Київський національний університет імені Тараса Шевченка
факультет радіофізики, електроніки та комп’ютерних систем

Лабораторна робота 1

Роботу виконав
студент 3 курсу
Зажидько Дмитро

Київ 2019

Мета: Дослідити імовірнісні параметри української мови для оцінки кількості інформації текстів. Дослідити вплив різних методів кодування інформації на її кількість.

Хід виконання роботи:

1. Дослідження кількості інформації в тексті

1. 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

text1 (BMTH - Can u feel my heart)

```
Frequency of appearence for charachter '
': 0.030685920577617327
Frequency of appearence for charachter 'c': 0.01444043321299639
Frequency of appearence for charachter 'x': 0.002707581227436823
Frequency of appearence for charachter 'g': 0.019855595667870037
Frequency of appearence for charachter 'i': 0.021660649819494584
Frequency of appearence for charachter 'u': 0.02075812274368231
Frequency of appearence for charachter 'S': 0.0090025270758122744
Frequency of appearence for charachter 'd': 0.01444043321299639
Frequency of appearence for charachter 's': 0.026173285198555957
Frequency of appearence for charachter 'b': 0.008122743682310469
Frequency of appearence for charachter ',': 0.009025270758122744
Frequency of appearence for charachter 'k': 0.011732851985559567
Frequency of appearence for charachter 'n': 0.05595667870036101
Frequency of appearence for charachter 'y': 0.0351985559566787
Frequency of appearence for charachter 'T': 0.002707581227436823
Frequency of appearence for charachter 'h': 0.045126353790613372
Frequency of appearence for charachter 'r': 0.0388086642599278
Frequency of appearence for charachter ' ': 0.18050541516245489
Frequency of appearence for charachter ''': 0.009927797833935019
Frequency of appearence for charachter 'w': 0.013537906137184115
Frequency of appearence for charachter 'o': 0.06859205776173286
Frequency of appearence for charachter 'I': 0.01895306859205776
Frequency of appearence for charachter 'a': 0.05595667870036101
Frequency of appearence for charachter '?': 0.017148014440433214
Frequency of appearence for charachter 'F': 0.0090025270758122744
Frequency of appearence for charachter 'p': 0.0018050541516245488
Frequency of appearence for charachter 'v': 0.0036101083032490976
Frequency of appearence for charachter 'f': 0.02075812274368231
Frequency of appearence for charachter 'm': 0.023465703971119134
Frequency of appearence for charachter 'e': 0.1092057761732852
Frequency of appearence for charachter 'l': 0.046028880866425995
Frequency of appearence for charachter 'C': 0.015342960288808664
Frequency of appearence for charachter 'W': 0.0018050541516245488
Frequency of appearence for charachter 't': 0.05415162454873646
```

text2(Вильям шекспир - соннет 1)

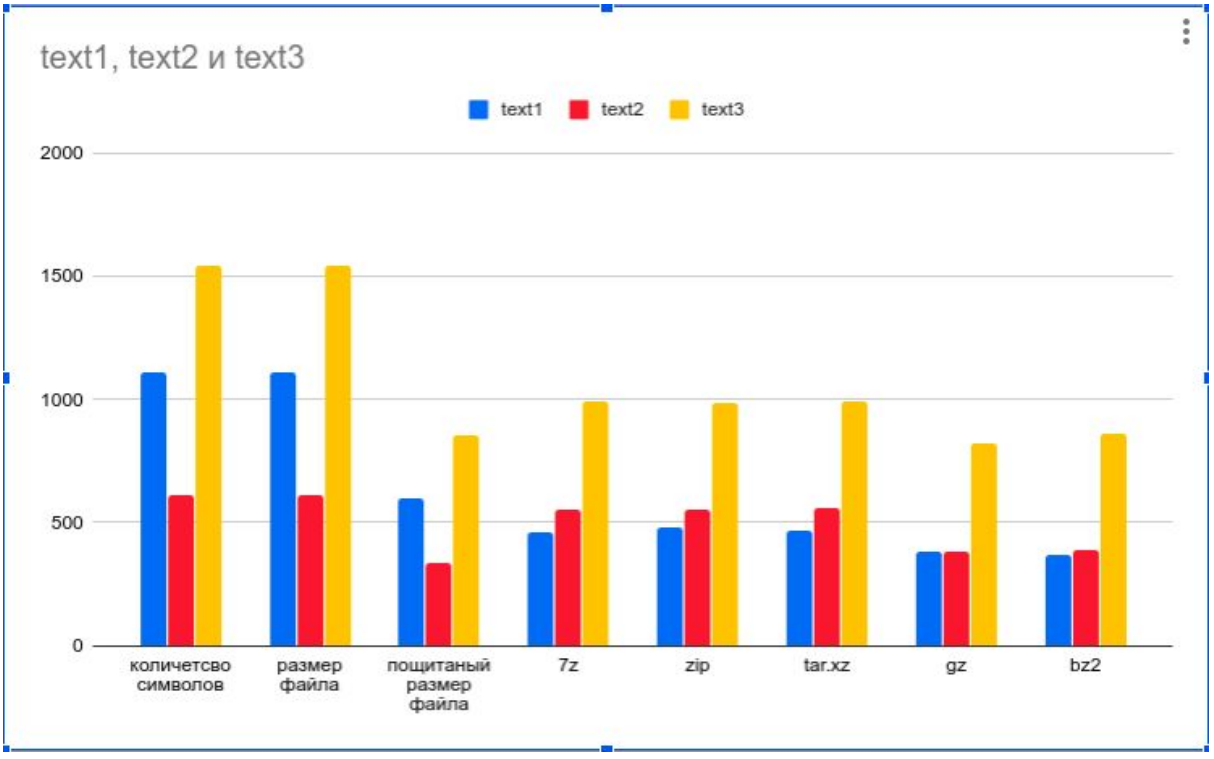
```
Frequency of appearence for charachter 'o': 0.040983606557377046
Frequency of appearence for charachter 'm': 0.01639344262295082
Frequency of appearence for charachter ':': 0.004918032786885246
Frequency of appearence for charachter 'w': 0.018032786885245903
Frequency of appearence for charachter 'W': 0.001639344262295082
Frequency of appearence for charachter 'i': 0.047540983606557376
Frequency of appearence for charachter 'f': 0.014754098360655738
Frequency of appearence for charachter 'H': 0.001639344262295082
Frequency of appearence for charachter '.': 0.001639344262295082
Frequency of appearence for charachter 'e': 0.11147540983606558
Frequency of appearence for charachter 'p': 0.003278688524590164
Frequency of appearence for charachter 'u': 0.027868852459016394
Frequency of appearence for charachter 'c': 0.014754098360655738
Frequency of appearence for charachter 'v': 0.003278688524590164
Frequency of appearence for charachter ',': 0.021311475409836064
Frequency of appearence for charachter 't': 0.08852459016393442
Frequency of appearence for charachter 'F': 0.003278688524590164
Frequency of appearence for charachter 'h': 0.054098360655737705
Frequency of appearence for charachter 'P': 0.001639344262295082
Frequency of appearence for charachter 'a': 0.04590163934426229
Frequency of appearence for charachter 'r': 0.054098360655737705
Frequency of appearence for charachter 'M': 0.001639344262295082
Frequency of appearence for charachter 'B': 0.003278688524590164
Frequency of appearence for charachter 'd': 0.03278688524590164
Frequency of appearence for charachter ' ': 0.15081967213114755
Frequency of appearence for charachter 'n': 0.047540983606557376
Frequency of appearence for charachter 'g': 0.019672131147540985
Frequency of appearence for charachter 's': 0.04918032786885246
Frequency of appearence for charachter 'y': 0.022950819672131147
Frequency of appearence for charachter 'k': 0.003278688524590164
Frequency of appearence for charachter 'b': 0.018032786885245903
Frequency of appearence for charachter ''': 0.009836065573770493
Frequency of appearence for charachter '-': 0.001639344262295082
Frequency of appearence for charachter '
': 0.022950819672131147
Frequency of appearence for charachter 'A': 0.003278688524590164
Frequency of appearence for charachter 'T': 0.006557377049180328
Frequency of appearence for charachter 'l': 0.029508196721311476
```

text3(PCI express text)

```
Frequency of appearence for charachter 'x': 0.0032404406999351912
Frequency of appearence for charachter '5': 0.0012961762799740765
Frequency of appearence for charachter 's': 0.0524951393389501
Frequency of appearence for charachter '9': 0.0019442644199611147
Frequency of appearence for charachter 'M': 0.0006480881399870382
Frequency of appearence for charachter '0': 0.005184705119896306
Frequency of appearence for charachter 'G': 0.0012961762799740765
Frequency of appearence for charachter 'r': 0.04212572909915749
Frequency of appearence for charachter 'w': 0.009721322099805573
Frequency of appearence for charachter 'A': 0.0019442644199611147
Frequency of appearence for charachter '—': 0.0006480881399870382
Frequency of appearence for charachter '’': 0.0012961762799740765
Frequency of appearence for charachter 'E': 0.0006480881399870382
Frequency of appearence for charachter '6': 0.0012961762799740765
Frequency of appearence for charachter 'O': 0.0006480881399870382
Frequency of appearence for charachter 'd': 0.0304601425793908
Frequency of appearence for charachter 'g': 0.019442644199611146
Frequency of appearence for charachter 'T': 0.002592352559948153
Frequency of appearence for charachter 'b': 0.011665586519766688
Frequency of appearence for charachter 'L': 0.0006480881399870382
Frequency of appearence for charachter 'f': 0.014906027219701879
Frequency of appearence for charachter 'u': 0.014257939079714841
Frequency of appearence for charachter 'W': 0.0012961762799740765
Frequency of appearence for charachter 'n': 0.06934543097861309
Frequency of appearence for charachter 'a': 0.06934543097861309
Frequency of appearence for charachter 'e': 0.084899546338302
Frequency of appearence for charachter '-': 0.0038885288399222295
Frequency of appearence for charachter 'k': 0.0064808813998703824
Frequency of appearence for charachter 'l': 0.026571613739468567
Frequency of appearence for charachter '
': 0.0019442644199611147
Frequency of appearence for charachter ' ': 0.15294880103694103
Frequency of appearence for charachter '2': 0.0006480881399870382
Frequency of appearence for charachter 'i': 0.07193778353856124
Frequency of appearence for charachter '4': 0.0006480881399870382
Frequency of appearence for charachter '.': 0.007128969539857421
Frequency of appearence for charachter 'c': 0.04342190537913156
Frequency of appearence for charachter '1': 0.0032404406999351912
Frequency of appearence for charachter 'v': 0.007128969539857421
Frequency of appearence for charachter 't': 0.07712248865845756
Frequency of appearence for charachter '8': 0.0006480881399870382
Frequency of appearence for charachter 'q': 0.0006480881399870382
Frequency of appearence for charachter 'C': 0.0012961762799740765
Frequency of appearence for charachter 'P': 0.0012961762799740765
Frequency of appearence for charachter 'K': 0.0006480881399870382
Frequency of appearence for charachter 'h': 0.01814646791963707
Frequency of appearence for charachter 'y': 0.009721322099805573
Frequency of appearence for charachter 'o': 0.061568373298763863
Frequency of appearence for charachter 'I': 0.0019442644199611147
Frequency of appearence for charachter 'm': 0.026571613739468567
Frequency of appearence for charachter '/': 0.0012961762799740765
Frequency of appearence for charachter ',': 0.010369410239792612
Frequency of appearence for charachter 'p': 0.017498379779650033
Frequency of appearence for charachter 'B': 0.0012961762799740765
Frequency of appearence for charachter 'U': 0.0006480881399870382
```

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).

	text1	text2	text3
количетсво символов	1108	610	1543
энтропия	4.341873071695859	4.379146675429261	4.421311504550075
размер файла	1108	610	1543
пощитаный размер файла	601	333	852
7z	463	553	996
zip	480	551	989
tar.xz	464	556	996
gz	383	383	821
bz2	371	386	861



Як бачимо, кращим методом кодування, виявився gzip. Також можемо помітити, що розміри архівів більше ніж розмір порахованої інформації, крім першого тексту, з ним відбулось щось незрозуміле.

2. Дослідження способів кодування інформації на прикладі Base64

1. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)
    a. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)

text1.txt

online (https://www.base64encode.org/)

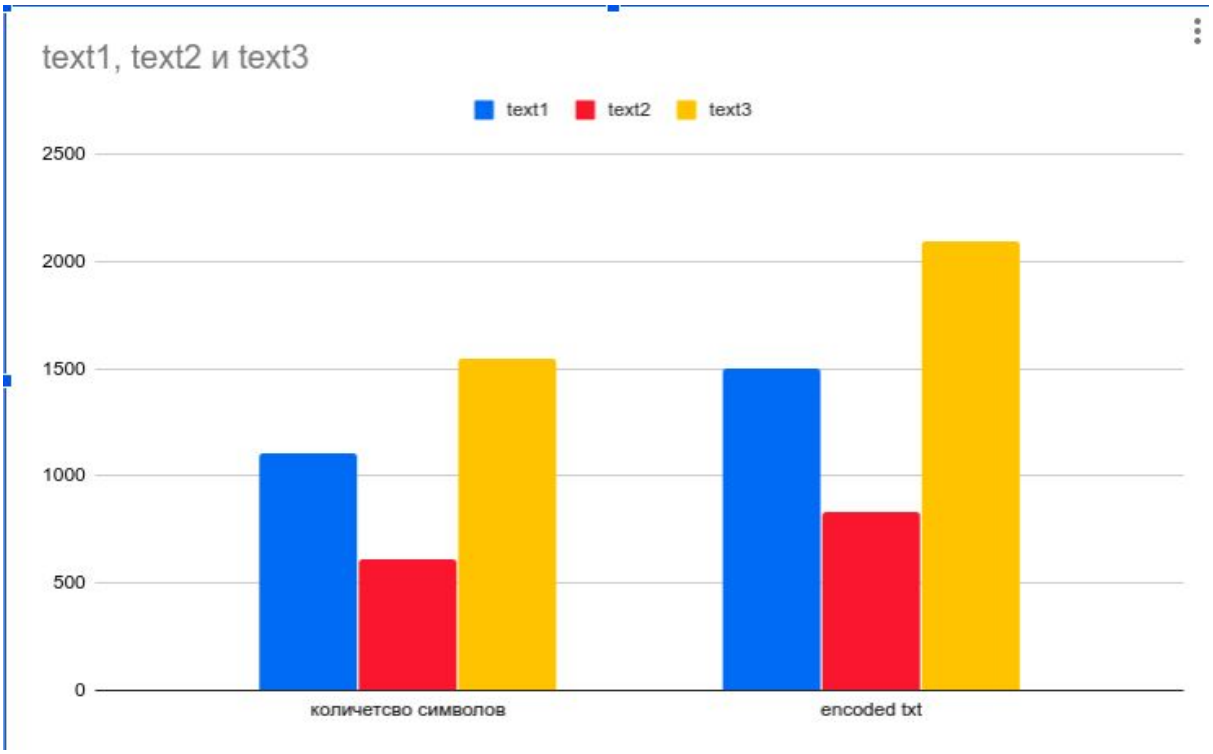
1. Закодуйте в Base64 обрані вами текстові файли
    a. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
    b. Порівняйте отримане значення з кількістю інформації вихідного файлу
    c. Зробіть висновки з отриманого результату

text2.txt

YiJGcm9tIGZhaXJlc3QgY3JlYXR1cmVzIHdlIGRlc2lyZSBpbmNyZWFzZSxcblRo
YXQgdGhlcmVieSBiZWF1dHkncyByb3NlIG1pZ2h0IG5ldmVyIGRpZSxcbkJ1dCBh
cyB0aGUgcmlwZXIgc2hvdWxkIGJ5IHRpbWUgZGVjZWFzZSxcbkhpcyB0ZW5kZXIg
aGVpciBtaWdodCBiZWFyIGhpcyBtZW1vcnk6XG5CdXQgdGhvdSBjb250cmFjdGVk
IHRvIHRoaW5lIG93biBicmlnaHQgZXllcyxcbkZlZWQnc3QgdGh5IGxpZ2h0J3Mg
ZmxhbWUgd2l0aCBzZWxmLXN1YnN0YW50aWFsIGZ1ZWwsXG5NYWtpbmcgYSBmYW1p
bmUgd2hlcmUgYWJ1bmRhbmNlIGxpZXMsXG5UaHkgc2VsZiB0aHkgZm9lLCB0byB0
aHkgc3dlZXQgc2VsZiB0b28gY3J1ZWw6XG5UaG91IHRoYXQgYXJ0IG5vdyB0aGUg
d29ybGQncyBmcmVzaCBvcm5hbWVudCxcbkFuZCBvbmx5IGhlcmFsZCB0byB0aGUg
Z2F1ZHkgc3ByaW5nLFxuV2l0aGluIHRoaW5lIG93biBidWQgYnVyaWVzdCB0aHkg
Y29udGVudCxcbkFuZCB0ZW5kZXIgY2h1cmwgbWFrJ3N0IHdhc3RlIGluIG5pZ2dh
cmRpbmc6XG5QaXR5IHRoZSB3b3JsZCwgb3IgZWxzZSB0aGlzIGdsdXR0b24gYmUs
XG5UbyBlYXQgdGhlIHdvcmxkJ3MgZHVlLCBieSB0aGUgZ3JhdmUgYW5kIHRoZWUu
XG4i

text3.txt

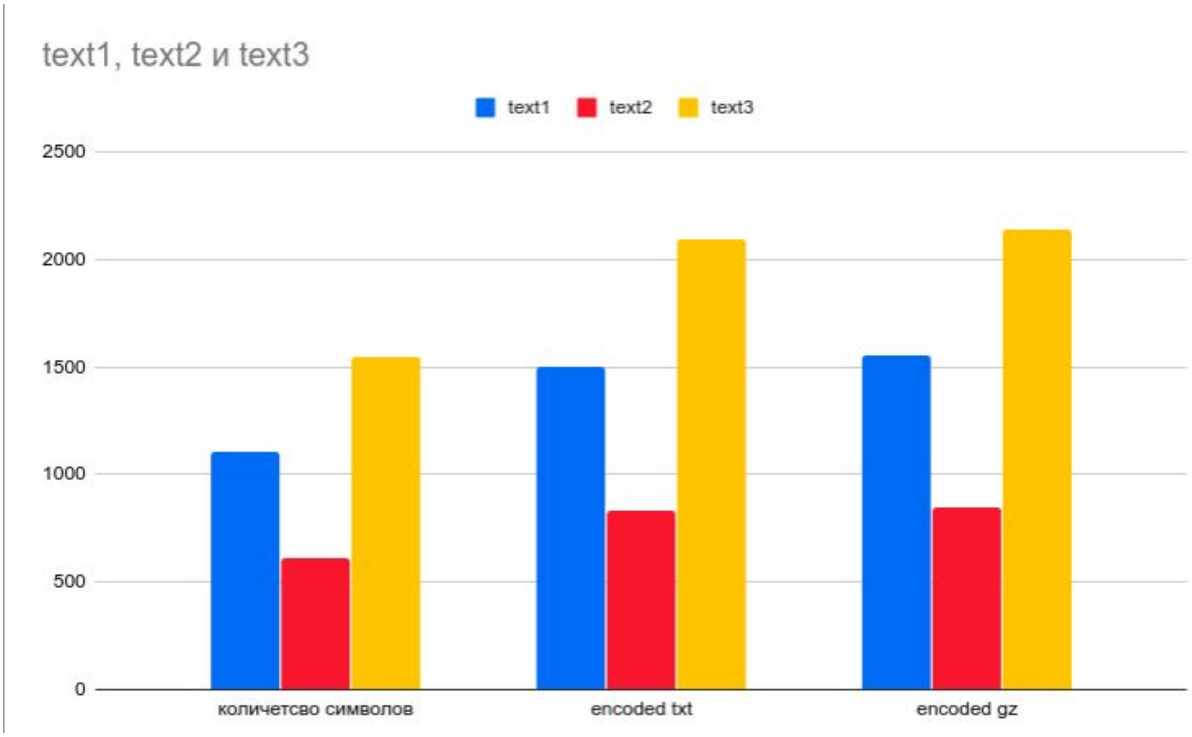
YidXaXRoIHRoZSB0ZWNobm9sb2d5IGltcHJvdmVtZW50cywgbWFzc2l2ZSBkYXRh
IHN0b3JhZ2UgYW5kIGZhc3QgcHJvY2Vzc29ycyBhcmUgZW5hYmxpbmcgbmV3IGFu
ZCBleGNpdGluZyBjYXBhYmlsaXRpZXMuIEFydGlmaWNpYWwgaW50ZWxsaWdlbmNl
LCB2aXJ0dWFsIHJlYWxpdHksIGFuZCBzZWxmLWRyaXZpbmcgY2FycyB3ZXJlIG9u
bHkgaW4gc2NpZW5jZSBmaWN0aW9uIG1vdmllc1x4ZTJceDgwXHg5NG5vdyBpbiAy
MDE4IHdlIGFyZSBleHBlcmllbmNpbmcgdGhlbSBmaXJzdCloYW5kLiBXaXRoIG1h
bnkgb2YgdGhlc2UgYXBwbGljYXRpb25zLCBkYXRhIGFuZCB0aGUgYWJpbGl0eSB0
byBwcm9jZXNzIHRoYXQgZGF0YSB3aXRoIGV4dHJlbWVseSBsb3cgbGF0ZW5jeSBo
YXMgYmVlbiBjcml0aWNhbCB0byBtYWtpbmcgdGhlbSBwb3NzaWJsZS5cbkxldFx4
ZTJceDgwXHg5OXMgZGlnIGRlZXBlciBpbnRvIHNvbWUgb2YgdGhlIHRlY2hub2xv
Z2ljYWwgYWR2YW5jZW1lbnRzLCBpbiBwYXJ0aWN1bGFyIHN0b3JhZ2UsIG11bHRp
LWNvcmUgQ1BVcyBhbmQgUENJIEV4cHJlc3MsIHRoYXQgYXJlIGJyaW5naW5nIHNj
aWVuY2UgZmljdGlvbiBmYW50YXNpZXMgdG8gbGlmZS5cbk9uZSBvZiB0aGUgY3Jp
dGljYWwgY29udHJpYnV0b3JzIHRvIHRoaXMgd29ybGQgb2YgbmV3IGNhcGFiaWxp
dHkgaXMgdGhlIHRlY2hub2xvZ3kgYWR2YW5jZW1lbnRzIG1hZGUgaW4gZGF0YSBj
b21tdW5pY2F0aW9ucy4gVGFraW5nIHlvdSBiYWNrIHRvIDE5OTksIGNvbm5lY3Rp
bmcgdG8gdGhlIGludGVybmV0IHRvIGFjY2VzcyBkYXRhIG9jY3VycmVkIG92ZXIg
Y29wcGVyIHRlbGVwaG9uZSBsaW5lcyBjb25uZWN0ZWQgdG8gYSA1NiBLQiBtb2Rl
bS4gQXQgNTYsMDAwIGJpdHMgcGVyIHNlY29uZCwgaXQgd291bGQgdGFrZSBhcHBy
b3hpbWF0ZWx5IDEwIG1pbnV0ZXMgdG8gZG93bmxvYWQgYW4gYXZlcmFnZSA0IE1C
IHNvbmcuXG5Ub2RheSwgZmliZXIgb3B0aWMgY29tbXVuaWNhdGlvbnMgd2l0aCBt
dWx0aS1tb2RlIGZpYmVyIG9wdGljIG5ldHdvcmtzIHJ1bm5pbmcgYXQgc3BlZWRz
IHVwIHRvIDEwIEdiL3MgYW5kIHNpbmdsZS1tb2RlIGZpYmVyIG9wdGljIG5ldHdv
cmtzIGFyZSBhYmxlIHRvIHJlYWNoIHNwZWVkcyB1cCB0byAxMDAgR2IvcyB4gVGhh
dCBzb25nIG5vdyBkb3dubG9hZHMgaW4gZnJhY3Rpb25zIG9mIGEgc2Vjb25kLiBU
aGVzZSBkYXRhIGNvbW11bmljYXRpb24gYWR2YW5jZW1lbnRzLCBsZXQgdXMgdGFr
ZSBsYXJnZSBhbW91bnRzIG9mIGRhdGEgZ2VuZXJhdGVkIGZyb20gc2Vuc29ycyBp
biBtYW55IGRldmljZXMsIHF1aWNrbHkgcHJvY2VzcyBpdCBpbiBkYXRhIGNlbnRl
cnMgYW5kIHRyYW5zZmVyIGl0IGJhY2sgdG8gZW5kIHBvaW50cyBmb3IgcmVhbC10
aW1lIGRlY2lzaW9uIG1ha2luZy4gSW4gdHVybiwgdGhpcyBpcyBlbmFibGluZyBh
IHRyYW5zaXRpb24gZnJvbSBodW1hbiB0cmlnZ2VyZWQgY29tbXVuaWNhdGlvbnMg
dG8gQUkgdHJpZ2dlcmVkIGNvbW11bmljYXRpb25zLCBmYWNpbGl0YXRpbmcgdG9k
YXlceGUyXHg4MFx4OTlzIHNtYXJ0IGNpdGllcywgaG9tZXMgYW5kIGNhcnMuJw==



як бачимо, розмір файлів зріс, це пов’язано з кодуванням BASE64.

1. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
    a. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
    b. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу[1]
    c. Зробіть висновки з отриманого результату

1



як бачимо, після закодовування в BASE64 наших стисненних файлів, їх розмір став більше ніж просто Base64 файлів, закодованих з простого текстового файлу.

Висновок:Було досліджено вплив різних методів кодування інформації на її кількість. Також було порівнянно алгоритми стиснення та було обрано кращий з них для випадків, коли треба буде зекономити місце на носії (gzip). Ознайомився з алгоритмом кодування Base64.