Звіт з лабораторної роботи №1

З курсу «Комп’ютерні системи»

Студента 3 курсу КІ-СА

Папідохи Владислава

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)

Було обрано 3 файли:

* + 1. Пісня Павла Зіброва «Жінки й вино»
    2. Вірш Ліни Костенко
    3. Біблійна притча

1. 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
2. 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).
3. 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)

Кількість інформації - мінімальна кількість біт, необхідна для кодування всіх можливих значень повідомлення. Вона буде меншою за розмір наший файлів.

З розповсюджених визначень, можна навести таке:

Кількість інформаціі  можна розгядати як міру зменшення невизначенності знання

При отриманні інформаційних поідомень.

|  |  |  |  |
| --- | --- | --- | --- |
| Файл | text1.txt, bytes | text2.txt, bytes | text3.txt, bytes |
| rar | 726 | 377 | 1089 |
| tar | 3584 | 2048 | 4096 |
| bz2 | 572 | 268 | 852 |
| zip | 778 | 436 | 1131 |
| xz | 704 | 352 | 1048 |
| ORIGINAL | 1563 | 462 | 2448 |
| Кількість інформації | 413 | 124 | 671 |

У результаті ідеального стиснення розмір файлу повинен бути рівним кількості інформації. Реально ж отримуємо, що розміри архівованих файлів у більшості своїй дещо більші за кількість інформації. Проте, у випадку використання більшого тексту, ми спостерігаємо більший рівень стиснення тексту. Так відбувається тому, що алгоритми архіваторів побудовані таким чином аби використати повторювані частини тексту. Виходячи з цього, формула розрахунку кількості інформації що ми рахували у роботі не є досконалою. Вона не враховує частоту появи символів та передбачення наступного шматочку текста.

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)

перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)



Figure Приклад закодованого тексту

1. 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

Кількість інформації у закодованих файлах стало більше. Це пов'язано з особливостями base64 кодування, яке збільшує кліькість символів у тексті на третину.

1. 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
     3. Зробіть висновки з отриманого результату

Найкраще стиснені файли були за допомоги інструмента для компресії bz2

У кодуванні Base64 розмір текстового файлу зростає у приблизно 4/3 рази.

**Посилання на код програм:**

<https://github.com/vpapidokha/Computer_systems/tree/master/1_Lab>

Використані текстові файли знаходяться у репозиторії