Лабораторна робота №1

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.

Вибрані тексти:

* вірш OmegaFadizTertas ”Ten Thousand Years” (Ten Thousand Years.txt)
* Волков «Чарівник Смарагдового міста» (kazka.txt)
* специфікація протоколу PPP RFC1331 (rfc1331.txt)

Результати аналізу програми та розміри після стиснення

|  |  |  |  |
| --- | --- | --- | --- |
| Ім’я файлу | Ten Thousand Years.txt | kazka.txt | rfc1331.txt |
| Середня ентропія | 4.281 | 4.753 | 4.684 |
| Кількість інформації, bytes | 459.15 | 101977.02 | 52692.64 |
| Розмір файлу, bytes | 1066 | 376016 | 129892 |
| Відношення інформація/розмір, % | 43.07 | 27.12 | 40.57 |
| Після стиснення | | | |
| zip | 755 | 95340 | 29948 |
| %\* | 60.81 | 106.96 | 175.95 |
| xz | 696 | 82168 | 27424 |
| % | 65.97 | 124.11 | 192.14 |
| gz | 618 | 95216 | 29822 |
| % | 74.30 | 107.10 | 176.69 |
| bzip2 | 593 | 67748 | 26529 |
| % | 77.43 | 150.52 | 198.62 |
| rar | 702 | 88389 | 29441 |
| % | 65.40 | 115.37 | 178.97 |

\*відношення інформація/розмір файлу

Кількість інформації та розміри файлів

Можна зробити наступні висновки:

* Низьке відношення інформації до розміру файлу kazka.txt пояснюється тим, що в ньому кириличний текст в кодуванні UTF8. У інших файлів також таке кодування, але для символів, що є стандартними для ASCII, кодування і відповідно кількість байт співпадає і в UTF8 і в ASCII
* Це відношення значно покращується при стисненні інформації. І чим більше файл, тим ефективніше. Наприклад відношення для кириличного тексту зросло приблизно в 7 раз, а для файлу rfc1331.txt досягло 198% з 40%.
* Найефективнішим алгоритмом стиснення виявився bzip2. Але інші алгоритми приблизно на його рівні.

2. Дослідження способів кодування інформації на прикладі Base64

Зміни після кодування файлів в Base64

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ім’я файлу | Середня ентропія | Кількість інформації, bytes | Розмір файлу, bytes | Відношення інформація/розмір, % |
| Ten Thousand Years.txt | 4.688 | 459.15 | 1066 | 43.07 |
| encoded\_Ten Thousand Years.txt | 4.795 | 853.55 | 1424 | 59.94 |
| encoded\_Ten Thousand Years.txt.bz2 | 5.158 | 510.63 | 792 | 64.47 |
| kazka.txt | 4.753 | 101977 | 376016 | 27.12 |
| encoded\_kazka.txt | 4.498 | 281862 | 501356 | 56.22 |
| encoded\_kazka.txt.bz2 | 5.174 | 58417.94 | 90332 | 64.67 |
| rfc1331.txt | 4.684 | 52692 | 129892 | 40.57 |
| encoded\_rfc1331.txt | 4.688 | 101497 | 173192 | 58.60 |
| encoded\_rfc1331.txt.bz2 | 5.163 | 22830.25 | 35372 | 64.54 |

При кодуванні текстових файлів в Base64 розмір файлів та кількість обчисленої інформації зростає, причому остання особливо. Частково це пояснюється особливістю самого алгоритму кодування, і також тим що в алгоритмі підрахунку інформації не враховуються whitespace символи, але вони також кодуються і перетворюються на послідовності символів.

Стиснення значно зменшує обсяг інформації, при чому не залежно від того текст на кирилиці чи латиниці, і кодування стиснених файлів не настільки збільшує обсяг інформації інформацію, як для не стиснених файлів.

Додаток. Приклад виводу програми аналізатора.

------------------------------

rfc1331.txt

n 0.066

e 0.105

t 0.088

w 0.007

o 0.069

r 0.053

k 0.007

i 0.068

g 0.014

u 0.022

p 0.027

. 0.029

s 0.052

m 0.023

q 0.003

f 0.023

c 0.042

: 0.001

1 0.007

3 0.005

d 0.031

a 0.061

y 0.010

b 0.010

l 0.033

7 0.002

, 0.009

2 0.004

9 0.004

h 0.029

- 0.035

( 0.002

) 0.002

v 0.006

" 0.001

z 0.001

x 0.005

@ 0.000

[ 0.001

] 0.001

4 0.002

5 0.002

6 0.003

0 0.007

8 0.002

j 0.001

' 0.000

/ 0.001

+ 0.016

| 0.003

\* 0.001

; 0.000

> 0.000

^ 0.000

< 0.000

= 0.001

! 0.000

{ 0.000

} 0.000

# 0.000

& 0.000

\ 0.000

% 0.000

? 0.000

entropy info size relation

4.684 52692.64 129892 40.57