# МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА ФАКУЛЬТЕТ РАДІОФІЗИКИ, ЕЛЕКТРОНІКИ ТА КОМП’ЮТЕРНИХ СИСТЕМ

Кафедра комп’ютерної інженерії

# Звіт

**з лабораторної роботи №1**

«**Дослідження кількості інформації при різних варіантах кодування**»

Роботу виконав : студент 3-го курсу

напряму підготовки: «Комп’ютерна інженерія» спеціалізація: «Системний адміністратор»

Соснов Владислав

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 (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI)

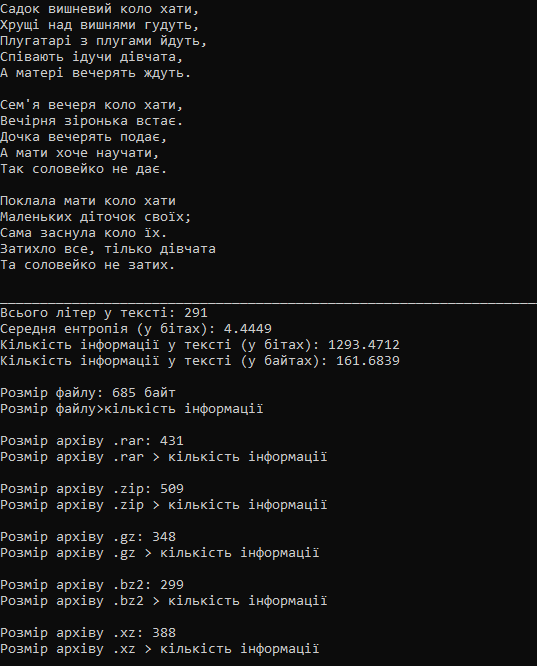
* [Terrasoft](https://uk.wikipedia.org/wiki/Terrasoft)
* [DHCP](http://a-yak.com/shho-take-dhcp/)
* [«Садок вишневий коло хати»](http://poetyka.uazone.net/kobzar/sadok_vyshnevyi.html%20-%20%D0%B2%D1%96%D1%80%D1%88%20%D0%A2.%20%D0%93.%20%D0%A8%D0%B5%D0%B2%D1%87%D0%B5%D0%BD%D0%BA%D0%B0)

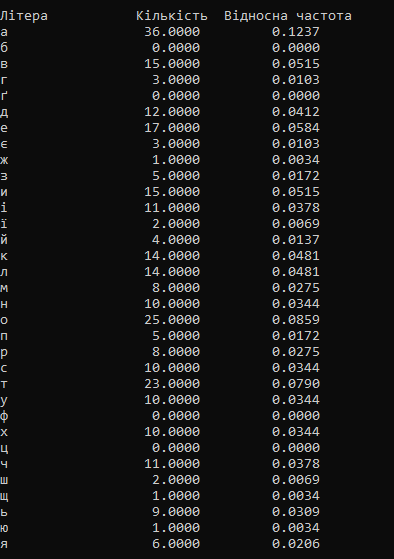
1. Переконайтесь, що тексти, які ви використовуєте є унікальними і не повторюються у ваших колег! Використовуйте наявні електронні засоби зв’язку та документообігу, щоб уникнути дублювання! Вдруге аналіз того самого тексту не зараховується!

[Перевірено](https://docs.google.com/spreadsheets/d/1lv_-aLfkekkrWQPtk8MoTJPQMLCKwFAy083K3bsuEKc/edit#gid%3D0)!

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 Приклад роботи створеної [програми](https://github.com/dekom8/Computer-Systems/blob/master/Entropy.cs%20(Sosnov_Lab1_part1)) для обрахунку кількості інформації :

«Садок вишневий коло хати»





1. 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).

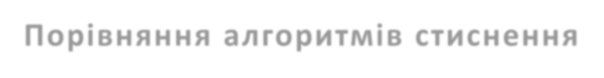
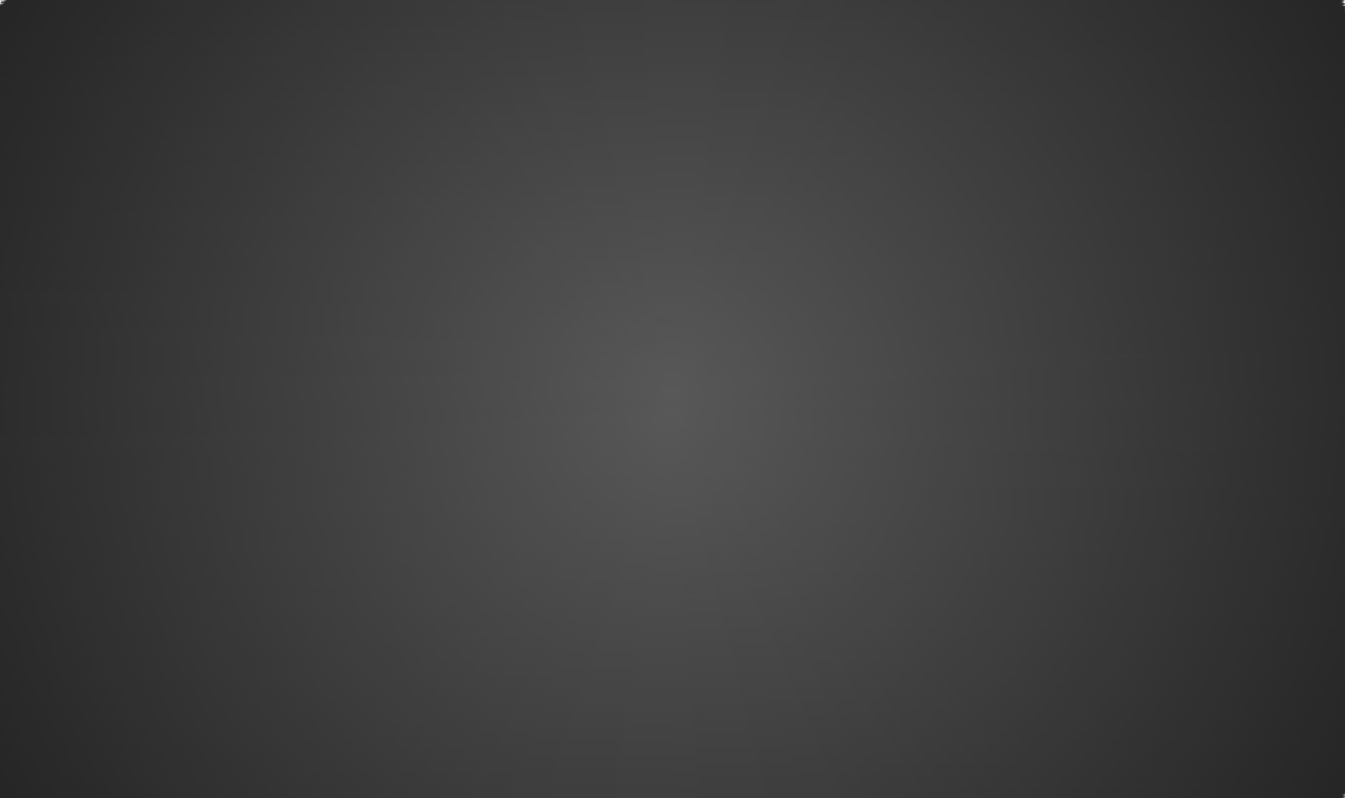
За допомогою програми WinRAR та [цього пАтужного сайту](https://online-converting.com/) файли було стиснено у zip,

rar, gzip, bzip2 та xz.

1. 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)

|  |  |  |  |
| --- | --- | --- | --- |
| Файл | DHCP.txt, Байт | Садок вишневий .txt, Байт | terrasoft.txt, Байт |
| rar | 923 | 431 | 762 |
| zip | 980 | 509 | 811 |
| gz | 844 | 348 | 784 |
| xz | 876 | 388 | 748 |
| bz2 | 721 | 299 | 720 |
| Оригінал | 1763 | 685 | 1789 |
| Кількість інформації | 425 | 162 | 405 |

*У всіх випадках найкраще стискав bzip2. Кількість інформації для всіх стиснених файлів менша за їхні фактичні розміри.*



rar zip gz xz bz2 Оригінал Кількість

terrasoft.txt, Байт

Садок вишневий .txt, Байт

DHCP.txt, Байт

0

162

299

405

388

348

200

425 431

509

400

784 748 720

685

762

721

600

811

876

844

923

800

980

1400

1200

1000

1763

1600

1789

2000

1800

**Порівняння алгоритмів стиснення**

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.

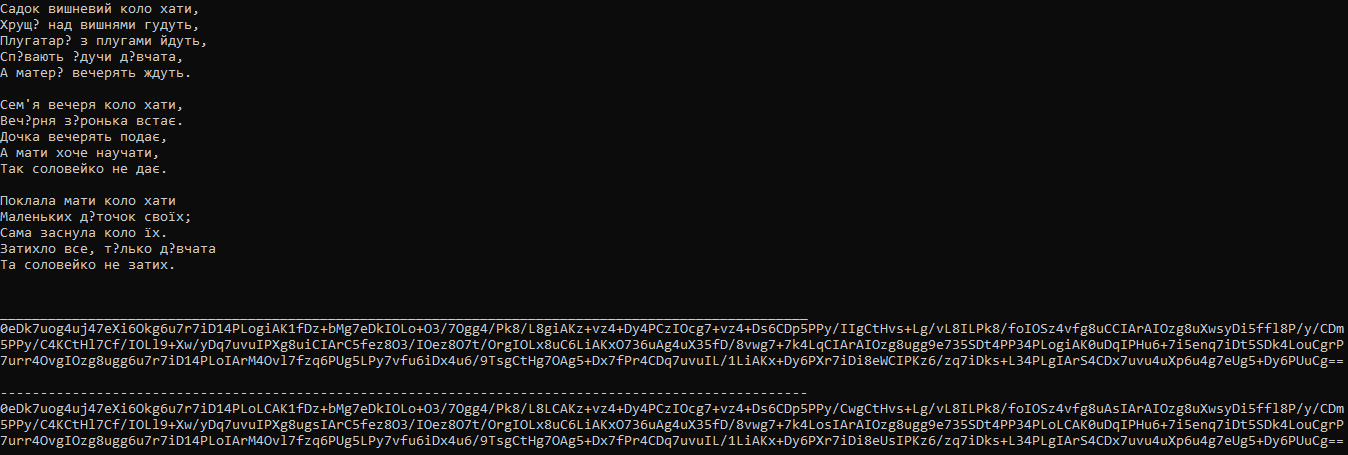
Це відбувається тому, що алгоритми архіваторів побудовані таким чином аби використати повторювані частини тексту. Виходячи з цього, формула розрахунку кількості інформації, використана для програми, не є досконалою, бо вона не враховує передбачення наступного уривку текста.

## *Дослідження способів кодування інформації на прикладі Base64*

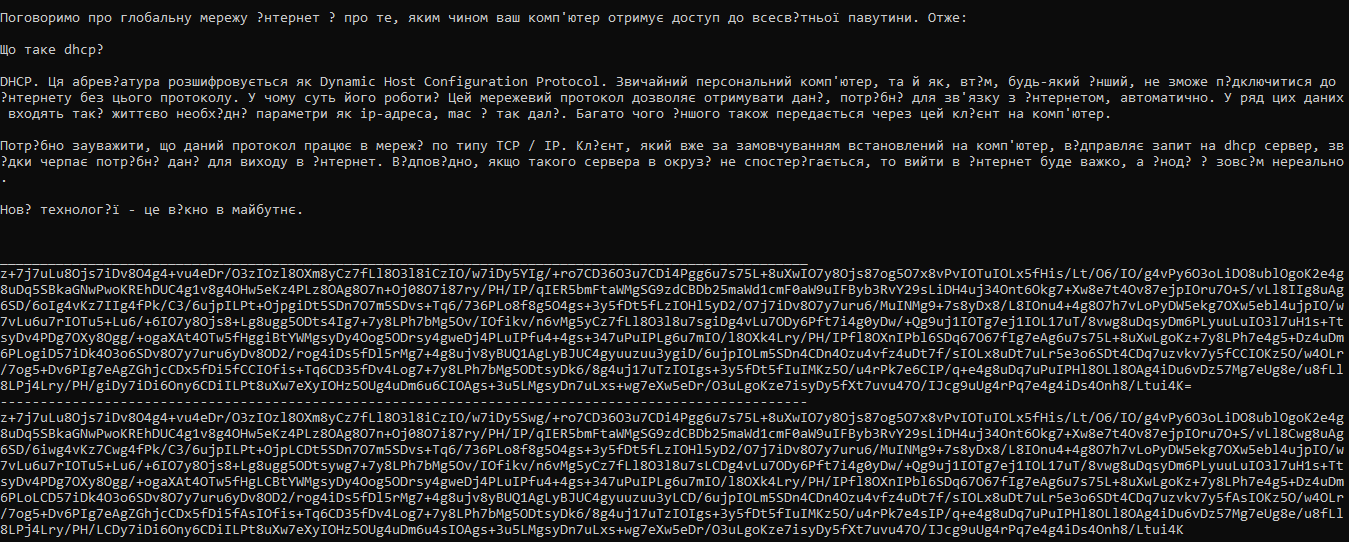
1. Ознайомтесь зі стандартом [RFC4648](https://tools.ietf.org/html/rfc4648)
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)

Верхнє кодування - моя програма. Знизу – використовуючи вже існуючу функцію у Visual Studio.

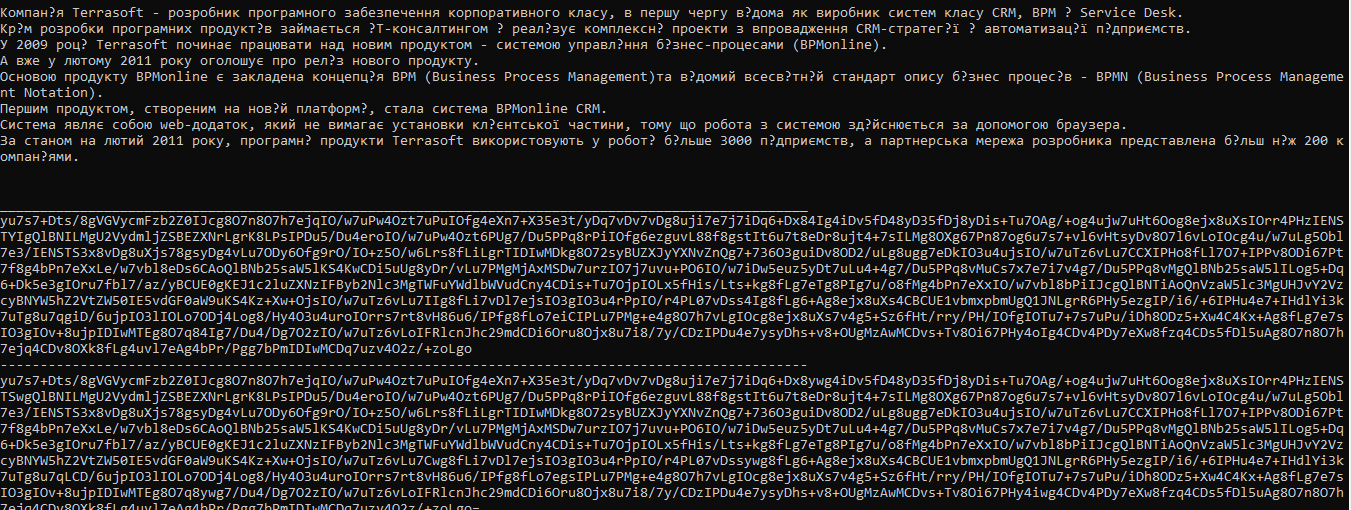
«Садок вишневий коло хати»



DHCP

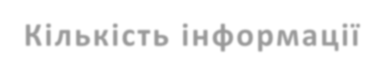
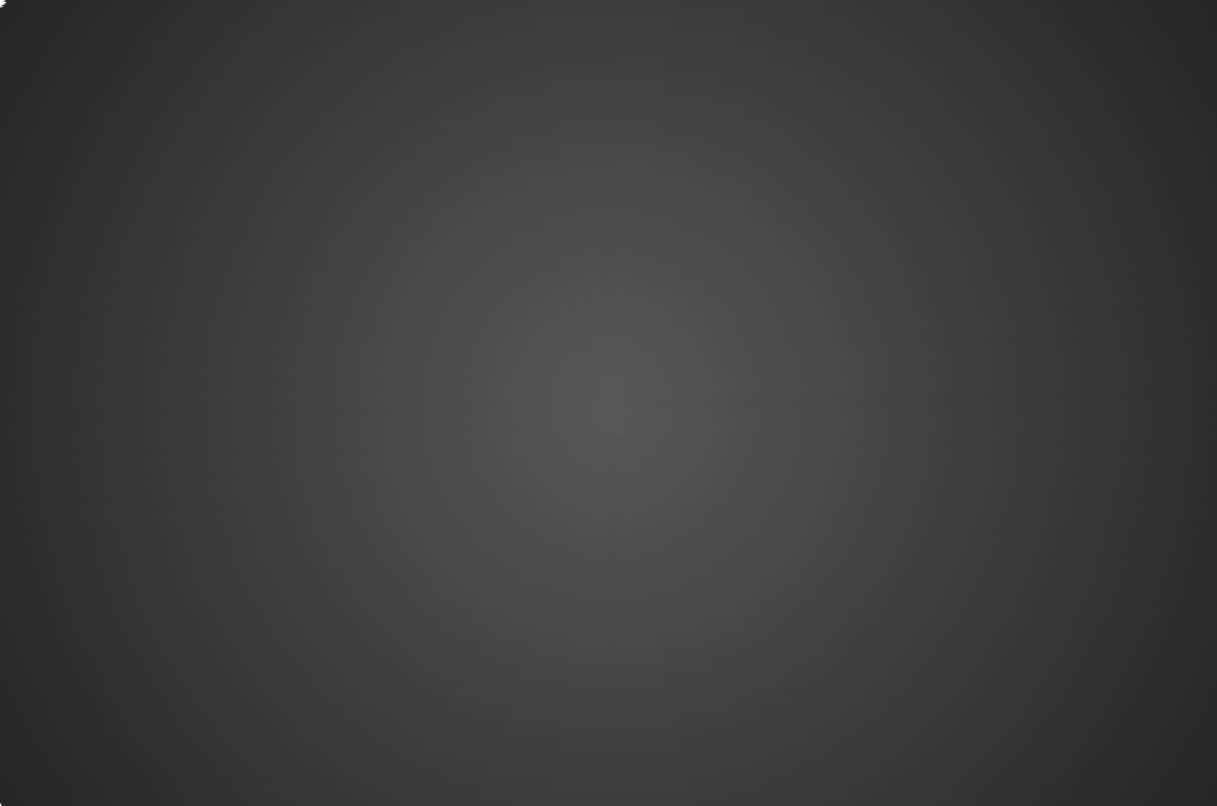


Terrasoft



Приклад роботи створеної [програми](https://github.com/dekom8/Computer-Systems/blob/master/Encoder_base64.cs%20(Sosnov_Lab1_part2)) для кодування в Base64.

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



terrasoft.txt, Байт

Садок вишневий .txt, Байт

DHCP.txt, Байт

0

100

200

300

400

500

600

700

800

**Кількість інформації**

Base64 Оригінал

3

2

7

0

65

5

2

6

4

0

4

8

26

1

6

1

|  |  |  |
| --- | --- | --- |
| **Файл** | **Кількість інф. у вихідному файлі** | **Кількість інф. у base64** |
| Садок вишневий.txt | 161.6839 | 267.8070 |
| DHCP.txt | 425.1303 | 722.4728 |
| terrasoft.txt | 405.9767 | 650.33 |

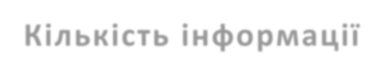
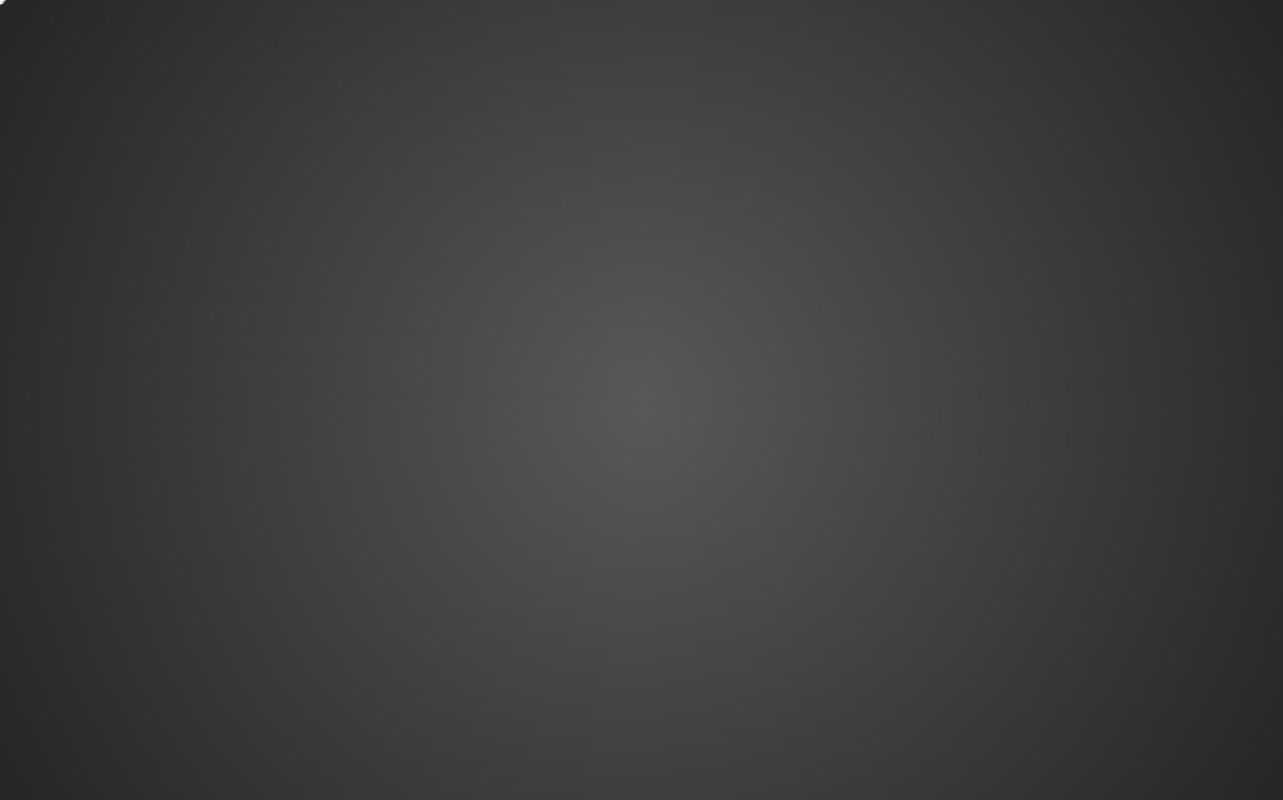
|  |  |  |  |
| --- | --- | --- | --- |
|  | DHCP.txt, Байт | Садок вишневий .txt, Байт | terrasoft.txt, Байт |
| Base64 | 723 | 268 | 650 |
| Оригінал | 425 | 161 | 406 |

* 1. Зробіть висновки з отриманого результату

Кількість інформації у Base64 закодованому файлі у середньому у 1.65 рази більше ніж кількість інформації у вихідному файлі.

1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
2. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
3. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-

закодованого файлу



**Кількість інформації**

551

terrasoft.txt, Байт

406

617

Садок вишневий .txt, Байт

425

249

DHCP.txt, Байт

162

0

100

200

300

400

500

600

700

Кількість інф. у base64 стисненого файлу bz2

Кількість інф. у вихідному файлі

|  |  |  |
| --- | --- | --- |
| **Файл** | **Кількість інф. у вихідному файлі** | **Кількість інф. у base64**  **стисненого файлу bz2** |
| Садок вишневий.txt | 161.6839 | 248.2375 |
| DHCP.txt | 425.1303 | 617.0215 |
| terrasoft.txt | 405.9767 | 551,2435 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | DHCP.txt, Байт | Садок вишневий .txt, Байт | terrasoft.txt, Байт |
| Кількість інф. у base64 стисненого файлу bz2 | 249 | 617 | 551 |
| Кількість інф. у вихідному файлі | 162 | 425 | 406 |

1. Зробіть висновки з отриманого результату

Кількість інформації у Base64 стисненого файлу у bz2 у середньому у 1.33 рази більше, ніж кількість інформації у вихідному файлі

**Висновок**

Протягом виконання даної лаб. роботи я дізнався, що таке ентропія інформації та про пов’язані з нею поняття. Теоретичних знань вже майже не залишилися, практичні знання виконав з болем. Також порівняв алгоритми стиснення – з усіх bz2 виявився найкращим. Також працював з алгоритмом кодування Base64, читав про його переваги та недоліки.

[All information related to this lab is here. (Push me, satisfaction)](https://github.com/dekom8/Computer-Systems)