

**Тема:** 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.

### 1. Дослідження кількості інформації в тексті

- ☐ Оберіть 3 текстових файли різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про ріпку” Леся Подерв'янського та специфікацію інтерфейсу PCI).
- ☐ вірш Тараса Шевченка “Мені тринадцятий минало”
- ☐ “Казка про ріпку”
- ☐ специфікацію інтерфейсу USB
- ☐ 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
 Програма написана мовою програмування C# її результати

Мені тринадцятий минало.txt	Казка про ріпку.txt	USB.txt
Frequency: 0.07788 Frequency: 0.07788 Letter: Frequency: 0.11604 Letter: ! Frequency: 0.00701 Letter: , Frequency: 0.02336 Letter: . Frequency: 0.02804 Letter: / Frequency: 0.00156 Letter: 3 Frequency: 0.00078 Letter: 7 Frequency: 0.00078 Letter: : Frequency: 0.00078 Letter: ? Frequency: 0.00078 Letter: [ Frequency: 0.00078 Letter: ] Frequency: 0.00078 Letter: I Frequency: 0.00545 Letter: A Frequency: 0.00234 Letter: Б Frequency: 0.00467 Letter: Г Frequency: 0.00156 Letter: 3 Frequency: 0.00078 Letter: Л Frequency: 0.00078 Letter: М Frequency: 0.00467	Frequency: 0.00762 Frequency: 0.00762 Letter: Frequency: 0.1664 Letter: ! Frequency: 0.00481 Letter: , Frequency: 0.03569 Letter: - Frequency: 0.0008 Letter: . Frequency: 0.00642 Letter: : Frequency: 0.00241 Letter: ; Frequency: 0.0008 Letter: « Frequency: 0.00241 Letter: » Frequency: 0.00241 Letter: А Frequency: 0.0008 Letter: Б Frequency: 0.0004 Letter: В Frequency: 0.0016 Letter: Д Frequency: 0.0004 Letter: 3 Frequency: 0.0004 Letter: К Frequency: 0.002 Letter: М Frequency: 0.00241 Letter: П Frequency: 0.00281 Letter: Р Frequency: 0.0008	Frequency: 0.00283 Frequency: 0.00283 Letter: Frequency: 0.1304 Letter: ' Frequency: 0.00354 Letter: ( Frequency: 0.00283 Letter: ) Frequency: 0.00283 Letter: + Frequency: 0.00071 Letter: , Frequency: 0.00992 Letter: - Frequency: 0.00709 Letter: . Frequency: 0.00921 Letter: 0 Frequency: 0.00496 Letter: 1 Frequency: 0.00071 Letter: 2 Frequency: 0.00071 Letter: 3 Frequency: 0.00071 Letter: 5 Frequency: 0.00213 Letter: 7 Frequency: 0.00071 Letter: 9 Frequency: 0.00071 Letter: : Frequency: 0.00071 Letter: В Frequency: 0.00709 Letter: D Frequency: 0.00071

<p>Letter: H Frequency: 0.00779  Letter: O Frequency: 0.00078  Letter: П Frequency: 0.00312  Letter: C Frequency: 0.00078  Letter: T Frequency: 0.00389  Letter: Y Frequency: 0.00234  Letter: Ч Frequency: 0.00467  Letter: Я Frequency: 0.00156  Letter: а Frequency: 0.05374  Letter: б Frequency: 0.01246  Letter: в Frequency: 0.02336  Letter: г Frequency: 0.01869  Letter: д Frequency: 0.01791  Letter: е Frequency: 0.04361  Letter: ж Frequency: 0.00467  Letter: з Frequency: 0.00935  Letter: и Frequency: 0.03349  Letter: й Frequency: 0.00623  Letter: к Frequency: 0.0148  Letter: л Frequency: 0.04439  Letter: м Frequency: 0.01636  Letter: н Frequency: 0.04206  Letter: о Frequency: 0.07866  Letter: п Frequency: 0.01558  Letter: р Frequency: 0.02025  Letter: с Frequency: 0.02804  Letter: т Frequency: 0.02181  Letter: у Frequency: 0.01558  Letter: х Frequency: 0.00234  Letter: ц Frequency: 0.00467  Letter: ч Frequency: 0.01168  Letter: ш Frequency: 0.00078  Letter: щ Frequency: 0.00078  Letter: ь Frequency: 0.00935  Letter: ю Frequency: 0.00935  Letter: я Frequency: 0.02181  Letter: е Frequency: 0.00312  Letter: і Frequency: 0.02882  Letter: ї Frequency: 0.00234  Letter: — Frequency: 0.00156  Letter: ' Frequency: 0.00078</p>	<p>Letter: C Frequency: 0.0016  Letter: T Frequency: 0.0008  Letter: Y Frequency: 0.00281  Letter: X Frequency: 0.00361  Letter: Ч Frequency: 0.0004  Letter: Щ Frequency: 0.0004  Letter: а Frequency: 0.0846  Letter: б Frequency: 0.02446  Letter: в Frequency: 0.03047  Letter: г Frequency: 0.01524  Letter: д Frequency: 0.03769  Letter: е Frequency: 0.02045  Letter: ж Frequency: 0.00521  Letter: з Frequency: 0.01684  Letter: и Frequency: 0.04852  Letter: й Frequency: 0.0016  Letter: к Frequency: 0.05253  Letter: л Frequency: 0.01564  Letter: м Frequency: 0.01925  Letter: н Frequency: 0.03168  Letter: о Frequency: 0.05533  Letter: п Frequency: 0.03047  Letter: р Frequency: 0.03729  Letter: с Frequency: 0.02285  Letter: т Frequency: 0.02085  Letter: у Frequency: 0.04932  Letter: х Frequency: 0.002  Letter: ц Frequency: 0.00441  Letter: ч Frequency: 0.02285  Letter: ш Frequency: 0.01043  Letter: ь Frequency: 0.01524  Letter: ю Frequency: 0.00601  Letter: я Frequency: 0.01604  Letter: е Frequency: 0.0012  Letter: і Frequency: 0.03649  Letter: ї Frequency: 0.0004  Letter: — Frequency: 0.00601</p>	<p>Letter: G Frequency: 0.00071  Letter: N Frequency: 0.00071  Letter: S Frequency: 0.00709  Letter: U Frequency: 0.00709  Letter: « Frequency: 0.00142  Letter: » Frequency: 0.00142  Letter: A Frequency: 0.00213  Letter: B Frequency: 0.00283  Letter: Д Frequency: 0.00071  Letter: З Frequency: 0.00142  Letter: К Frequency: 0.00071  Letter: С Frequency: 0.00071  Letter: Ц Frequency: 0.00142  Letter: Я Frequency: 0.00071  Letter: а Frequency: 0.05386  Letter: б Frequency: 0.00921  Letter: в Frequency: 0.03969  Letter: г Frequency: 0.0085  Letter: д Frequency: 0.02622  Letter: е Frequency: 0.04536  Letter: ж Frequency: 0.01063  Letter: з Frequency: 0.01134  Letter: и Frequency: 0.0404  Letter: й Frequency: 0.00496  Letter: к Frequency: 0.03189  Letter: л Frequency: 0.0241  Letter: м Frequency: 0.01772  Letter: н Frequency: 0.0652  Letter: о Frequency: 0.07583  Letter: п Frequency: 0.02481  Letter: р Frequency: 0.04678  Letter: с Frequency: 0.02835  Letter: т Frequency: 0.04678  Letter: у Frequency: 0.02055  Letter: ф Frequency: 0.00354  Letter: х Frequency: 0.01134  Letter: ц Frequency: 0.0085  Letter: ч Frequency: 0.00567  Letter: ш Frequency: 0.00496  Letter: щ Frequency: 0.00425  Letter: ь Frequency: 0.0163  Letter: ю Frequency: 0.01347  Letter: я Frequency: 0.02126  Letter: е Frequency: 0.01063  Letter: і Frequency: 0.03827  Letter: ї Frequency: 0.0078  Letter: — Frequency: 0.00142</p>
4	4	4
782	1469	875

□ Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір)

В таблиці розмір файлів

	Мені тринадцятий минало.txt	Казка про пірку.txt	USB.txt
original	2,142 bytes	4,426 bytes	2,535 bytes
rar	1,128 bytes	1,182 bytes	1,025 bytes
zip	1,184 bytes	1,311 bytes	1,178 bytes
bz2	983 bytes	1,100 bytes	847 bytes
7z	1,129 bytes	1,188 bytes	1,026 bytes
gz	1,112 bytes	1,156 bytes	992 bytes
Обчислена кількість інформації	782	1469	875



Згідно з результатами першої частини лабораторної роботи можна побачити, що у другому файлі об'єм оригінального тексту найбільш сильно відрізняється від кількості інформації в ньому. Крім того для файлів 2 та 3 кількість інформації дещо більше за об'єм архіву. Найкращі результати при архівуванні показав алгоритм bz2 для всіх трьох файлів.

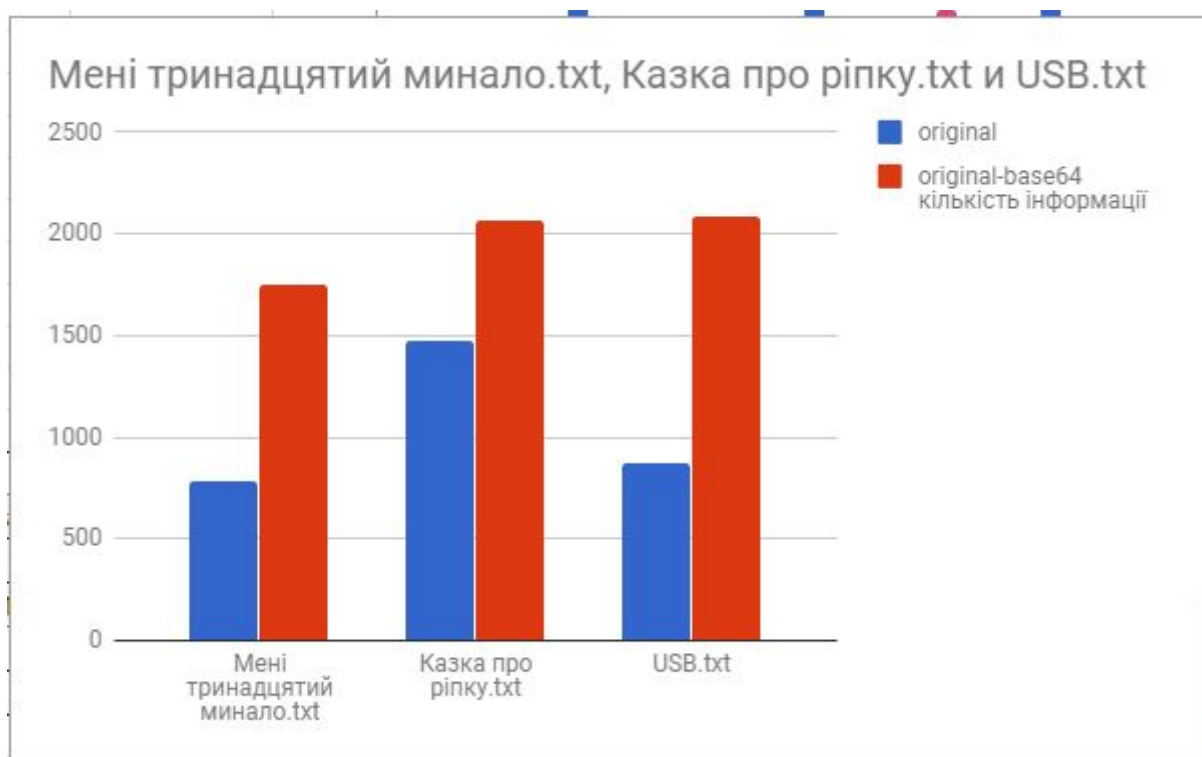
## 2. Дослідження способів кодування інформації на прикладі Base64

Програма написана мовою C#, коректність роботи перевірена за допомогою ресурсу

<http://www.onedollardata.com/encoder.php>

1. Закодуйте в Base64 обрані вами текстові файли
  - а. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
  - б. Порівняйте отримане значення з кількістю інформації вихідного файлу
  - в. Зробіть висновки з отриманого результату
2. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
  - а. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
  - б. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу<sup>1</sup>
  - в. Зробіть висновки з отриманого результату

	Мені тринадцятий минало.txt	Казка про піпку.txt	USB.txt
original	782	1469	875
original-base 64 кількість інформації	1754	2063	2081
bz2	517	653	600
bz2-base64	646	706	664



<sup>1</sup> Для кращого сприйняття інформації подайте отримані значення у вигляді таблиці, що містить всі варіанти значення обрахованої кількості інформації та відповідні діаграми на основі табличних даних



Відповідно до отриманих даних можна зробити висновок, що при використанні алгоритму Base-64 до вихідного файлу кількість інформації в ньому збільшується. Це можна пояснити тим, що при кодуванні у base64 розмір файлу збільшується у відношенні 4:3.

Кількість інформації у архівованому найкращим (як виявилось експериментальним шляхом) архіватором bz2 і закодованого потім в base64 – мала різний вигляд для файлів.

Всі файли, їх архіви, закодоване у Base64 і вихідні коди - [https://github.com/tet-zu/Computer\\_Systems](https://github.com/tet-zu/Computer_Systems)