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:

$$H = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{m} p_i \log_2 p_i$$

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:
    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 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]

[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

2. Дослідження способів кодування інформації на прикладі Base64

1. Ознайомтесь зі стандартом 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 (наприклад, `openssl enc -base64`)
3. Закодуйте в Base64 обрані вами текстові файли
    a. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
    b. Порівняйте отримане значення з кількістю інформації вихідного файлу
    c. Зробіть висновки з отриманого результату
4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
    a. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
    b. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу[2]
    c. 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

[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