Лабораторна робота №1

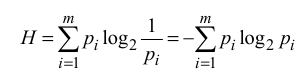
Дослідження кількості інформації при різних варіантах кодування

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

# Теоретичні відомості

***Відносна частота появи символу*** - імовірність появи певного символу в певному місці тексту - відношення числа появи символу в тексті до загальної кількості символів.

***Середня ентропія нерівноймовірного алфавіту***:



де *m* - кількість символів алфавіту, *p* - імовірність появи символу

Ентропія вимірюється в **БІТАХ** (як представлення кількості можливих варіантів).

***Кількість інформації в тексті*** - середня ентропія вихідного алфавіту помножена на кількість символів тексту. (***HINT***: результат обрахунку для порівняння значення з розміром файлів треба перевести з бітів в байти)

# 1. 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)
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:
   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
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 для стиснення).
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)[[1]](#footnote-0)

# 2. 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)
   1. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)
3. Закодуйте в Base64 обрані вами текстові файли
   1. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
   2. Порівняйте отримане значення з кількістю інформації вихідного файлу
   3. Зробіть висновки з отриманого результату
4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
   1. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
   2. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу[[2]](#footnote-1)
   3. Зробіть висновки з отриманого результату

## Вихідні коди розроблених програм завантажте в свій репозиторій на GitHub.

## В Moodle завантажте звіт, що містить:

* результати проведеного аналізу кількості інформації обраних текстів (самі тексти в вигляді посилань або в додатках)
* посилання на програму в GitHub
* приклад роботи створеної програми для підрахунку кількості інформації
* приклад роботи створеної програми для кодування в Base64

1. Для кращого сприйняття інформації **обов’язково** подайте отримані значення у вигляді таблиці, що містить всі варіанти значення обрахованої кількості інформації та **відповідні діаграми** на основі табличних даних [↑](#footnote-ref-0)
2. Для кращого сприйняття інформації **обов’язково** подайте отримані значення у вигляді таблиці, що містить всі варіанти значення обрахованої кількості інформації та **відповідні діаграми** на основі табличних даних [↑](#footnote-ref-1)