1. Дослідження кількості інформації в текст

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

|  |  |  |
| --- | --- | --- |
| Мені 13 минало | Казка про ріпку | PCI |
| Середня ентропія алфавіту = 4 bits  Кількість інформації 752 bytes  Частоту кожного символу:  = 0.07788  = 0.07788  = 0.11604  ! = 0.00701  , = 0.02336  . = 0.02804  / = 0.00156  3 = 0.00078  7 = 0.00078  : = 0.00078  ? = 0.00078  [ = 0.00078  ] = 0.00078  а = 0.05607  б = 0.01713  в = 0.02336  г = 0.02025  д = 0.01791  е = 0.04361  ж = 0.00467  з = 0.01012  и = 0.03349  й = 0.00623  к = 0.01480  л = 0.04517  м = 0.02103  н = 0.04984  о = 0.07944  п = 0.01869  р = 0.02025  с = 0.02882  т = 0.02570  у = 0.01791  х = 0.00234  ц = 0.00467  ч = 0.01636  ш = 0.00078  щ = 0.00078  ь = 0.00935  ю = 0.00935  я = 0.02336  є = 0.00312  і = 0.03427  ї = 0.00234  — = 0.00156  ’ = 0.00078 | Середня ентропія алфавіту = 4 bits  Кількість інформації 1439 bytes  Частоту кожного символу:  = 0.00801  = 0.00801  = 0.16627  ! = 0.00481  , = 0.03566  - = 0.00080  . = 0.00641  : = 0.00240  ; = 0.00080  « = 0.00240  » = 0.00240  а = 0.08534  б = 0.02484  в = 0.03205  г = 0.01522  д = 0.03806  е = 0.02043  ж = 0.00521  з = 0.01723  и = 0.04848  й = 0.00160  к = 0.05449  л = 0.01563  м = 0.02163  н = 0.03165  о = 0.05529  п = 0.03325  р = 0.03806  с = 0.02444  т = 0.02163  у = 0.05208  х = 0.00561  ц = 0.00441  ч = 0.02324  ш = 0.01042  щ = 0.00040  ь = 0.01522  ю = 0.00601  я = 0.01603  є = 0.00120  і = 0.03646  ї = 0.00040  — = 0.00601 | Середня ентропія алфавіту = 5 bits  Кількість інформації 1809 bytes  Частоту кожного символу:  = 0.00563  = 0.00563  = 0.13099  # = 0.00141  % = 0.00035  ' = 0.00211  ( = 0.00387  ) = 0.00387  , = 0.01408  - = 0.00106  . = 0.00493  / = 0.00070  0 = 0.00106  1 = 0.00458  2 = 0.00176  3 = 0.00070  4 = 0.00106  6 = 0.00141  8 = 0.00141  : = 0.00035  ; = 0.00106  a = 0.00282  b = 0.00176  c = 0.00599  d = 0.00106  e = 0.01197  g = 0.00106  h = 0.00070  i = 0.00810  k = 0.00035  l = 0.00282  m = 0.00035  n = 0.00352  o = 0.00070  p = 0.00951  q = 0.00141  r = 0.00634  s = 0.01021  t = 0.00176  v = 0.00035  x = 0.01056  « = 0.00070  » = 0.00070  а = 0.05528  б = 0.00775  в = 0.03556  г = 0.00775  д = 0.03099  е = 0.02993  ж = 0.00493  з = 0.01373  и = 0.04789  й = 0.00423  к = 0.03028  л = 0.02746  м = 0.02183  н = 0.05845  о = 0.06972  п = 0.03063  р = 0.04014  с = 0.02746  т = 0.03592  у = 0.01761  ф = 0.00458  х = 0.00880  ц = 0.00704  ч = 0.00951  ш = 0.00493  щ = 0.00176  ь = 0.01127  ю = 0.00915  я = 0.01831  є = 0.00845  і = 0.04296  ї = 0.00493 |

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)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Мені 13 минало | Казка про ріпку | PCI |
| .txt | 2,142 | 4,428 | 4,918 |
| .txt КІ | 752 | 1439 | 1809 |
| .bz2 | 743 | 980 | 1,449 |
| .gz | 914 | 1,073 | 1,667 |
| .zip | 1,098 | 1,233 | 1,664 |
| .xz | 936 | 1,096 | 1,789 |

# 2. Дослідження способів кодування інформації на прикладі 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
   3. Зробіть висновки з отриманого результату

|  |  |  |  |
| --- | --- | --- | --- |
|  | Мені 13 минало | Казка про ріпку | PCI |
| .txt base64 | 1,746 | 3,365 | 3,909 |
| .txt | 752 | 1439 | 1809 |

У закодованому варіанті base64 кількість інформації

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | Мені 13 минало | Казка про ріпку | PCI |
| .txt base64 | 1,746 | 3,365 | 3,909 |
| .txt | 752 | 1439 | 1809 |
| .bz2 base64 | 631 | 830 | 1,236 |

