Аналіз тексту

Для аналізу було взято переклад книги Mein Kampf.

Для зменшення впливу скорочень, спецсимволів і чисельно-символьно-буквених сполук, що не є "чистими" словами, за допомогою регексів з текстів видалено: дати виду 1927, 1934- $36\,$ гг., 1920-е, $1936\,$ z.; числа виду 12%, 12-13, 12/13; ініціали виду A., K.C. та інші символи, що не входять в рядок

chars = "абвгдежзийклмнопрстуфхцчшщэюяыьъ

Також замінено великі літери на малі та 'ë' на 'е', бо в повсякденному житті 'ë' майже не використовується. 'ь' та 'ъ' не замінювались, бо правила вживання 'ъ' можуть допомогти при аналізі.

Обсяг зі space-ом — 1 570 727 символів, без space-а — 1 354 233 символів

Аналіз літер

Аналіз біграм з перетином (crs) і без (seq)

```
letters - with '
                                        bgs, crs - with
     0.1378310808
                   216495
                                        'и '
                                              0.017563216 27587
                                                                     'и '
o'
    0.0967443738
                   151959
                                        'o '
                                              0.0174976412 27484
                                                                     'o '
    0.0747545563
                  117419
                                              0.015983055 25105
    0.0694296335 109055
                                              0.0153260339 24073
'a'
    0.0625538365 98255
                                         п'
                                              0.0150592783 23654
'н'
    0.0615453863 96671
                                          н'
                                              0.014145688 22219
    0.0576064459 90484
                                              0.0138025346 21680
    0.0504390642 79226
                                              0.0135262293
                                                            21246
'p'
    0.0421626419
                   66226
                                        'но'
                                              0.0130156374
                                                            20444
                                                                     'но'
    0.0401412849 63051
                                        'на'
                                              0.0118130088
                                                            18555
                                                                     'на'
    H1 = 4.4111
                                           H2 = 3.9868
letters - without ' '
                                        bgs, crs - without ' '
'o'
    0.1122104632 151959
                                        'ст'
                                              0.0179422861 24298
                                                                     'ст'
'e'
    0.0867052322 117419
                                        'но'
                                              0.0151694947
                                                            20543
                                                                     'но'
'и'
    0.0805290379 109055
                                        'ен'
                                              0.0142604917 19312
                                                                     'ен'
'a'
    0.072554038 98255
                                        'то'
                                              0.013770915 18649
                                                                     'на'
                                        'на'
'н'
    0.0713843714 96671
                                              0.0137266094 18589
                                                                     'то'
                                        'ни'
                                              0.012039305 16304
    0.0668157302 90484
                                                                     'oc'
                                        'oc'
'c'
                                              0.0119270641 16152
    0.0585025313
                   79226
                                                                     'ни'
                                        'ов'
                                              0.0118480525 16045
'р'
    0.0489029945
                   66226
                                                                     'ов'
                                        'po'
                                              0.010688723 14475
'в'
    0.0465584922
                   63051
                                                                     'po'
    0.0370099067
                                        'пр'
                                              0.0100795211 13650
                   50120
                                                                     'пр'
                                           H2 = 4.1008
   H1 = 4.4453
```

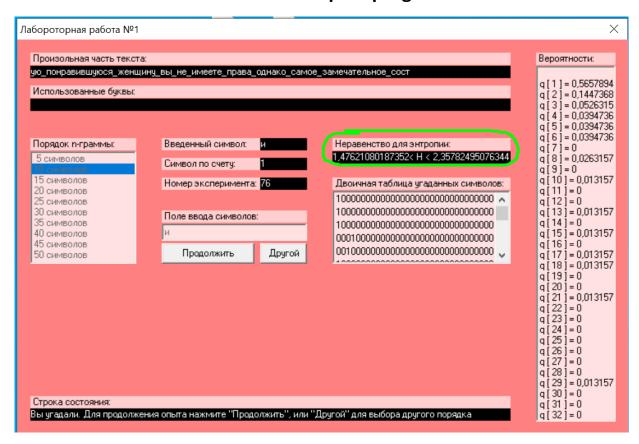
```
bgs, seq - with ' '
     0.017642797 13856
     0.0173779513
                   13648
     0.0159620456
                   12536
     0.0153292172 12039
     0.0148568242
                  11668
     0.0141539645 11116
     0.0139731564 10974
     0.0135084031
                   10609
     0.0129838049 10197
     0.0119116892
                   9355
   H2 = 3.9868
bgs, seq - without ' '
     0.0180072543 12193
     0.0152809858 10347
     0.0142516201
                   9650
     0.0137613053
                   9318
     0.0137037081
                   9279
     0.0119920368
                   8120
                   8042
     0.0118768424
     0.0117631248
                   7965
     0.0108120322
                   7321
     0.0100617915
                   6813
-- H2 = 4.1001
```

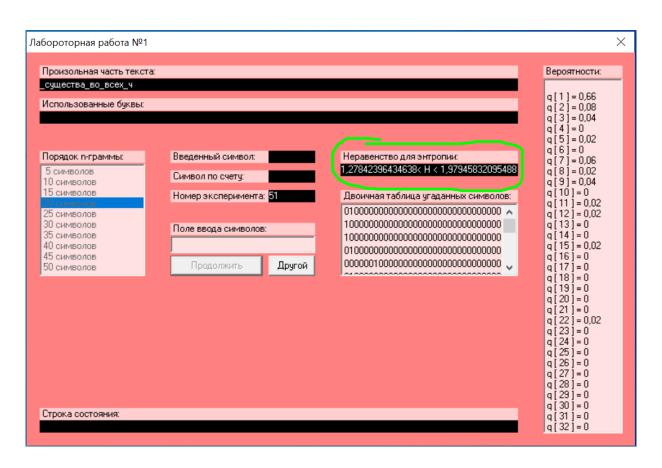
Видно, що зі space-ами Н в 1- і 2-граммі менше, бо space символ зустрічається помітно частіше за інші і, відповідно, якщо починати вгадувати у малих n-грамах з нього, то шанс вгадати найвищий.

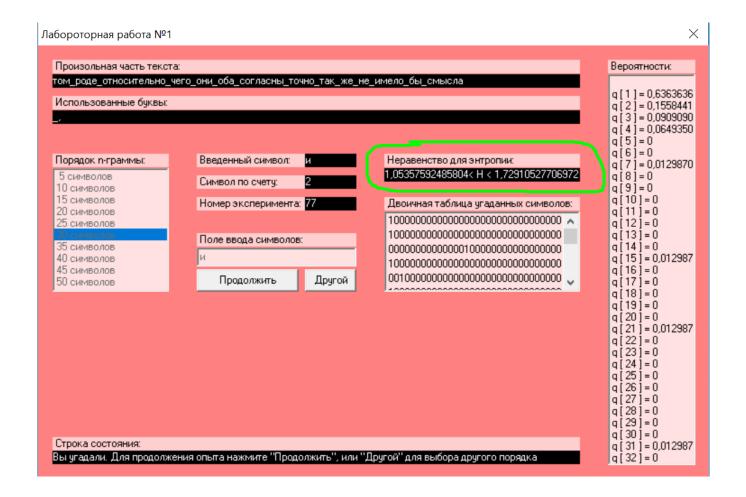
Також, через великий обсяг тексту, рахування з перетином і без мають майже однакові результати



Cool pink program







Помітно, що зі збільшенням довжини п-грами, збільшується "швидкість" вгадування і зменшується Н

Також, для великих n збільшення довжини n-грами до (n+k) дає відносно невеликий приріст шансу вгадування не пробільних символів (не $'_'$). Якщо ж треба вгадати наступний символ після $'_'$, то це зробити приблизно однаково складно для будь-яких великих n.

Надлишковість

Нехай
$$H_{\infty} \sim H^{30}, H^{(30)} \in (1.055; 1.729), H_0 = \log_2 32 = 5 \Longrightarrow$$

$$R \in \left(\left(1 - \frac{1.729}{5} \right); \left(1 - \frac{1.055}{5} \right) \right) \Longrightarrow R \in (\mathbf{0}.\mathbf{65}; \mathbf{0}.\mathbf{79})$$