Київський національний університет імені Тараса Шевченка

Факультет радіофізики, електроніки та комп’ютерних систем

Лабораторна робота № 1

на тему:«Дослідження кількості інформації при різних варіантах кодування»

Роботу виконала

студентка 3 курсу

КІ-СА

Крішен Єлизавета

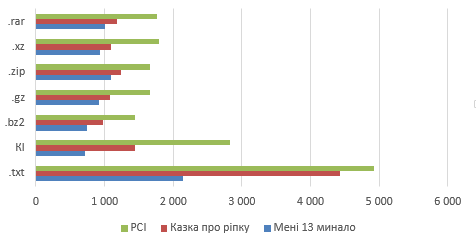
Київ 2018

**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**

|  |  |  |
| --- | --- | --- |
| Мені 13 минало | Казка про ріпку | PCI |
| Середня ентропія: 4,858306479813 bits  Кількість символів: 1184  Кількість інформації: 719,029359012324 bytes  Частоту кожного символу:  = 0,08446  = 0,12584  ! = 0,00760  , = 0,02534  . = 0,03041  / = 0,00169  3 = 0,00084  7 = 0,00084  : = 0,00084  ? = 0,00084  [ = 0,00084  ] = 0,00084  ? = 0,00591  А = 0,00253  Б = 0,00507  Г = 0,00169  З = 0,00084  Л = 0,00084  М = 0,00507  Н = 0,00845  О = 0,00084  П = 0,00338  С = 0,00084  Т = 0,00422  У = 0,00253  Ч = 0,00507  Я = 0,00169  а = 0,05828  б = 0,01351  в = 0,02534  г = 0,02027  д = 0,01943  е = 0,04730  ж = 0,00507  з = 0,01014  и = 0,03632  й = 0,00676  к = 0,01605  л = 0,04814  м = 0,01774  н = 0,04561  о = 0,08530  п = 0,01689  р = 0,02196  с = 0,03041  т = 0,02365  у = 0,01689  х = 0,00253  ц = 0,00507  ч = 0,01267  ш = 0,00084  щ = 0,00084  ь = 0,01014  ю = 0,01014  я = 0,02365  є = 0,00338  ? = 0,03125  ї = 0,00253  - = 0,00169  ' = 0,00084 | **Середня ентропія:**  4,68732772131689 bits  **Кількість символів:** 2476  **Кількість інформації:** 1450,72792974758 bytes  **Частоту кожного символу:**  = 0,00808  = 0,16761  ! = 0,00485  , = 0,03595  - = 0,00081  . = 0,00646  : = 0,00242  ; = 0,00081  < = 0,00242  > = 0,00242  А = 0,00081  Б = 0,00040  В = 0,00162  Д = 0,00040  З = 0,00040  К = 0,00202  М = 0,00242  П = 0,00283  Р = 0,00081  С = 0,00162  Т = 0,00081  У = 0,00283  Х = 0,00363  Ч = 0,00040  Щ = 0,00040  а = 0,08522  б = 0,02464  в = 0,03069  г = 0,01535  д = 0,03796  е = 0,02060  ж = 0,00525  з = 0,01696  и = 0,04887  й = 0,00162  к = 0,05291  л = 0,01575  м = 0,01939  н = 0,03191  о = 0,05574  п = 0,03069  р = 0,03756  с = 0,02302  т = 0,02100  у = 0,04968  х = 0,00202  ц = 0,00444  ч = 0,02302  ш = 0,01050  ь = 0,01535  ю = 0,00606  я = 0,01616  є = 0,00121  ? = 0,03675  ї = 0,00040  - = 0,00606 | **Середня ентропія алфаіту:** 5,15731900492174 bits  **Кількість символів**: 2824  **Кількість інформацї**  1820,53360873738 bytes  **Частоту кожного символу:**  = 0,00567  = 0,13173  # = 0,00142  % = 0,00035  ' = 0,00212  ( = 0,00390  ) = 0,00390  , = 0,01416  - = 0,00106  . = 0,00496  / = 0,00071  0 = 0,00106  1 = 0,00460  2 = 0,00177  3 = 0,00071  4 = 0,00106  6 = 0,00142  8 = 0,00142  : = 0,00035  ; = 0,00106  A = 0,00035  B = 0,00106  C = 0,00567  D = 0,00106  E = 0,00531  G = 0,00035  H = 0,00035  I = 0,00673  L = 0,00035  N = 0,00071  O = 0,00071  P = 0,00531  Q = 0,00142  R = 0,00142  S = 0,00142  T = 0,00071  V = 0,00035  a = 0,00248  b = 0,00071  c = 0,00035  e = 0,00673  g = 0,00071  h = 0,00035  i = 0,00142  k = 0,00035  l = 0,00248  m = 0,00035  n = 0,00283  p = 0,00425  r = 0,00496  s = 0,00885  t = 0,00106  x = 0,01062  < = 0,00071  > = 0,00071  В = 0,00071  Д = 0,00071  З = 0,00035  К = 0,00035  Н = 0,00071  П = 0,00071  С = 0,00035  У = 0,00035  а = 0,05559  б = 0,00779  в = 0,03506  г = 0,00779  д = 0,03045  е = 0,03010  ж = 0,00496  з = 0,01346  и = 0,04816  й = 0,00425  к = 0,03010  л = 0,02762  м = 0,02195  н = 0,05807  о = 0,07011  п = 0,03010  р = 0,04037  с = 0,02727  т = 0,03612  у = 0,01735  ф = 0,00460  х = 0,00885  ц = 0,00708  ч = 0,00956  ш = 0,00496  щ = 0,00177  ь = 0,01133  ю = 0,00921  я = 0,01841  є = 0,00850  ? = 0,04320  ї = 0,00496 |

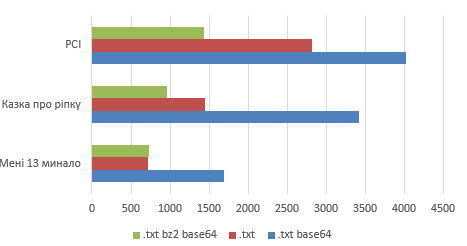
|  |  |  |  |
| --- | --- | --- | --- |
|  | Мені 13 минало | Казка про ріпку | PCI |
| .txt | 2,142 | 4,428 | 4,918 |
| К-сть інф. | 719 | 1450 | 2824 |
| .bz2 | 743 | 980 | 1,449 |
| .gz | 914 | 1,073 | 1,667 |
| .zip | 1,098 | 1,233 | 1,664 |
| .xz | 936 | 1,096 | 1,789 |
| .rar | 1,013 | 1,177 | 1,759 |



**До першого завдання можна зробити висновок, що**архівуючи текст ми можемо зменшити його об’єм, але він ніколи не буде менше за кількість інформації, оскільки це мінімальний розмір, який необхідний, щоб певна інформація не була пошкоджена і могла бути отримана і використана. Використовуючи різні архіватори, мі можемо знайти саме той, що найбільше прямує до величини кількості інформації, але через не ідеальність алгоритмів архіваторів, отримані експериментальні значення не дорівнюють кількості інформації.

**2. Дослідження способів кодування інформації на прикладі Base64**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Мені 13 минало | Казка про ріпку | PCI |
| .txt base64 | 1686 | 3421 | 4023 |
| .txt | 719 | 1450 | 2824 |
| .txt bz2 base64 | 733 | 964 | 1435 |



**Висновки:** кількість інформації в Base64 більше, ніж в оригінальному файлі. Це – логічно, оскільки при кодуванні у base64 розмір файлу збільшується у відношенні 4:3. Кількість інформаціі у архівованому найкращим (як виявилося експериментальним шляхом – найкраще стиснув текст пісні) архіватором gz і закодованого потім в base64 – мала різний вигляд для файлів. Це пояснюється тим, що архів – це бінарний файл, а не текстовий. <https://github.com/lizakrishen/computer_systems/tree/master/Lab1>