Національний технічний університет України
“Київський політехнічний інститут”

Лабораторна робота №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали студенти:
Групи ФІ-93
Шашенок Микита
Медведь Михайло
Варінт №11

Київ 2022

Мета роботи

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела

Постановка задачі

Написати програми для обрахунку частот букв та біграм для тексту російською мовою (як з пробілами, так і без них) та відповідних значень H1 та H2, де ймовірності літер замінити частотами.
За допомогою програми CoolPinkProgram отримати експериментальні значення H(10), H(20), H(30). Використовуючи отримані дані, оцінити надлишковість російської мови у різних моделях.

Хід роботи

1.Частоти літер та біграм (з пробілами та без пробілів, з перетином біграм та без перетину):

```
probabilities_monograms_no_space
Letter = о Pr = 0.10596335628428055
Letter = а Pr = 0.08915859493523165
Letter = е Pr = 0.08402380674524448
Letter = и Pr = 0.06465165130120201
Letter = н Pr = 0.06278445559575213
Letter = т Pr = 0.058933364453261756
Letter = л Pr = 0.04901388726805928
Letter = в Pr = 0.04714669156260941
Letter = с Pr = 0.045746294783522
Letter = р Pr = 0.045162796125568914
Letter = к Pr = 0.03606021706150076
Letter = д Pr = 0.031508927529466685
Letter = у Pr = 0.029641731824016804
Letter = п Pr = 0.02684093826584199
Letter = м Pr = 0.025790640681526433
Letter = я Pr = 0.025557241218345197
Letter = г Pr = 0.023573345781304704
Letter = ь Pr = 0.019605554907223714
Letter = ч Pr = 0.018321857859726922
Letter = б Pr = 0.01715486054382075
Letter = з Pr = 0.01668806161745828
Letter = ы Pr = 0.01540436456996149
Letter = й Pr = 0.012020072353833585
Letter = ш Pr = 0.01096977476951803
Letter = ж Pr = 0.010269576379974325
Letter = х Pr = 0.009919477185202474
Letter = ю Pr = 0.004667989263624693
Letter = ц Pr = 0.003734391410899755
Letter = э Pr = 0.003734391410899755
Letter = щ Pr = 0.0030341930213560507
Letter = ф Pr = 0.0010502975843155562
Letter = ъ Pr = 0.00046679892636246936
```

```
probabilities_monograms_with_space
Letter = о Pr = 0.0880954690986708
Letter = а Pr = 0.07412438148830891
Letter = е Pr = 0.06985543805180945
Letter = и Pr = 0.05374987872319783
Letter = н Pr = 0.05219753565537984
Letter = т Pr = 0.04899582807800524
Letter = л Pr = 0.040749005530222177
Letter = в Pr = 0.03919666246240419
Letter = с Pr = 0.0380324051615407
Letter = р Pr = 0.03754729795284758
Letter = к Pr = 0.029979625497234888
Letter = д Pr = 0.026195789269428543
Letter = у Pr = 0.024643446201610558
Letter = п Pr = 0.022314931599883573
Letter = м Pr = 0.021441738624235956
Letter = я Pr = 0.02124769574075871
Letter = г Pr = 0.019598331231202096
Letter = ь Pr = 0.01629960221208887
Letter = ч Pr = 0.015232366352964006
Letter = б Pr = 0.014262151935577764
Letter = з Pr = 0.013874066168623266
Letter = ы Pr = 0.012806830309498399
Letter = й Pr = 0.009993208499078296
Letter = ш Pr = 0.009120015523430678
Letter = ж Pr = 0.008537886872998933
Letter = х Pr = 0.00824682254778306
Letter = ю Pr = 0.0038808576695449695
Letter = ц Pr = 0.0031046861356359757
Letter = э Pr = 0.0031046861356359757
Letter = щ Pr = 0.00252255748520423
Letter = ф Pr = 0.0008731929756476181
Letter = ъ Pr = 0.00038808576695449696
```

```
probabilities_bigram_no_space_no_overlap
Bigram = ов Pr = 0.015637764033142723
Bigram = не Pr = 0.01493756564359902
Bigram = то Pr = 0.013770568327692845
Bigram = ал Pr = 0.012370171548605438
Bigram = на Pr = 0.012370171548605438
Bigram = ел Pr = 0.011319873964289882
Bigram = по Pr = 0.011086474501108648
Bigram = го Pr = 0.010736375306336796
Bigram = ер Pr = 0.010619675574746178
Bigram = ра Pr = 0.010619675574746178
Bigram = ко Pr = 0.01050297584315556
Bigram = ен Pr = 0.009802777453611857
Bigram = он Pr = 0.009452678258840004
Bigram = ет Pr = 0.00921927879565877
Bigram = ст Pr = 0.008869179600886918
Bigram = ро Pr = 0.0087524798692963
Bigram = ка Pr = 0.008635780137705683
Bigram = но Pr = 0.008635780137705683
Bigram = ни Pr = 0.008519080406115065
Bigram = ак Pr = 0.00840238067452445
Bigram = ва Pr = 0.00840238067452445
Bigram = ос Pr = 0.008285680942933832
Bigram = ор Pr = 0.008052281479752597
Bigram = во Pr = 0.007935581748161979
Bigram = ло Pr = 0.007935581748161979
Bigram = од Pr = 0.007935581748161979
Bigram = ит Pr = 0.007702182284980745
Bigram = ть Pr = 0.00746878282179951
Bigram = ла Pr = 0.007352083090208892
Bigram = та Pr = 0.007118683627027658
Bigram = ин Pr = 0.00700198389543704
Bigram = ат Pr = 0.006885284163846423
Bigram = об Pr = 0.006885284163846423
```

```
probabilities_bigram_no_space_with_overlap
Bigram = ов Pr = 0.015637764033142723
Bigram = не Pr = 0.01493756564359902
Bigram = то Pr = 0.013770568327692845
Bigram = ал Pr = 0.012370171548605438
Bigram = на Pr = 0.012370171548605438
Bigram = ел Pr = 0.011319873964289882
Bigram = по Pr = 0.011086474501108648
Bigram = го Pr = 0.010736375306336796
Bigram = ер Pr = 0.010619675574746178
Bigram = ра Pr = 0.010619675574746178
Bigram = ко Pr = 0.01050297584315556
Bigram = ен Pr = 0.009802777453611857
Bigram = он Pr = 0.009452678258840004
Bigram = ет Pr = 0.00921927879565877
Bigram = ст Pr = 0.008869179600886918
Bigram = ро Pr = 0.0087524798692963
Bigram = ка Pr = 0.008635780137705683
Bigram = но Pr = 0.008635780137705683
Bigram = ни Pr = 0.008519080406115065
Bigram = ак Pr = 0.00840238067452445
Bigram = ва Pr = 0.00840238067452445
Bigram = ос Pr = 0.008285680942933832
Bigram = ор Pr = 0.008052281479752597
Bigram = во Pr = 0.007935581748161979
Bigram = ло Pr = 0.007935581748161979
Bigram = од Pr = 0.007935581748161979
Bigram = ит Pr = 0.007702182284980745
Bigram = ть Pr = 0.00746878282179951
Bigram = ла Pr = 0.007352083090208892
Bigram = та Pr = 0.007118683627027658
Bigram = ин Pr = 0.00700198389543704
Bigram = ат Pr = 0.006885284163846423
Bigram = об Pr = 0.006885284163846423
```

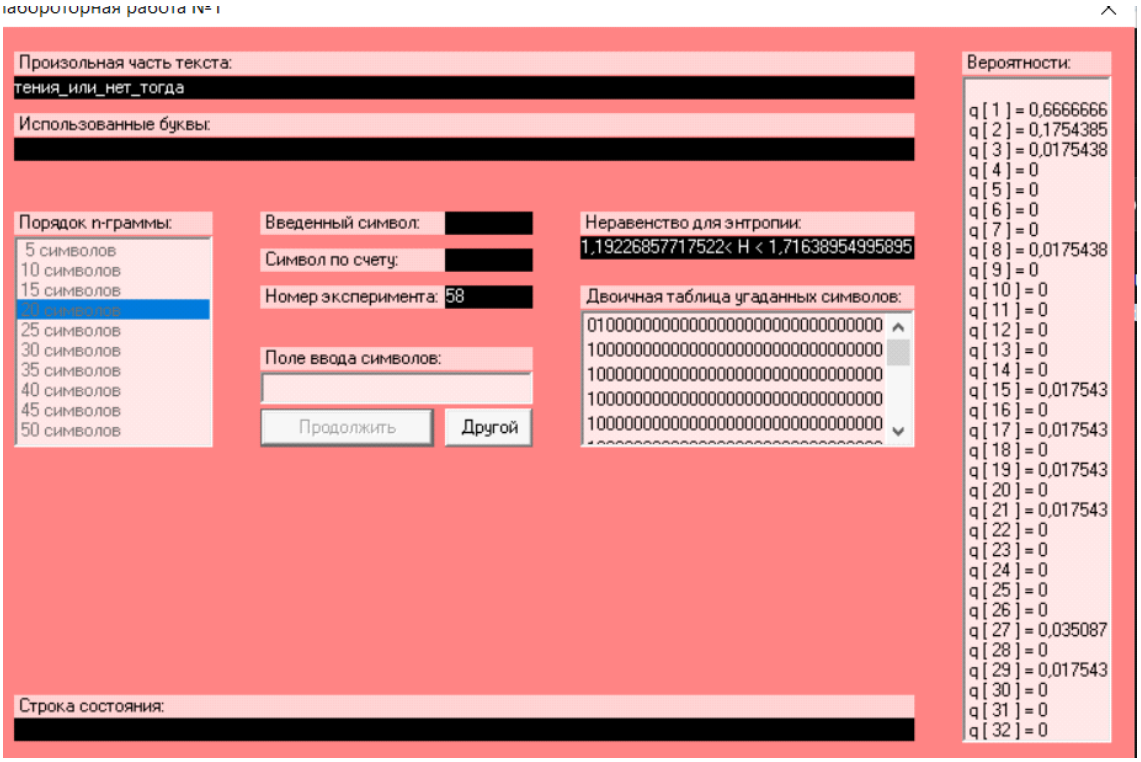
```
probabilities_bigram_with_space_no_overlap
Bigram = не Pr = 0.012418744542543903
Bigram = то Pr = 0.010963422916464538
Bigram = ал Pr = 0.010284272824294168
Bigram = на Pr = 0.010187251382555544
Bigram = ов Pr = 0.010187251382555544
Bigram = по Pr = 0.009217036965169302
Bigram = ел Pr = 0.009022994081692054
Bigram = ра Pr = 0.008828951198214805
Bigram = го Pr = 0.008731929756476181
Bigram = ер Pr = 0.008634908314737557
Bigram = ко Pr = 0.008343843989521683
Bigram = ро Pr = 0.007276608130396818
Bigram = ка Pr = 0.007179586688658194
Bigram = ст Pr = 0.007179586688658194
Bigram = ни Pr = 0.006888522363442321
Bigram = но Pr = 0.006791500921703697
Bigram = ет Pr = 0.006694479479965073
Bigram = ва Pr = 0.006597458038226448
Bigram = ак Pr = 0.0064034151547491995
Bigram = ен Pr = 0.0064034151547491995
Bigram = ор Pr = 0.0064034151547491995
Bigram = во Pr = 0.0063063937130105755
Bigram = ть Pr = 0.006209372271271951
Bigram = ло Pr = 0.0061123508295333265
Bigram = од Pr = 0.006015329387794702
Bigram = ит Pr = 0.005918307946056078
Bigram = та Pr = 0.005821286504317454
Bigram = ла Pr = 0.005627243620840205
Bigram = он Pr = 0.005433200737362957
Bigram = че Pr = 0.005433200737362957
Bigram = за Pr = 0.00504511497040846
Bigram = ос Pr = 0.00504511497040846
```

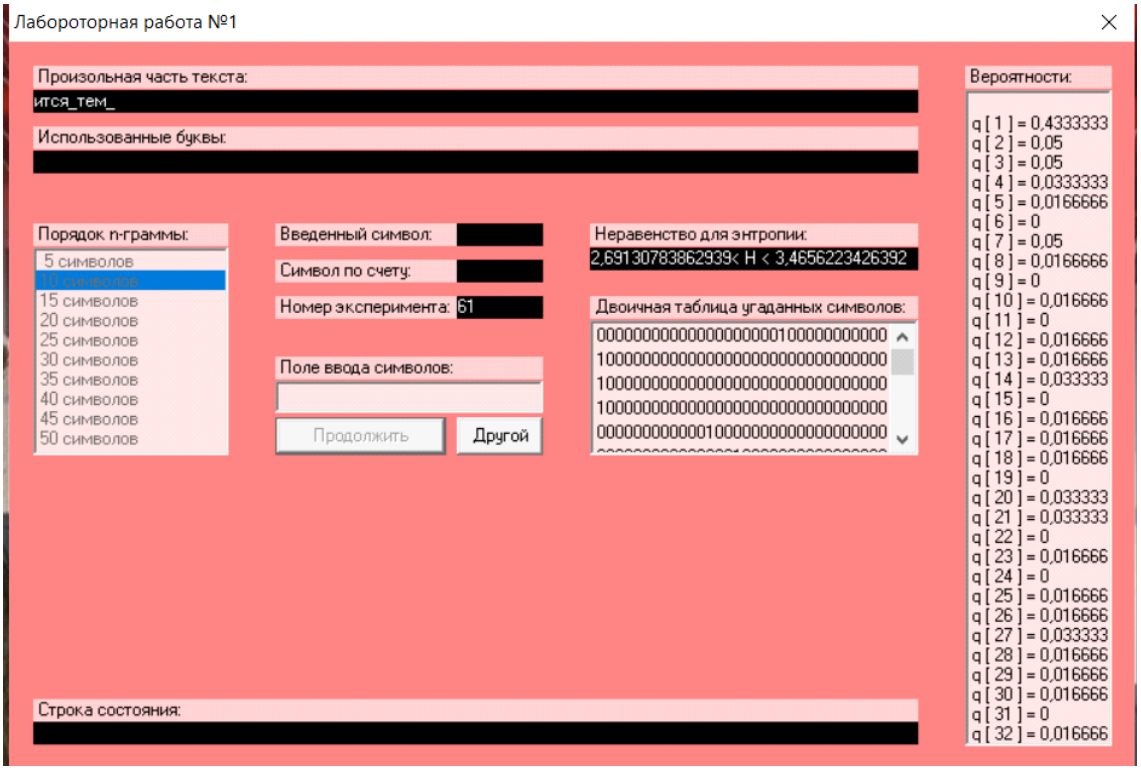
```
probabilities_bigram_with_space_with_overlap
Bigram = не Pr = 0.012418744542543903
Bigram = то Pr = 0.010963422916464538
Bigram = ал Pr = 0.010284272824294168
Bigram = на Pr = 0.010187251382555544
Bigram = ов Pr = 0.010187251382555544
Bigram = по Pr = 0.009217036965169302
Bigram = ел Pr = 0.009022994081692054
Bigram = ра Pr = 0.008828951198214805
Bigram = го Pr = 0.008731929756476181
Bigram = ер Pr = 0.008634908314737557
Bigram = ко Pr = 0.008343843989521683
Bigram = ро Pr = 0.007276608130396818
Bigram = ка Pr = 0.007179586688658194
Bigram = ст Pr = 0.007179586688658194
Bigram = ни Pr = 0.006888522363442321
Bigram = но Pr = 0.006791500921703697
Bigram = ет Pr = 0.006694479479965073
Bigram = ва Pr = 0.006597458038226448
Bigram = ак Pr = 0.0064034151547491995
Bigram = ен Pr = 0.0064034151547491995
Bigram = ор Pr = 0.0064034151547491995
Bigram = во Pr = 0.0063063937130105755
Bigram = ть Pr = 0.006209372271271951
Bigram = ло Pr = 0.0061123508295333265
Bigram = од Pr = 0.006015329387794702
Bigram = ит Pr = 0.005918307946056078
Bigram = та Pr = 0.005821286504317454
Bigram = ла Pr = 0.005627243620840205
Bigram = он Pr = 0.005433200737362957
Bigram = че Pr = 0.005433200737362957
Bigram = за Pr = 0.00504511497040846
Bigram = ос Pr = 0.00504511497040846
Bigram = пр Pr = 0.004851072086931212
```

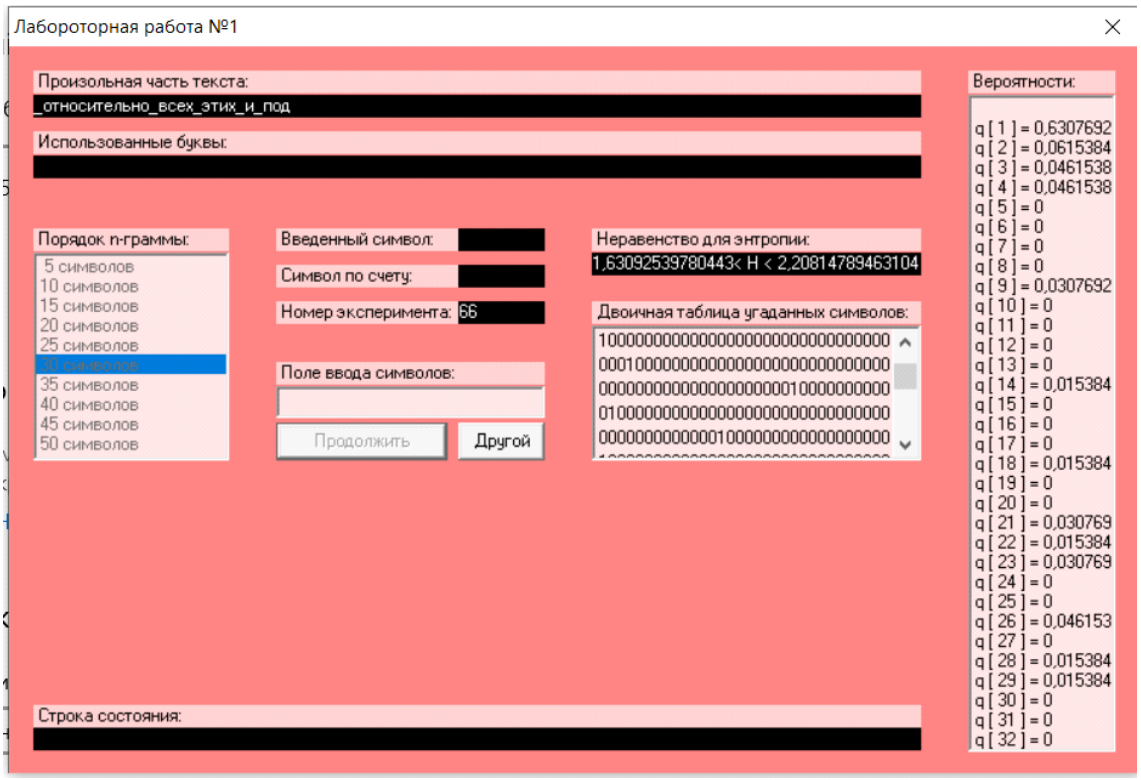
2. Значення ентропій:

```
H1 no spaces = 4.4750526739799
H1 with spaces = 3.941644634319368
H2 no spaces no overlap = 4.110658445344575
H2 no spaces with overlap = 4.112820840022576
H2 with spaces no overlap = 2.8064316035561676
H2 with spaces with overlap = 2.806944674738574
```

3.Експериментальні значення H(10), H(20), H(30):


Произольная часть текста:
тения_или_нет_тогда
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 58
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,19226857717522< H < 1,71638954995895
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,6666666
q [2] = 0,1754385
q [3] = 0,0175438
q [4] = 0
q [5] = 0
q [6] = 0
q [7] = 0
q [8] = 0,0175438
q [9] = 0
q [10] = 0
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0,017543
q [16] = 0
q [17] = 0,017543
q [18] = 0
q [19] = 0,017543
q [20] = 0
q [21] = 0,017543
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0,035087
q [28] = 0
q [29] = 0,017543
q [30] = 0
q [31] = 0
q [32] = 0


Лабороторная работа №1
Произольная часть текста:
ится_тем_
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 61
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
2,69130783862939< H < 3,4656223426392
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,4333333
q [2] = 0,05
q [3] = 0,05
q [4] = 0,0333333
q [5] = 0,0166666
q [6] = 0
q [7] = 0,05
q [8] = 0,0166666
q [9] = 0
q [10] = 0,016666
q [11] = 0
q [12] = 0,016666
q [13] = 0,016666
q [14] = 0,033333
q [15] = 0
q [16] = 0,016666
q [17] = 0,016666
q [18] = 0,016666
q [19] = 0
q [20] = 0,033333
q [21] = 0,033333
q [22] = 0
q [23] = 0,016666
q [24] = 0
q [25] = 0,016666
q [26] = 0,016666
q [27] = 0,033333
q [28] = 0,016666
q [29] = 0,016666
q [30] = 0,016666
q [31] = 0
q [32] = 0,016666



4. Оцінка значення R:

* 0.37 < R < 0.46 при H(10)
* 0.56 < R < 0.67 при H(30)
* 0.65 < R < 0.76 при H(20)

Висновки

Наша пригада реалізувала програмний підрахунок частот літер, біграм та значень ентропій для різних моделей тексту. Також були отримані експериментальні дані обчислення надлишковості тексту російською мовою. Було встановлено, що ентропії текстів з вилученими пробілами можуть достатньо відрізнятись, в той час як ентропія сукупного розподілу біграм на тексті у випадках коли ми рахуємо неперетинні біграми та біграми що перетинаються майже співпадають. З таблиці результатів видно, що найуживанішими літерами російської мови - це "о", "п", "р", "с", а найчастішими біграмами є "не", "то", "ов", "на" , "ст" і так далі.