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

Лабораторна робота №1
з курсу «Комп’ютерні системи»
студента 3 курсу
групи СА-КІ
Колоскова Нікіти

Київ
2019

1. Дослідження кількості інформації в тексті

Обрані тексти:

1. Ray Kurzweil «The Singularity is Near» (prologue)
2. Structured Procrestination
3. 3D Reconstruction of Human Face Based on Single or Several Images (up to section 2.4)

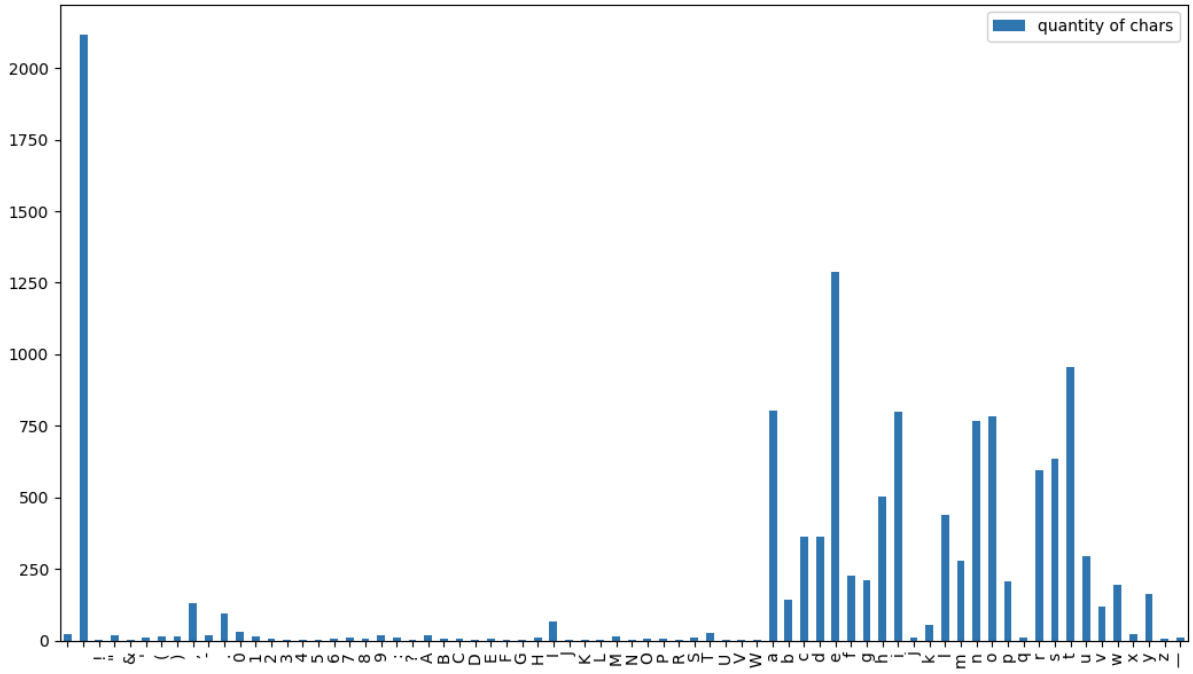
Посилання на GitHub репозиторій: https://github.com/Carterochka/cs

1.Ray Kurzweil «The Singularity is Near» (prologue)

```
Chars amounts in "The Singularity is Near" Prologue

: 24
 : 2117
!: 1
": 19
&: 1
': 10
(: 13
): 13
,: 129
-: 20
.: 96
0: 29
1: 16
2: 8
3: 4
4: 3
5: 2
6: 8
7: 9
8: 5
9: 17
:: 11
?: 1
A: 18
B: 6
C: 5
D: 4
E: 5
F: 2
G: 2
H: 10
I: 68
J: 4
K: 2
L: 1
M: 16
N: 1
O: 7
P: 7
R: 4
S: 12
T: 28
U: 1
V: 1
W: 4
a: 802
b: 141
c: 362
d: 364
e: 1288
f: 225
g: 209
h: 502
i: 799
j: 10
k: 54
l: 441
m: 278
n: 768
o: 782
p: 205
q: 12
r: 595
s: 635
t: 956
u: 294
v: 117
w: 193
x: 21
y: 164
z: 8
—: 11
```

Кількість символів у тексті

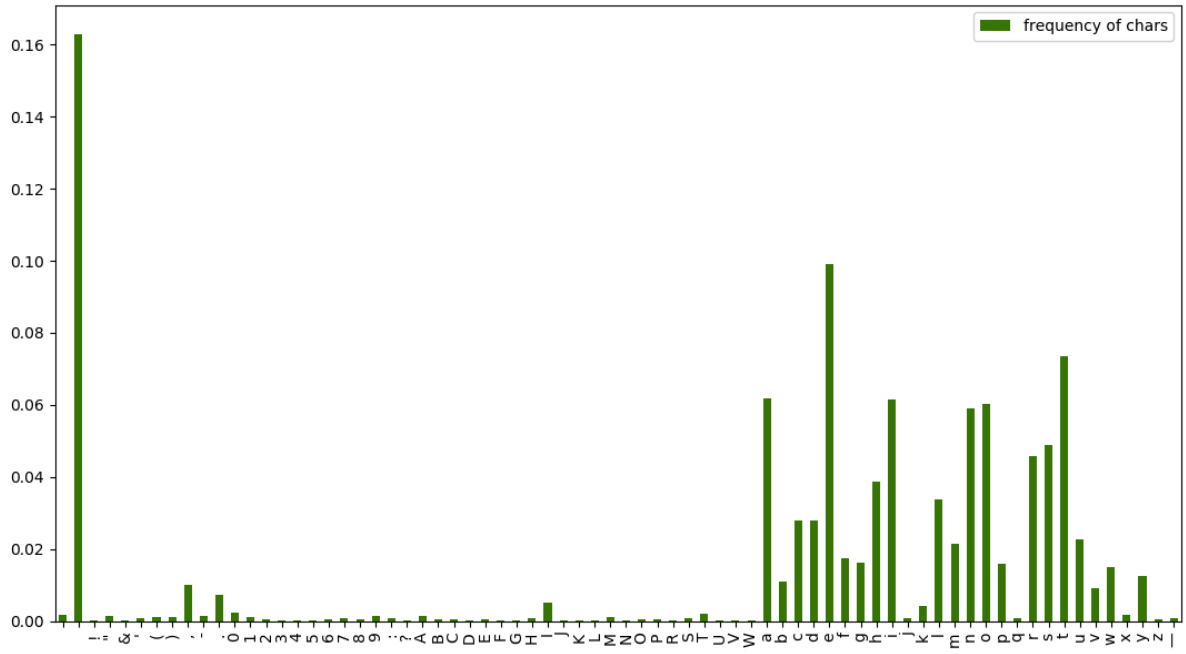


Гістограма кількості символів у тексті

```
Chars frequency in "The Singularity is Near" Prologue

: 0.001846
 : 0.162846
!: 0.000077
": 0.001462
&: 0.000077
': 0.000769
(: 0.001000
): 0.001000
,: 0.009923
-: 0.001538
.: 0.007385
0: 0.002231
1: 0.001231
2: 0.000615
3: 0.000308
4: 0.000231
5: 0.000154
6: 0.000615
7: 0.000692
8: 0.000385
9: 0.001308
:: 0.000846
?: 0.000077
A: 0.001385
B: 0.000462
C: 0.000385
D: 0.000308
E: 0.000385
F: 0.000154
G: 0.000154
H: 0.000769
I: 0.005231
J: 0.000308
K: 0.000154
L: 0.000077
M: 0.001231
N: 0.000077
O: 0.000538
P: 0.000538
R: 0.000308
S: 0.000923
T: 0.002154
U: 0.000077
V: 0.000077
W: 0.000308
a: 0.061692
b: 0.010846
c: 0.027846
d: 0.028000
e: 0.099077
f: 0.017308
g: 0.016077
h: 0.038615
i: 0.061462
j: 0.000769
k: 0.004154
l: 0.033923
m: 0.021385
n: 0.059077
o: 0.060154
p: 0.015769
q: 0.000923
r: 0.045769
s: 0.048846
t: 0.073538
u: 0.022615
v: 0.009000
w: 0.014846
x: 0.001615
y: 0.012615
z: 0.000615
—: 0.000846
```

Частота символів у тексті



Гістограма розподілу символів у тексті

Гістограми кількості та частот виглядають однаково, тому що по суті це ті самі ж значення, але у другому випадку пронормовані, щоб сума по усім символам дорівнювала одиниці.

```
Entropy: 4.412461
Quantity of information: 7170.248740
```

Ентропія та кількість інформації в оригінальному тексті

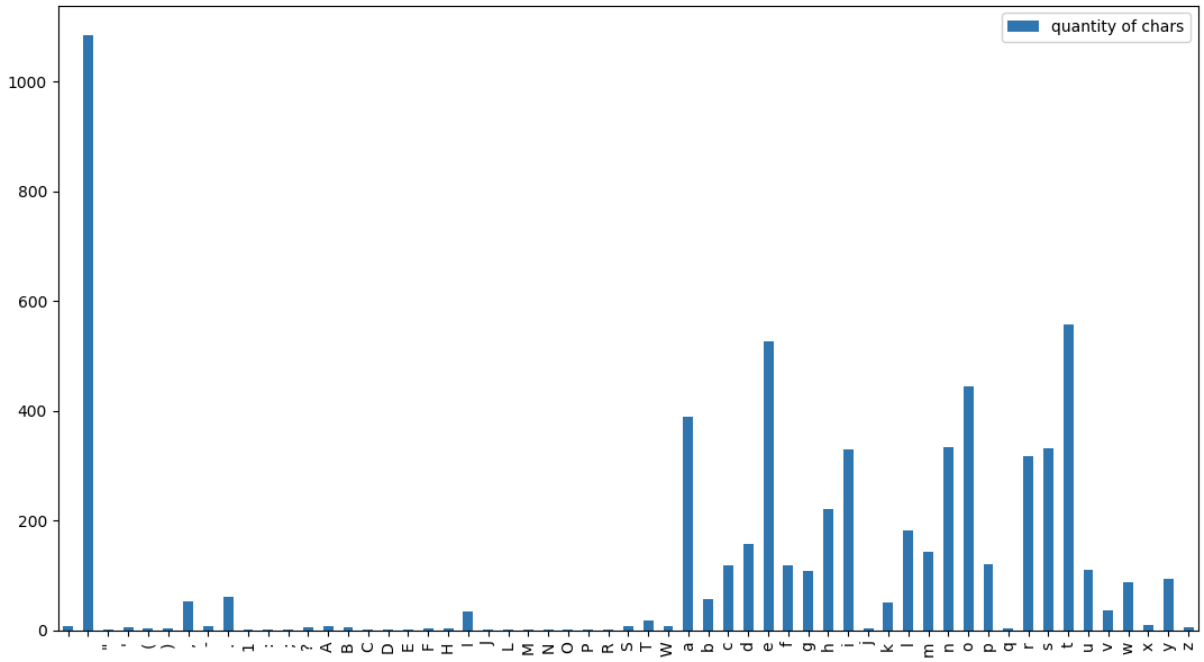
(цьо ще не всьо, див. далі)

2. Structured Procrastination

```
Chars amounts in "Structured Procrastination"

: 8
 : 1085
": 2
': 6
(: 3
): 3
,: 53
-: 7
.: 60
1: 1
:: 1
;: 2
?: 6
A: 8
B: 5
C: 1
D: 1
E: 2
F: 4
H: 3
I: 34
J: 1
L: 1
M: 1
N: 2
O: 2
P: 2
R: 2
S: 7
T: 18
W: 7
a: 390
b: 56
c: 119
d: 158
e: 527
f: 118
g: 109
h: 222
i: 329
j: 4
k: 50
l: 181
m: 144
n: 334
o: 444
p: 120
q: 3
r: 317
s: 331
t: 557
u: 110
v: 36
w: 87
x: 9
y: 94
z: 5
```

Кількість символів у тексті

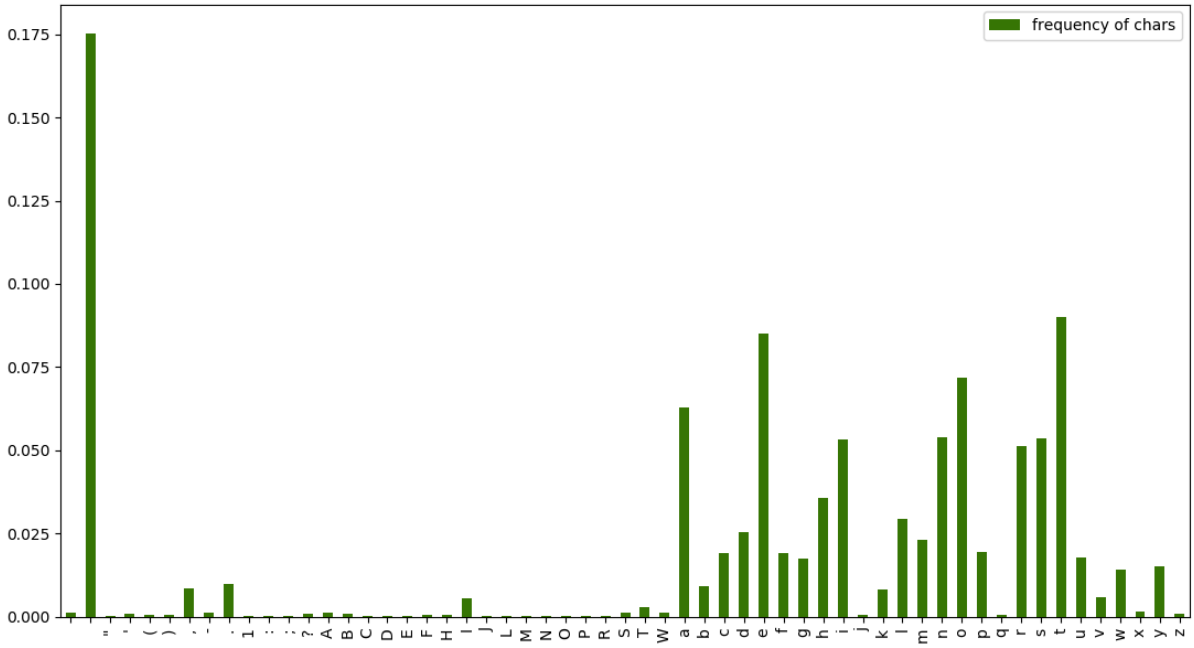


Гістограма кількості символів у тексті

```
Chars frequency in "Structured Procrastination"

: 0.001292
 : 0.175226
": 0.000323
': 0.000969
(: 0.000484
): 0.000484
,: 0.008559
-: 0.001130
.: 0.009690
1: 0.000161
:: 0.000161
;: 0.000323
?: 0.000969
A: 0.001292
B: 0.000807
C: 0.000161
D: 0.000161
E: 0.000323
F: 0.000646
H: 0.000484
I: 0.005491
J: 0.000161
L: 0.000161
M: 0.000161
N: 0.000323
O: 0.000323
P: 0.000323
R: 0.000323
S: 0.001130
T: 0.002907
W: 0.001130
a: 0.062984
b: 0.009044
c: 0.019218
d: 0.025517
e: 0.085110
f: 0.019057
g: 0.017603
h: 0.035853
i: 0.053133
j: 0.000646
k: 0.008075
l: 0.029231
m: 0.023256
n: 0.053941
o: 0.071705
p: 0.019380
q: 0.000484
r: 0.051195
s: 0.053456
t: 0.089955
u: 0.017765
v: 0.005814
w: 0.014050
x: 0.001453
y: 0.015181
z: 0.000807
```

Частота символів у тексті



Гістограма розподілу символів у тексті

```
Entropy: 4.305188
Quantity of information: 3332.215782
```

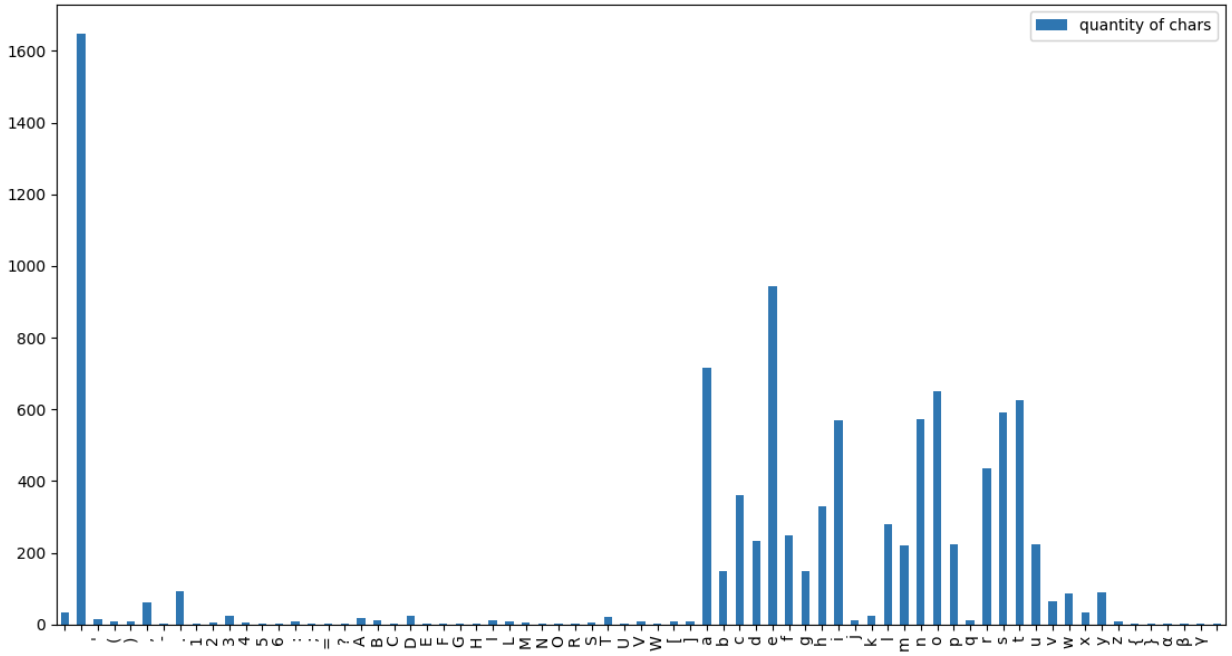
Ентропія та кількість інформації в оригінальному тексті

3. 3D Reconstruction of Human Face Based on Single or Several Images (up to section 2.4)

```
Chars amounts in "3D Reconstruction of Human Face..."

: 32
 : 1648
': 14
(: 9
): 9
,: 62
-: 3
.: 91
1: 2
2: 4
3: 24
4: 4
5: 1
6: 2
:: 9
;: 2
=: 1
?: 1
A: 19
B: 12
C: 3
D: 24
E: 1
F: 1
G: 1
H: 2
I: 12
L: 9
M: 5
N: 2
O: 1
R: 1
S: 6
T: 22
U: 1
V: 7
W: 3
[: 7
]: 7
a: 715
b: 149
c: 362
d: 232
e: 945
f: 248
g: 148
h: 329
i: 569
j: 10
k: 24
l: 278
m: 219
n: 574
o: 650
p: 223
q: 13
r: 435
s: 590
t: 626
u: 225
v: 65
w: 86
x: 32
y: 90
z: 8
{: 1
}: 1
α: 2
β: 1
γ: 2
 : 3
```

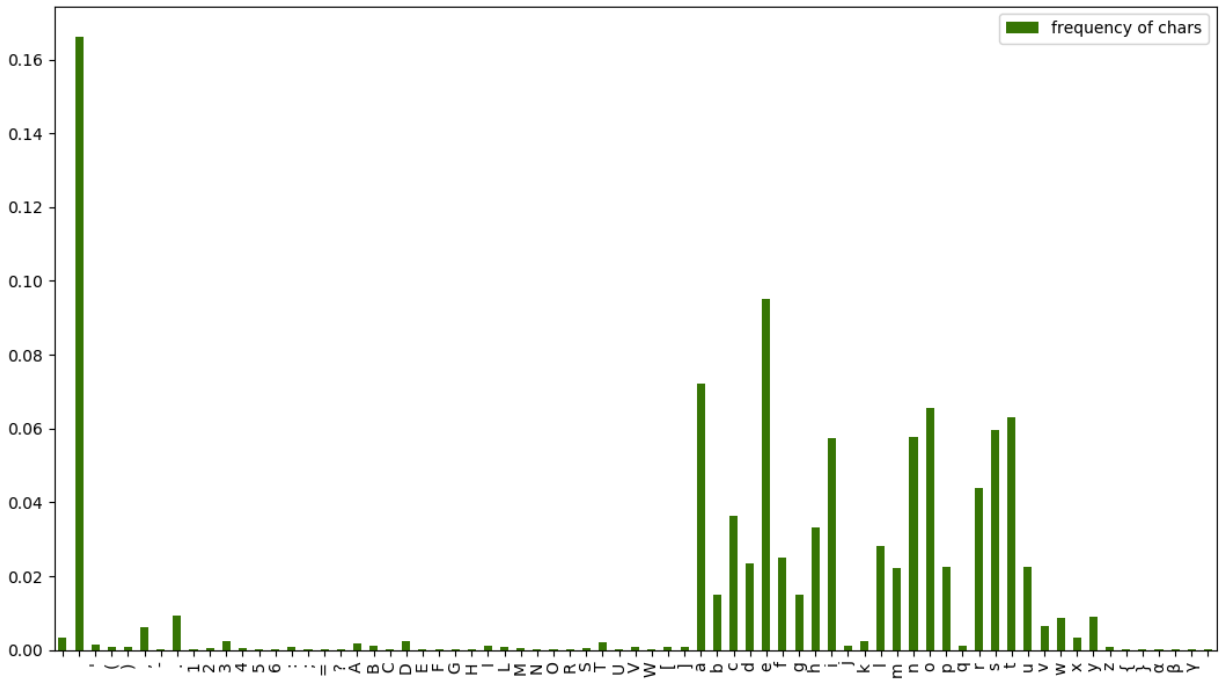
Кількість символів у тексті



Гістограма кількості символів

```
Chars frequency in "3D Reconstruction of Human Face..."
                                                        
: 0.003226                                              
 : 0.166146                                             
': 0.001411                                             
(: 0.000907                                             
): 0.000907                                             
,: 0.006251                                             
-: 0.000302                                             
.: 0.009174                                             
1: 0.000202                                             
2: 0.000403                                             
3: 0.002420                                             
4: 0.000403                                             
5: 0.000101                                             
6: 0.000202                                             
::: 0.000907                                            
;: 0.000202                                             
=: 0.000101                                             
?: 0.000101                                             
A: 0.001916                                             
B: 0.001210                                             
C: 0.000302                                             
D: 0.002420                                             
E: 0.000101                                             
F: 0.000101                                             
G: 0.000101                                             
H: 0.000202                                             
I: 0.001210                                             
L: 0.000907                                             
M: 0.000504                                             
N: 0.000202                                             
O: 0.000101                                             
R: 0.000101                                             
S: 0.000605                                             
T: 0.002218                                             
U: 0.000101                                             
V: 0.000706                                             
W: 0.000302                                             
[: 0.000706                                             
]: 0.000706                                             
a: 0.072084                                             
b: 0.015022                                             
c: 0.036496                                             
d: 0.023389                                             
e: 0.095272                                             
f: 0.025003                                             
g: 0.014921                                             
h: 0.033169                                             
i: 0.057365                                             
j: 0.001008                                             
k: 0.002420                                             
l: 0.028027                                             
m: 0.022079                                             
n: 0.057869                                             
o: 0.065531                                             
p: 0.022482                                             
q: 0.001311                                             
r: 0.043855                                             
s: 0.059482                                             
t: 0.063111                                             
u: 0.022684                                             
v: 0.006553                                             
w: 0.008670                                             
x: 0.003226                                             
y: 0.009073                                             
z: 0.000807                                             
{: 0.000101                                             
}: 0.000101                                             
α: 0.000202                                             
β: 0.000101                                             
γ: 0.000202                                             
 : 0.000302
```

Частота символів у тексті

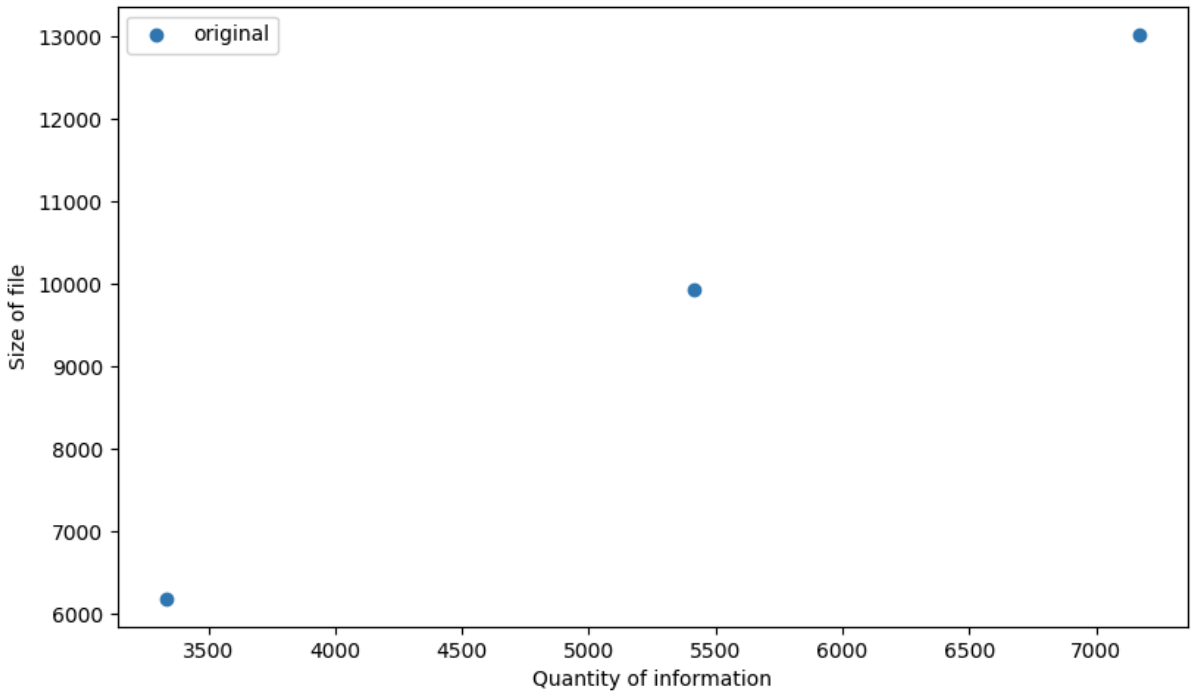


Гістограма розподілу символів у тексті

```
Entropy: 4.366933
Quantity of information: 5414.451134
```

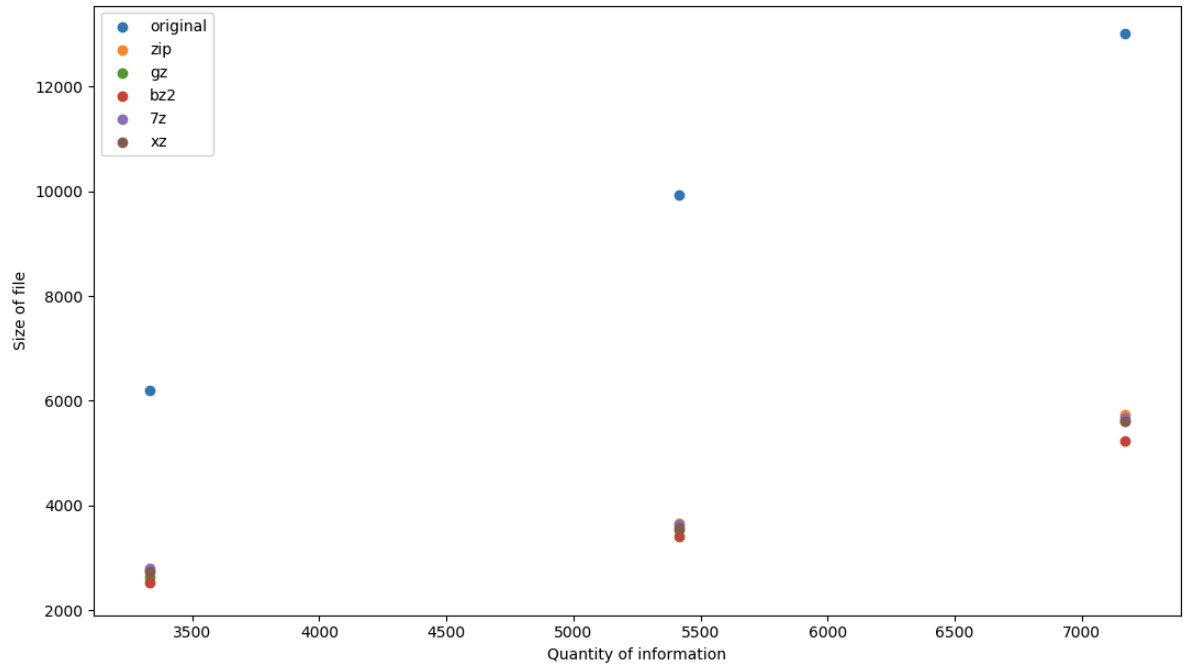
Ентропія та кількість інформації в оригінальному тексті

Для того, щоб можна було наглядно побачити тенденцію зміни розміру файлу в залежності від кількості інформації, тексти були вибрані таким чином, щоб їх довжина була різна.

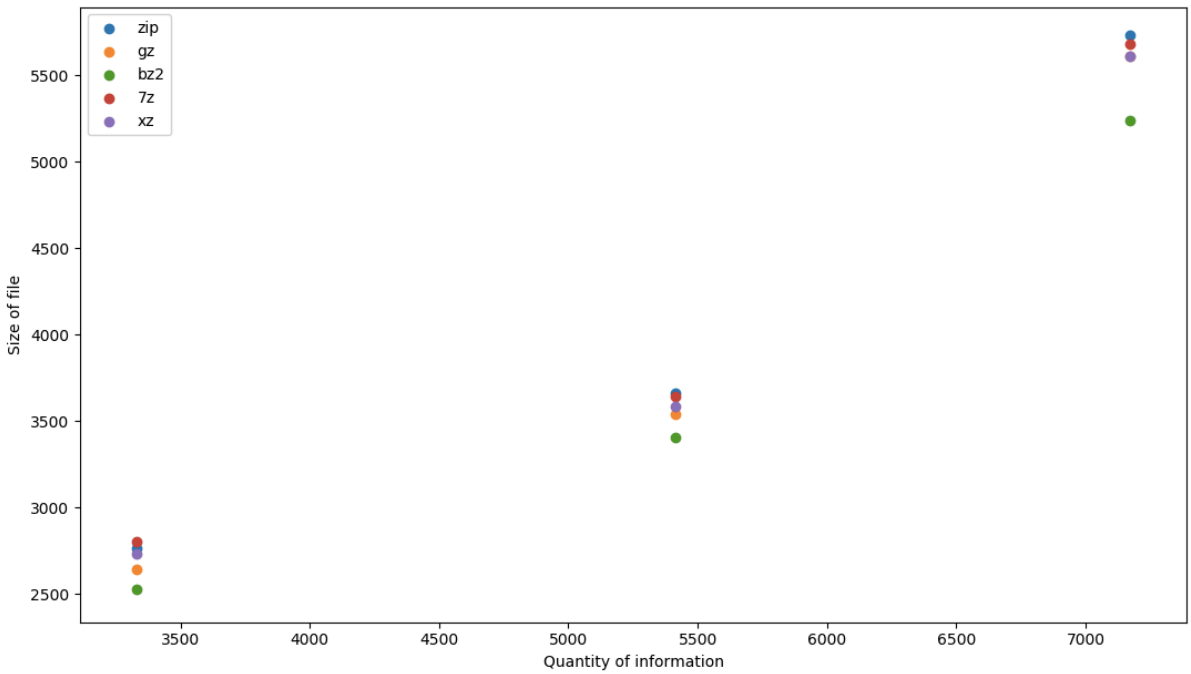


Залежність розміру файла (у байтах) від кількості інформації (у байтах)

Обрані алгоритми стиснення файлів: zip, gzip, bzip2, 7-zip, xz



Залежність розміру файла від кількості інформації в порівнянні з оригінальним файлом



та без порівняння з оригінальним файлом

З побаченого на останньому графіку можна зробити висновок, що:

1. ефективність різних алгоритмів змінюється із зміною розміру оригінального файлу (так, для малих розмірів оригінального файлу zip ефективніше за 7-zip, але зі збільшенням розміру 7-zip реалізує краще стискання. Аналогічна ситуація із gzip та xz - для малих розмірів xz стискає краще, але для найбільшого з наведених файлів їх результативність стала майже однаковою)
2. bzip2 усіх випереджає взагалі без шансів, і ця тенденція збільшується зі збільшенням розміру файлу

bzip2 - НАШ ПЕРЕМОЖЕЦЬ

3. для того, щоб зрозуміти ефективніше кореляцію між усіма алгоритмами, потрібно провести порівняння для ще декількох більших значеннях розміру файлу (більше даних - більш коректні висновки)

Загалом, алгоритм bzip2 і повинен стискати ефективніше за zip та gzip. Це зумовлено тим, що цей алгоритм використовує для стискання інший математичний апарат. Але за краще стискання доводиться платити більшим часом очікування та навантаженням на процесор.

Згідно інформації з вікіпедії, bzip2 повинен поступатися в ефективності алгоритму 7-zip, проте з отриманих результатів цього побачити не можна. Можливо, для більших розмірів файлів це дійсно було б так.

2. Дослідження способів кодування інформації на прикладі Base64

Перевірку роботи програми-кодувальника робив за допомогою онлайн-сервіса для кодування в base64 https://www.motobit.com/util/base64-decoder-encoder.asp та сервісу для порівняння двох текстів http://text.num2word.ru
Приклад роботи програми для кодування наведений в додатку А.

Результат порівняння:

(скріншоти фрагменту виводу порівняння на настурній сторінці)

Первый текст #FFFF00 **:**

vghyzwutzgltzw5zaw9uywwgkdneksbyzwnvbnn0cnvjdglvbibvzibhihnjzw5ligjhc2vkig9uigl0cybwag90b2dyyxb
ocybmb3jtcybhihzhc3qgc3ryyxr1bsbpbibjb250zw1wb3jhcnkgy29tchv0zxigdmlzaw9ulibtdgf0zsbvzib0agugyxj
0ihbyb3zpzgvzig5vignvbw1vbib0zwnobm9sb2d5ignhcgfibgugb2ygyxjiaxryyxj5ig9iamvjdcdzigrlbnnlidneihjly2
9uc3rydwn0aw9uigj5igegc2v0ig9mihbob3rvz3jhcghzihrha2vuigzyb20gdw5rbm93bib2awv3cg9pbnrzlibtcgfy
c2ugcmvjb25zdhj1y3rpb24gb2ygcmlnawqgc2nlbmvziglzihdlbgwgc3r1zgllzcbzbybmyxiuieegc29mdhdhcmugc
mvjb25zdhj1y3rpbmcgy2ftzxjhihryywply3rvcmllcybhbmqgm0qgbw9kzwwgb2ygysbzy2vuzsbhcybhihbvaw50i
gnsb3vkiglzigf2ywlsywjszsbbmv0uiehvd2v2zxigc3vic2vxdwvudcb0cmlhbmd1bgf0aw9uig9mihn1y2ggbw9kzw
xzignhbm5vdcbizsbhy2n1cmf0zwx5igrvbmugd2l0ag91dcbodw1hbibpbnrlcmfjdglvbibbml0uie1vcmvvdmvylcbp
dcbpcybub3qgywx3yxlzihbvc3npymxlihrvigfjaglldmugyxv0b21hdgljihjly29uc3rydwn0aw9uigv2zw4gc2lnbmlm
awnhbnrsesbyzwr1y2luzybhignsyxnzig9mig9iamvjdhmgdg8gymugchjvy2vzc2vkifszxs4gcjneihjly29uc3rydwn

Второй текст #00FF00 **:**

IG9mIGEgc2VnbWVudCBBQiB0byBhIHNlZ21lbnQgQSdCJywgdGhlbiB3ZSBnZXQgYSBjb3JyZWN0
IGF2ZXJhZ2luZyAoZmlnLiA2KS4gSW5kZWVkLCBhIHBvaW50IEEnJyBpcyBqdXN0IGhhbGYgd2F5
IGFsb25nIHRoZSBwb2ludHMgQSBhbmQgQScsIGFuZCBwb2ludCBCJycgaXMgZXhhY3RseSBpbiB0
aGUgbWlkZGxlIGJldHdlZW4gdGhlIHBvaW50cyBCIGFuZCBCJy4gClRodXMsIGluIG9yZGVyIHRv
IGdldCBhIHJlY3RhbmdsZSBieSBhdmVyYWdpbmcgdHdvIHJlY3RhbmdsZXMsIG9uZSBuZWVkcyB0
byBhdmVyYWdlIHNlbWFudGljYWxseSBpZGVudGljYWwsIGkuZS4gY29ycmVzcG9uZGVudCwgcG9p
bnRzLiBUaGlzIHJlYXNvbmluZyBpcyBhbHNvIGFwcGxpY2FibGUgZm9yIHRoZSBjYXNlIG9mIG11
Y2ggbW9yZSBjb21wbGljYXRlZCBjdXJ2ZXMgaW4gMkQgcGxhbmUgYW5kIHN1cmZhY2VzIGluIDNE
IHNwYWNlLiBJbiBwYXJ0aWN1bGFyLCBpdCBpcyBhcHBsaWNhYmxlIGZvciBzdXJmYWNlIG9mIGh1
bWFuIGZhY2UuIAoK

Поменять текст местами

Стереть

Время остановки сравнения: 3 секунд(ы)

Настройка результата отображения сравнения:

- (•) Семантическая очистка
- ( ) Эффективная очистка, значение редактирования (количество знаков суммирования): 4
- ( ) Без очистки

Сравнить

Тексты разные! В текстах присутствуют различающиеся фрагменты.

Результат сравнения:

===

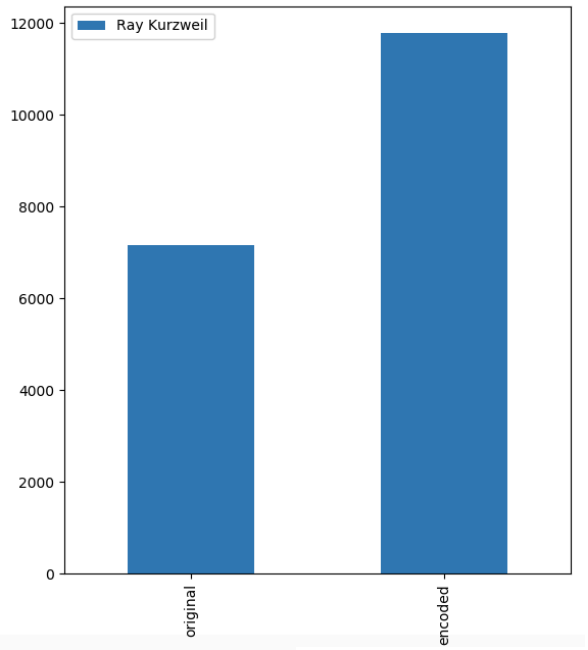
VGhyZWUtZGltZW5zaW9uYWwgKDNEKSByZWNvbnN0cnVjdGlvbiBvZiBhIHNjZW5lIGJhc2VkIG9u¶
IGl0cyBwaG90b2dyYXBocyBmb3JtcyBhIHZhc3Qgc3RyYXR1bSBpbiBjb250ZW1wb3JhcnkgY29t¶
cHV0ZXIgdmlzaW9uLiBTdGF0ZSBvZiB0aGUgYXJ0IHByb3ZpZGVzIG5vIGNvbW1vbiB0ZWNobm9s¶
b2d5IGNhcGFibGUgb2YgYXJiaXRyYXJ5IG9iamVjdCdzIGRlbnNlIDNEIHJlY29uc3RydWN0aW9u¶
IGJ5IGEgc2V0IG9mIHBob3RvZ3JhcGhzIHRha2VuIGZyb20gdW5rbm93biB2aWV3cG9pbnRzLiBT¶
cGFyc2UgcmVjb25zdHJ1Y3Rpb24gb2YgcmlnaWQgc2NlbmVzIGlzIHdlbGwgc3R1ZGllZCBzbyBm¶
YXIuIEEgc29mdHdhcmUgcmVjb25zdHJ1Y3RpbmcgY2FtZXJhIHRyYWplY3RvcmllcyBhbmQgM0Qg¶
bW9kZWwgb2YgYSBzY2VuZSBhcyBhIHBvaW50IGNsb3VkIGlzIGF2YWlsYWJsZSBbMV0uIEhvd2V2¶
bW9kZWwgb2YgYSBzY2VuZSBhcyBhIHBvaW50IGNsb3VkIGlzIGF2YWlsYWJsZSBbMV0uIEhvd2V2¶
ZXIgc3Vic2VxdWVudCB0cmlhbmd1bGF0aW9uIG9mIHN1Y2ggbW9kZWxzIGNhbm5vdCBiZSBhY2N1¶
cmF0ZWx5IGRvbmUgd2l0aG91dCBodW1hbiBpbnRlcmFjdGlvbiBbMl0uIE1vcmVvdmVyLCBpdCBp¶
cyBub3QgYWx3YXlzIHBvc3NpYmxlIHRvIGFjaGlldmUgYXV0b21hdGljIHJlY29uc3RydWN0aW9u¶
IGV2ZW4gc2lnbmlmaWNhbnRseSByZWR1Y2luZyBhIGNsYXNzIG9mIG9iamVjdHMgdG8gYmUgcHJv¶
Y2Vzc2VkIFszXS4gCjNEIHJlY29uc3RydWN0aW9uIG9mIGh1bWFuIGZhY2Ugb2NjdXBpZXMgYSBz¶
ZXBhcmF0ZSwgcmF0aGVyIHZhc3QsIG5pY2hlIGluIHRoZSBpbmRpY2F0ZWQgYXJlYS4gUmVzZWFy¶
Y2ggb24gdGhpcyB0b3BpYyBjYW4gYmUgZGl2aWRlZCBpbnRvIHR3byBsYXJnZSBjbGFzc2VzOiBn¶
ZW9tZXRyeSBvcmllbnRlZCBhbmQgbW9kZWwgYmFzZWQgbWV0aG9kcy4gR2VvbWV0cnkgb3JpZW50¶
b3JpZW50ZWQgYXBwcm9hY2hlcyBtb3N0bHkgcmVseSBvbiBlc3RhYmxpc2hpbmcgY29ycmVzcG9u¶
ZGVuY2UgYmV0d2VlbiBwb2ludHMgaW4gZGlmZmVyZW50IHBob3RvZ3JhcGhzIG9mIGFuIG9iamVj¶
dDsgc3VyZmFjZSByZWNvbnN0cnVjdGlvbiBpcyB0aGVuIHBlcmZvcm1lZCBieSBkZXRlcm1pbmlu¶
ZyBzdWNoIGxvY2F0aW9uIG9mIHBvaW50cyBpbiAzRCBzcGFjZSBhbmQgc3VjaCBjYW1lcmEgcGFy¶
YW1ldGVycyB3aGljaCBmaXQgaW4gdGhlIGJlc3Qgd2F5IHRvIHRoZSBwb2ludCBjb3JyZXNwb25k¶
ZW5jZS4gU3VjaCBhcHByb2FjaGVzIHJlcXVpcmUgZWl0aGVyIHN1YnN0YW50aWFsIG9yZ2FuaXph¶
dGlvbmFsIGRldmVsb3BtZW50OiBzcGVjaWFsIGNhbWVyYSBwb3NpdGlvbmluZyBhbmQgdHVuaW5n¶
LCBzcGVjaWFsIGlsbHVtaW5hdGlvbiBjb25kaXRpb25zLCBvciBzaWduaWZpY2FudCBhbW91bnQg¶
b2YgbWFudWFsIHdvcms6IGluZGljYXRpb24gb2YgcG9pbnQgY29ycmVzcG9uZGVuY2UsIHNlZ21l¶
bnRhdGlvbiBvZiBzY2VuZSBvYmplY3RzIGV0Yy4gCk1vcnBoYWJsZSBtb2RlbCBiYXNlZCB0ZWNo¶
bmlxdWVzIHV0aWxpemUgYW5hbHlzaXMtdGhyb3VnaC1zeW50aGVzaXMgYXBwcm9hY2guIApBIG1v¶
cnBoYWJsZSBtb2RlbCBvZiBodW1hbiBmYWNlIGNhbiBiZSBkZXBpY3RlZCBhcyBhIGJveCB3aXRo¶
IHR1bmFibGUgaGFuZGxlcy4gCkRlcGVuZGluZyBvbiBwb3NpdGlvbnMgb2YgdGhvc2UgaGFuZGxl¶
cyB0aGUgYm94IGdlbmVyYXRlcyBvbmUgb3IgYW5vdGhlciBwaWN0dXJlIG9mIG9uZSBvciBhbm90¶
aGVyIGh1bWFuIGZhY2UuIFRoZSBwcm9ibGVtIGlzIHRvIHNldCBoYW5kbGVzIG9mIHRob3NlIGJv¶
eCB0byBzdWNoIHBvc2l0aW9ucyB0aGF0IGl0IHdvdWxkIGdlbmVyYXRlIGV4YWN0bHkgdGhlIGlu¶
cHV0IGltYWdlLCBvciBhdCBsZWFzdCB0aGUgY2xvc2VzdCBwb3NzaWJsZS4gQSAzRCBzaGFwZSBv¶
ZiBhIGZhY2UgaXMgdW5pcXVlbHkgZGV0ZXJtaW5lZCBieSBwb3NpdGlvbnMgb2YgdGhvc2UgaGFu¶
ZGxlcyBvciwgc3BlYWtpbmcgbW9yZSBwcmVjaXNlbHksIGJ5IHBhcmFtZXRlcnMgb2YgdGhlIG1v¶
cnBoYWJsZSBtb2RlbC4gQW4gYWJpbGl0eSB0byByZWxpYWJseSByZWNvbnN0cnVjdCBhIDNEIHNo¶
YXBlIG9mIGEgaHVtYW4gZmFjZSBldmVuIGJ5IGEgc2luZ2xlIHBob3RvZ3JhcGggaXMgYSBjb2c2lk¶
ZXJhYmxlIGFkdmFudGFnZSBvZiBzdWNoIGFwcHJvYWNoZXMuIApWLiBCbGFueiBhbmQgVC4gVmV0¶
dGVyJ3MgcmVzZWFyY2ggWzNdIG9uIGRlc2lnbiBhbmQgdXNlIG9mIGEgbW9ycGhhYmxlIG1vZGVs¶
IG9mIGh1bWFuIGZhY2UgaXMgb25lIG9mIHRoZSBtb3N0IHdlbGwga25vd24gd29ya3MgaW4gdGhl¶
IGFyZWEuIFRoZXkgaGF2ZSBtYW5hZ2VkIHRvIGRlc2lnbiBhbiBhY2N1cmF0ZSBtb3JwaGFibGUg¶
bW9kZWwgYW5kIGltcGxlbWVudCBhbiBlZmZpY2llbnQgc2NoZW1lIG9mIHBhcnRpY3VsYXIgaW5k¶

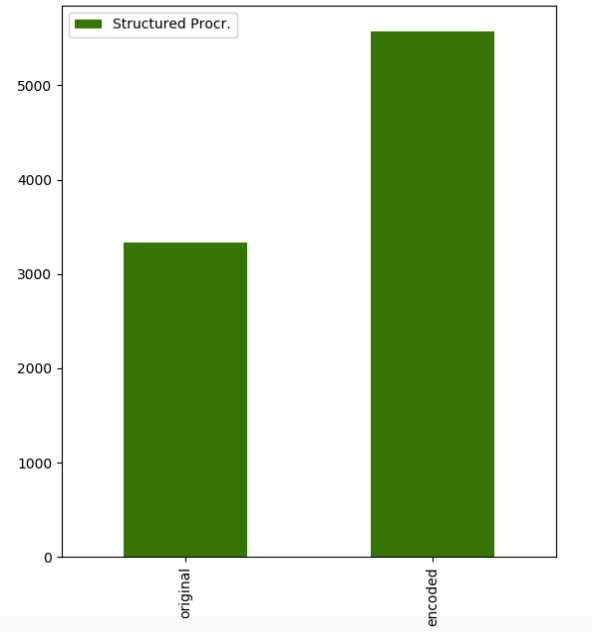
Як можна побачити, єдине, чим відрізнаються результати - це переходи на нову строку. Це зумовлено тим, що вивід написаної мною програми - сплошний текст, а результат на сайті розбитий на строки довжиною не більше 76 символів.

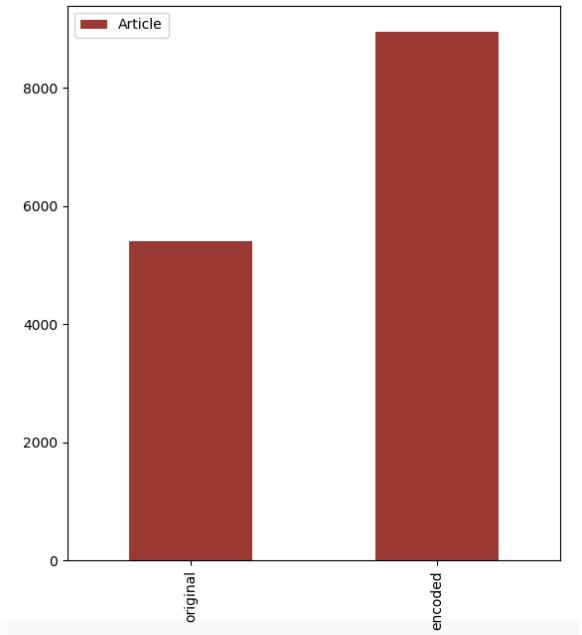
Отже, програма кодує вірно.

Кількість інформації в оригінальному та закодованому файлах:

	Original	Encoded
Ray Kurzweil	7170.248740	11778.310417
Structured Procr.	3332.215782	5567.081340
Article	5414.451134	8947.269930

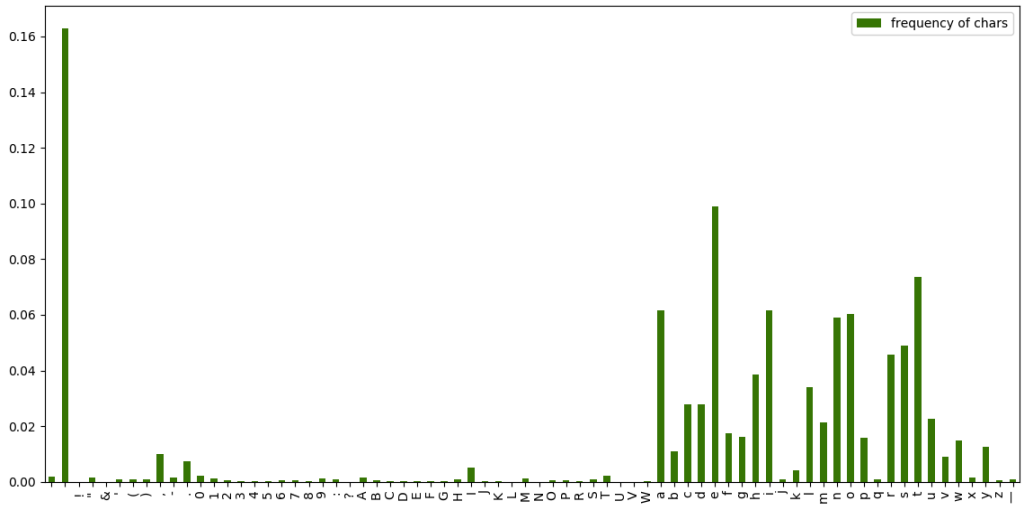




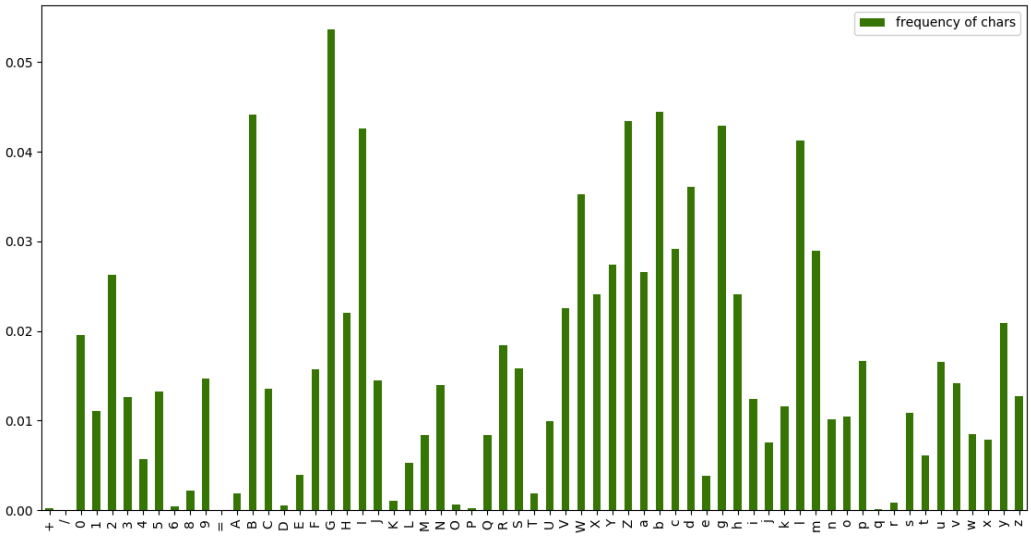


Отже, у кодуванні base64 кількість інформації збільшується, на прикладі текстів, що я використовував - приблизно в 1.65 разів. Це пов’язано з тим, що кількість символів, що міститься у закодованому файлі, більша, що випливає з опису кодування - грубо кажучі, кожним 8 оригінальним бітам відповідають 6 закодованих. Таким чином, збільшення кількості символів - 33%. Решта приросту може бути зумовлена тим, що кількість символів закодованого алгоритму більша за кількість символів оригінального.

Порівняння розподілу символів вихідного та закодованого алфавітів (на прикладі витягу з книги Рея Курцвела, для інших текстів ситуація аналогічна):



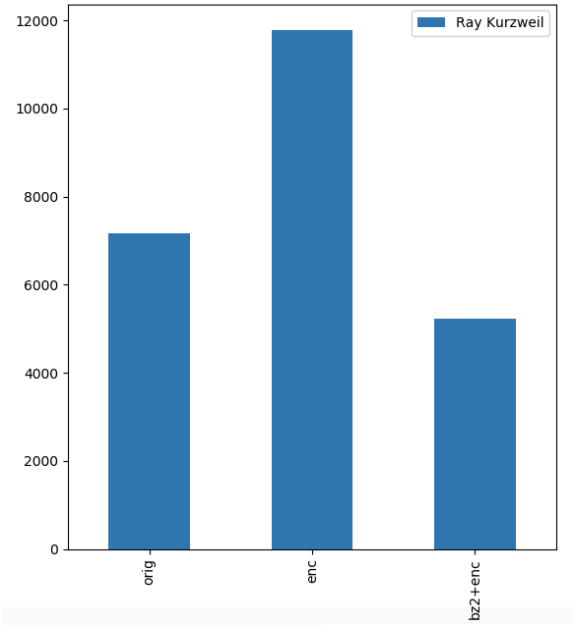
Розподіл символів алфавіту (оригінальний текст)

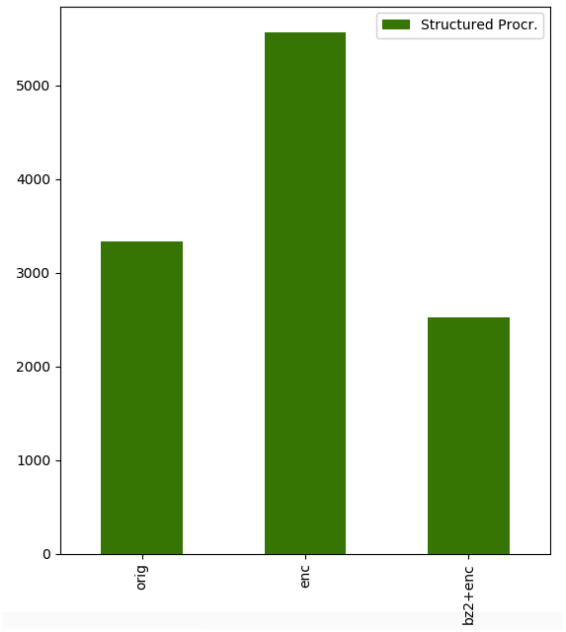


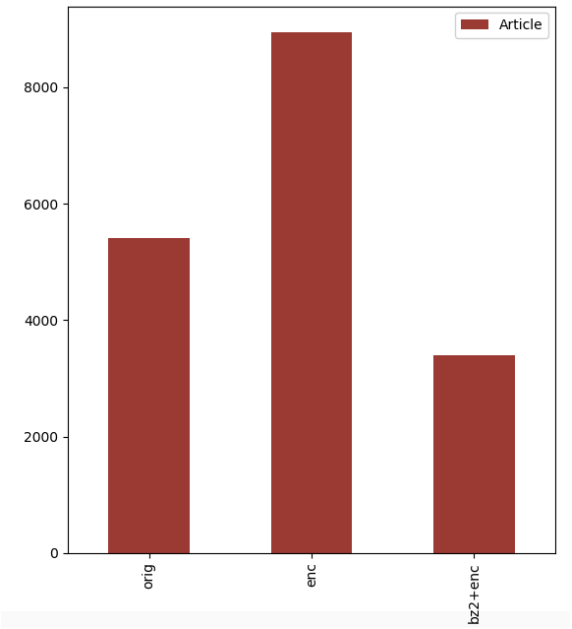
Закодований у base64 текст

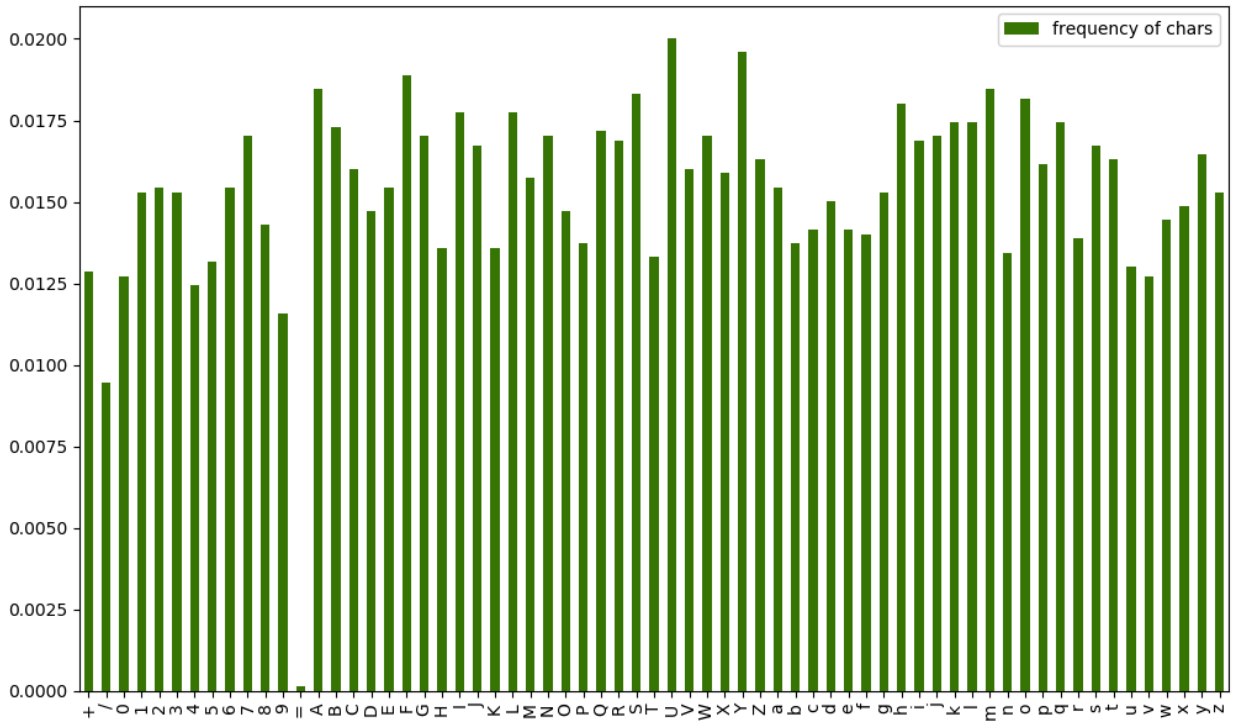
Порівняння кількості інформації в оригінальному, закодованому та закодованому стисненому файлах:

	Original	Encoded	Compressed+Encoded
Ray Kurzweil	7170.248740	11778.310417	5234.775139
Structured Procr.	3332.215782	5567.081340	2522.673411
Article	5414.451134	8947.269930	3391.485013









Розподіл символів алфавіту (стисений та закодований файл)

Отже, кількість інформації у стисненому та закодованому файлі становить ~0.7 від кількості інформації оригінального файлу. Зрозуміло, що кількість інформації повинна бути меншою -> bzip2 стискає розмір файлу більше ніж у 2,5 рази (судячи з графіків, отриманих у ч. 1 цієї лаби). Тоді виникає питання, чому ж тоді кількість інформації менша усього лише на ~0.3.

Можна побачити, що в закодованому після стиснення файлі усі символи алфавіту майже рівноймовірні (символ “=” зустрічається лише 1 раз наприкінці файлу, і означає, що кількість біт до кодування не була кратною 24).

При нерівноймовірному алфавіті $H = \sum_i p_i \, log_2 \frac{1}{p_i} \leq log_2 n$

При рівноймовірному алфавіті $H = log_2 n$

Таким чином, ентропія майже рівноймовірного алфавіту ближче до максимального значення, що також збільшило свій внесок у значення кількості інформації.

Висновки: за допомогою інструментів програмування можна обрахувати середню ентропію та кількість інформації в тексті та файлі. Ці величини пов’язані між собою наступним чином: кількість інформації дорівнює добутку середньої ентропії тексту на його кількість символів. При різних кодуваннях значення кількості інформації тексту буде відрізнятись; те саме можна сказати і про кодування після стискання.

ДОДАТОК А

Приклад роботи програми-кодувальника