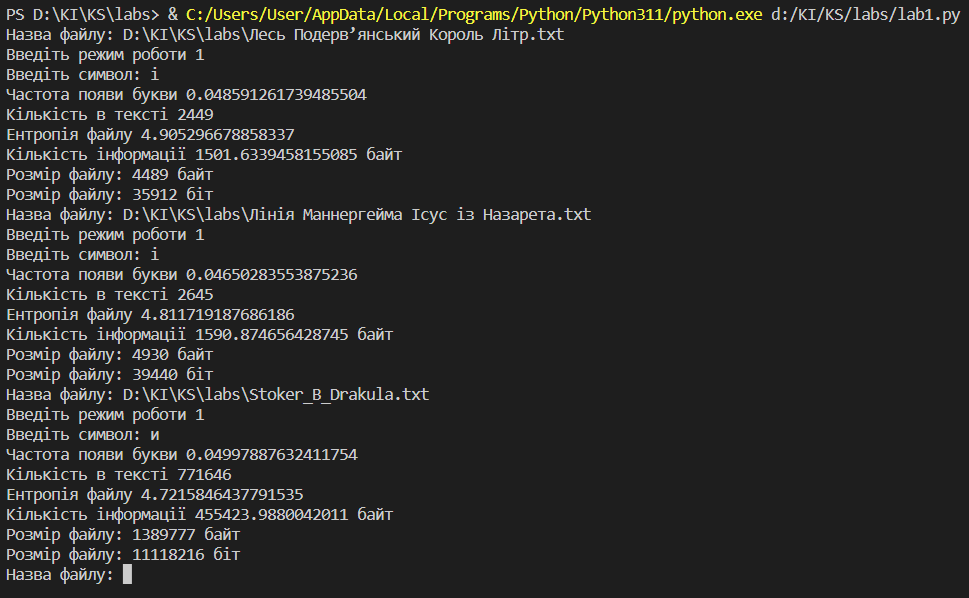
3. Закодуйте в Base64 обрані вами текстові файли

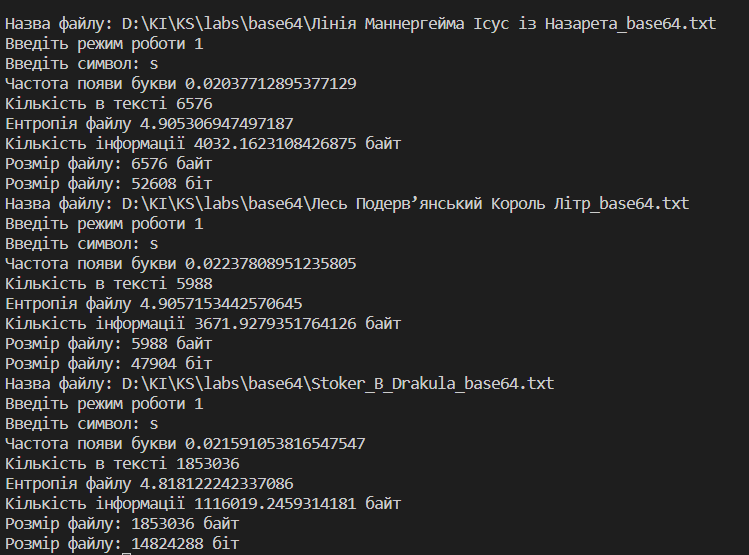


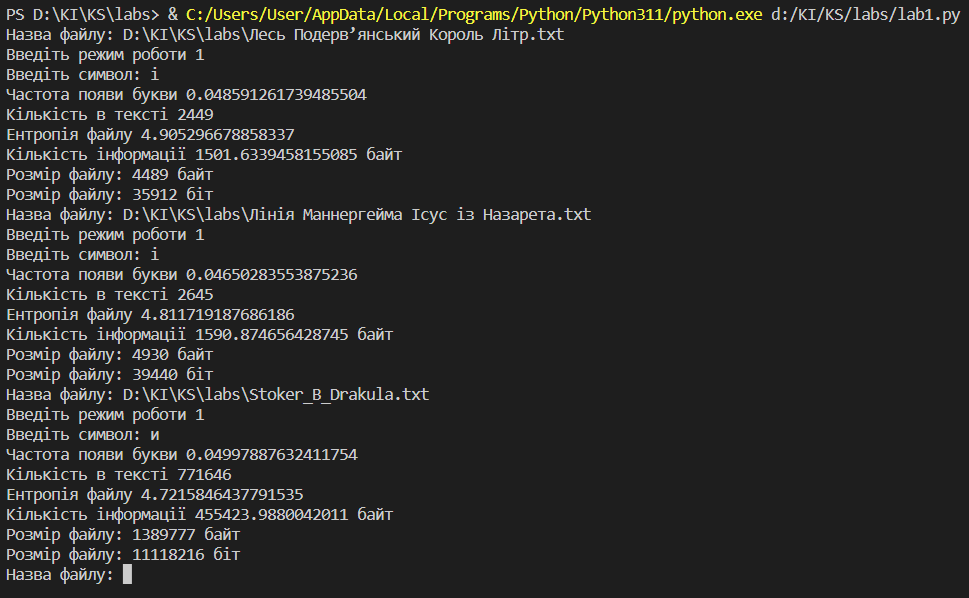
1. Обрахуйте кількість інформації в Base64-закодованому варіанті файлу

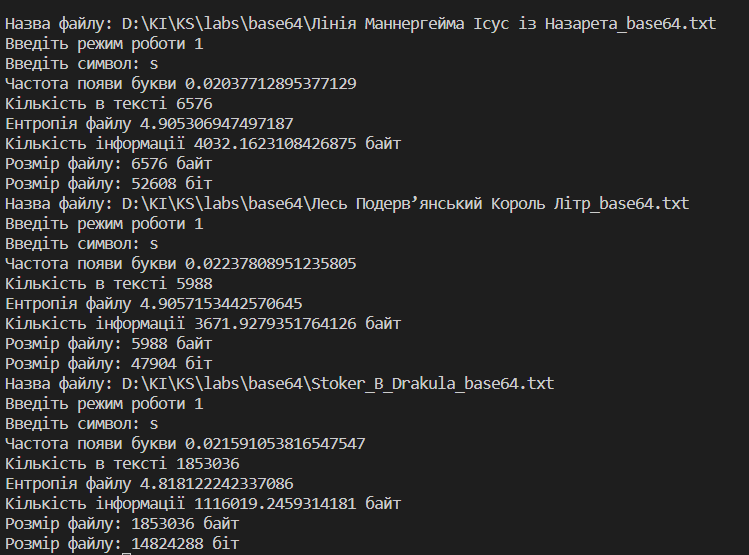


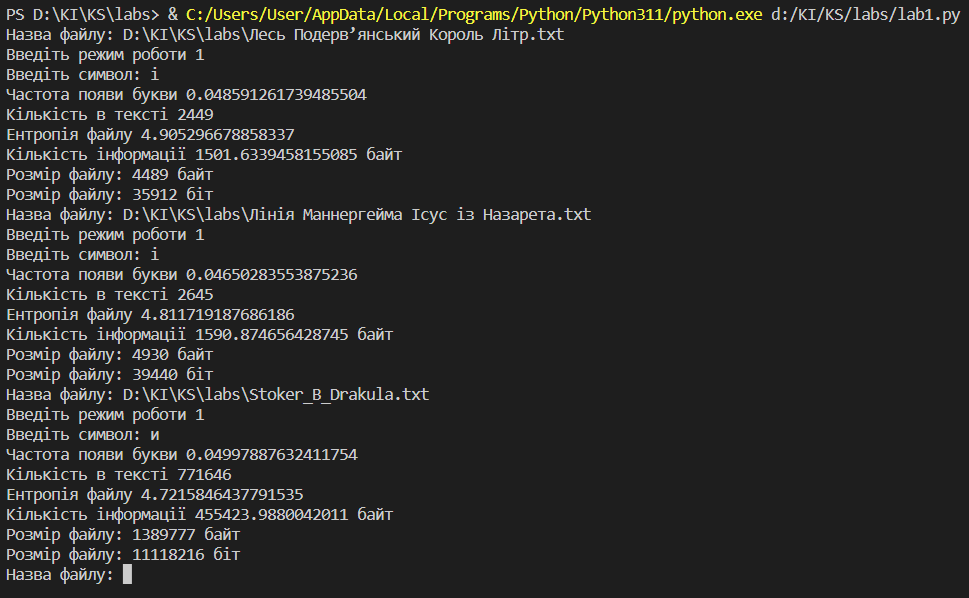
1. Порівняйте отримане значення з кількістю інформації вихідного файлу

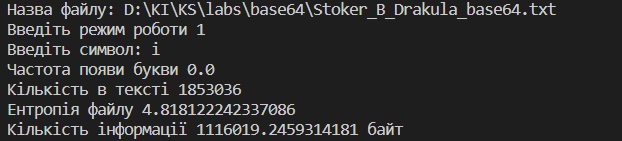








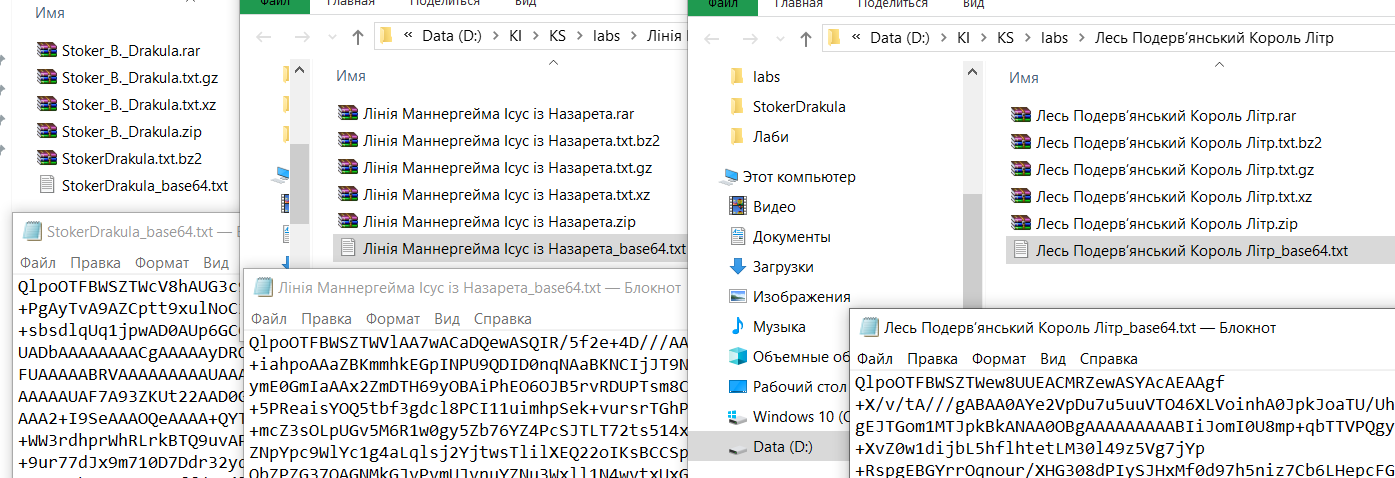




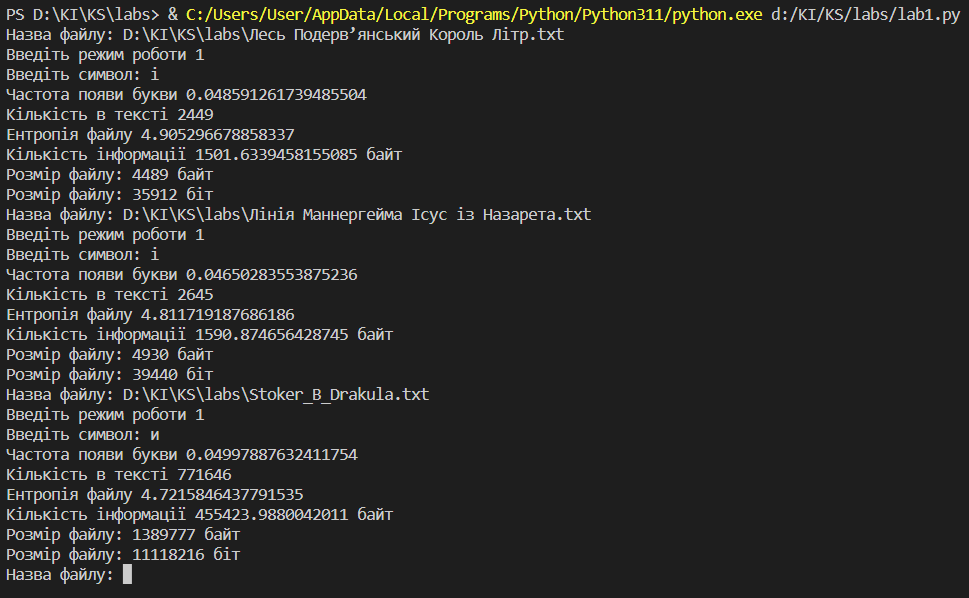
1. Зробіть висновки з отриманого результату

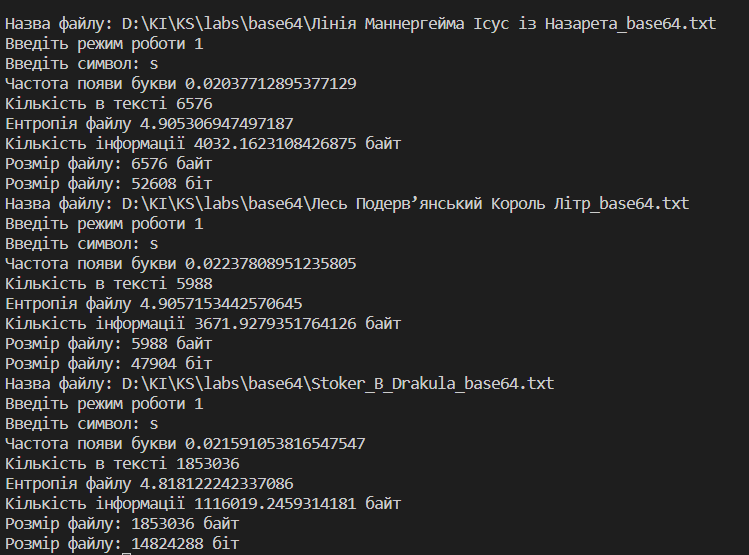
Можна побачити що значення кількості інформації збільшилось приблизно в 2.5 рази.

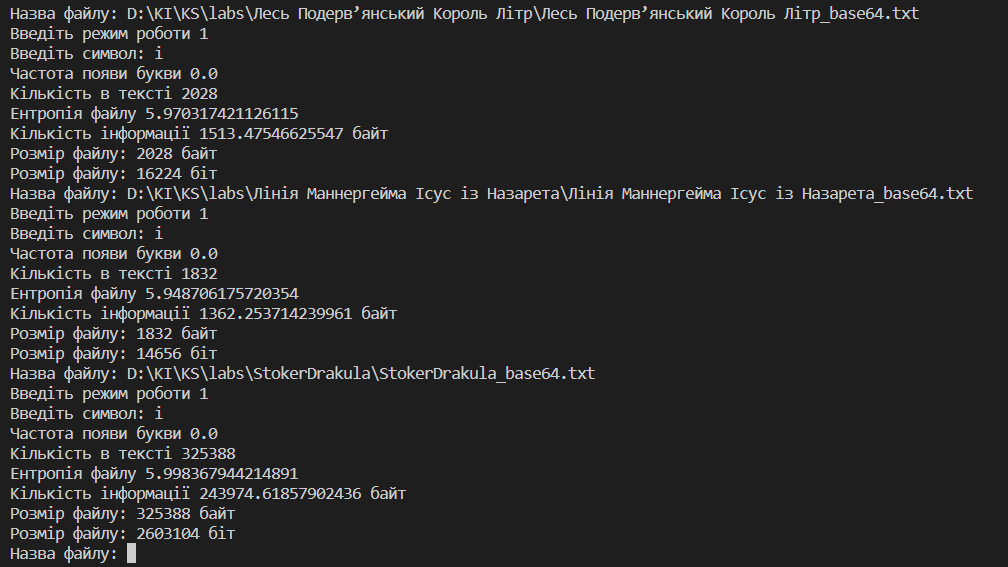
4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли

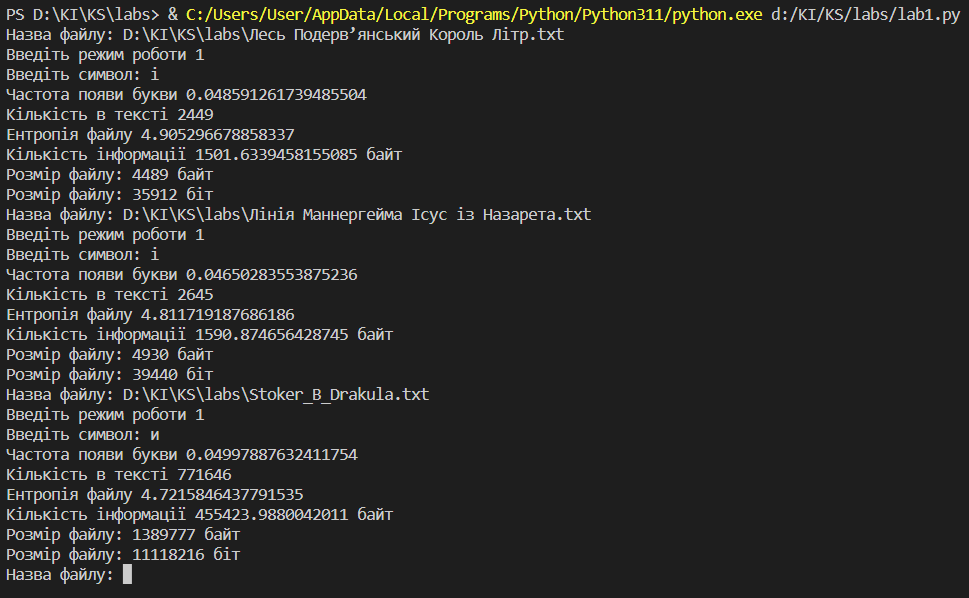


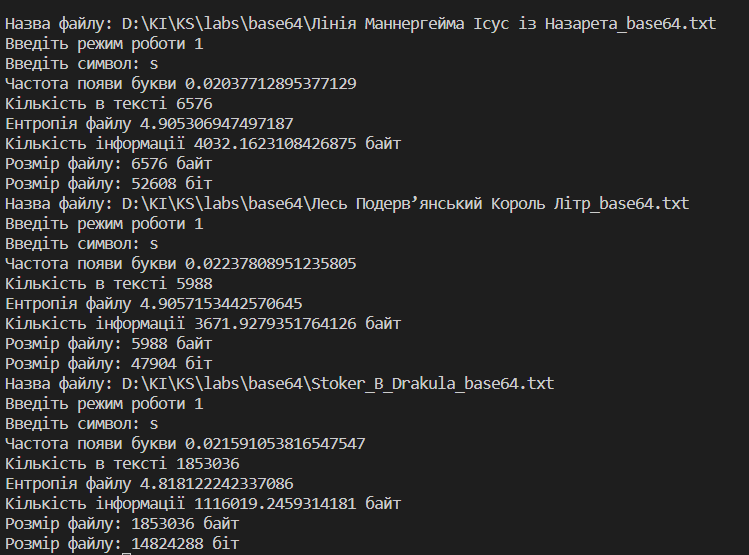
1. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу

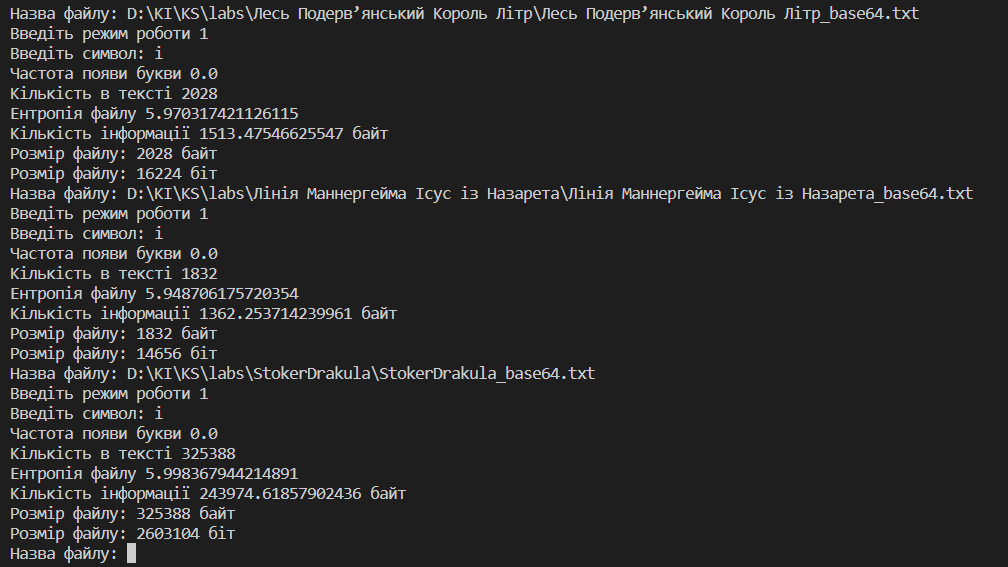


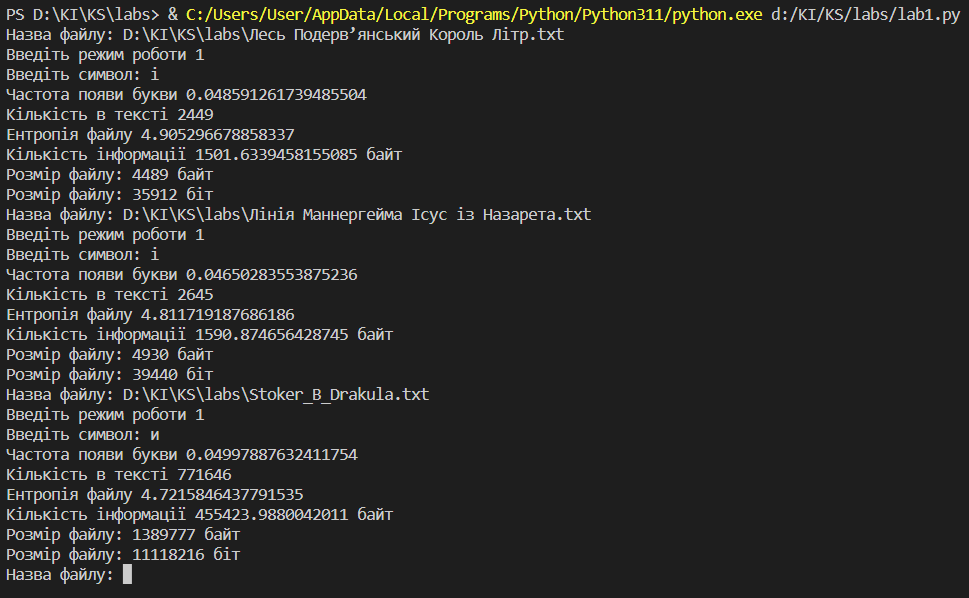


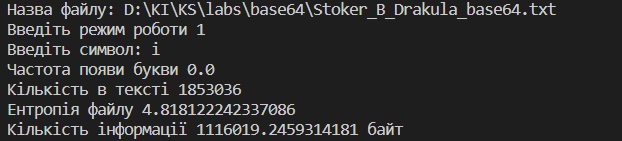


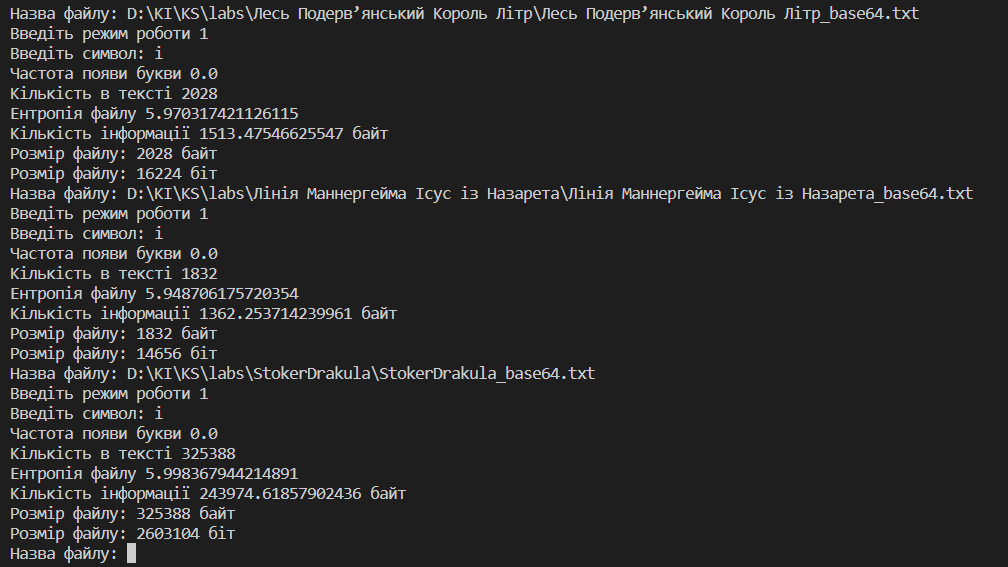












1. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу

c.      Зробіть висновки з отриманого результату

Вихідні коди розроблених програм завантажте в свій репозиторій на GitLab

В Moodle завантажте звіт, що містить:

·       Результати проведеного аналізу кількості інформації обраних текстів (самі тексти у вигляді посилань або в додатках)

·       Посилання на програму в GitLab [gl.vlabs.knu.ua](https://gl.vlabs.knu.ua/)

·       Приклад роботи створеної програми для підрахунку кількості інформації

·       Приклад роботи створеної програми для кодування в Base64

**Контрольні запитання**

1. Що таке кількість інформації?

2. Чим визначається межа стиснення без втрат?

3. Для чого використовується кодування Base64?

4. Чому розміри файлів, закодовані в Base64 більші в порівнянні з розмірами звичайних текстових файлів?

5. Якщо стиснути за допомогою zip, bz2, … медіафайл (фото/відео/аудіо) ти сильно буде відрізнятися розмір вихідного і стиснутого файлів, чому?

6. Як буде впливати повторне стиснення на розмір вже попередньо стиснутих файлів, поясніть чому?

7. В чому полягає суть алгоритму Base64?

8. Чому іноді в кінці тексту, закодованому за допомогою Base64 в з’являються символи  ‘=’?

9. Для чого використовують стиснення з втратами і без втрат?

10.  Який алгоритм стиснення найчастіше використовується для архівації zip?

**Критерії оцінювання**

1. Студент може пояснити всі дії, виконні в лабораторній роботі – 3 бали

2. Студент може пояснити алгоритм роботи алгоритму кодування Base64 – 2 бали

3. Студент може порівняти кількість інформації у вихідних файлах, закодованих, заархівованих і пояснити розбіжності – 3 бали

4. Студент може продемонструвати роботу програм в реальному часі – 2 бали