НТУУ "КПІ ім Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав:

студент групи ФБ-14 Хаща Іван

Київ 2023

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

Порядок виконання роботи

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли

З пробілами

```
" "   -->   0.14628218170216875
"о"   -->   0.1018578247065359
"е"   -->   0.07848984052400255
"т"   -->   0.06627971896850639
"и"   -->   0.061060402740910624
"а"   -->   0.060190050276148
"н"   -->   0.05349729169261166
"с"   -->   0.04978919839418571
"р"   -->   0.043814270702005026
"в"   -->   0.039622026354048746
"л"   -->   0.03165312396365789
"м"   -->   0.028401196104962827
"д"   -->   0.026867584687953455
"к"   -->   0.024908592001992576
"п"   -->   0.02300696981612055
"у"   -->   0.018171056442777125
"я"   -->   0.01663744502576775
"б"   -->   0.014632276083637794
"ы"   -->   0.014174711363416932
"ь"   -->   0.013907448789832268
"ч"   -->   0.012967132300571046
"з"   -->   0.012401822982622336
"г"   -->   0.01219472947010647
"й"   -->   0.009694215234526406
"ж"   -->   0.0075001434262502225
"х"   -->   0.007441373645671396
"ю"   -->   0.005830801801713559
"щ"   -->   0.004662402592586891
"ш"   -->   0.0043363702384234025
"ц"   -->   0.003500999785910085
"ф"   -->   0.0031693703097867075
"э"   -->   0.003057427870588943
```

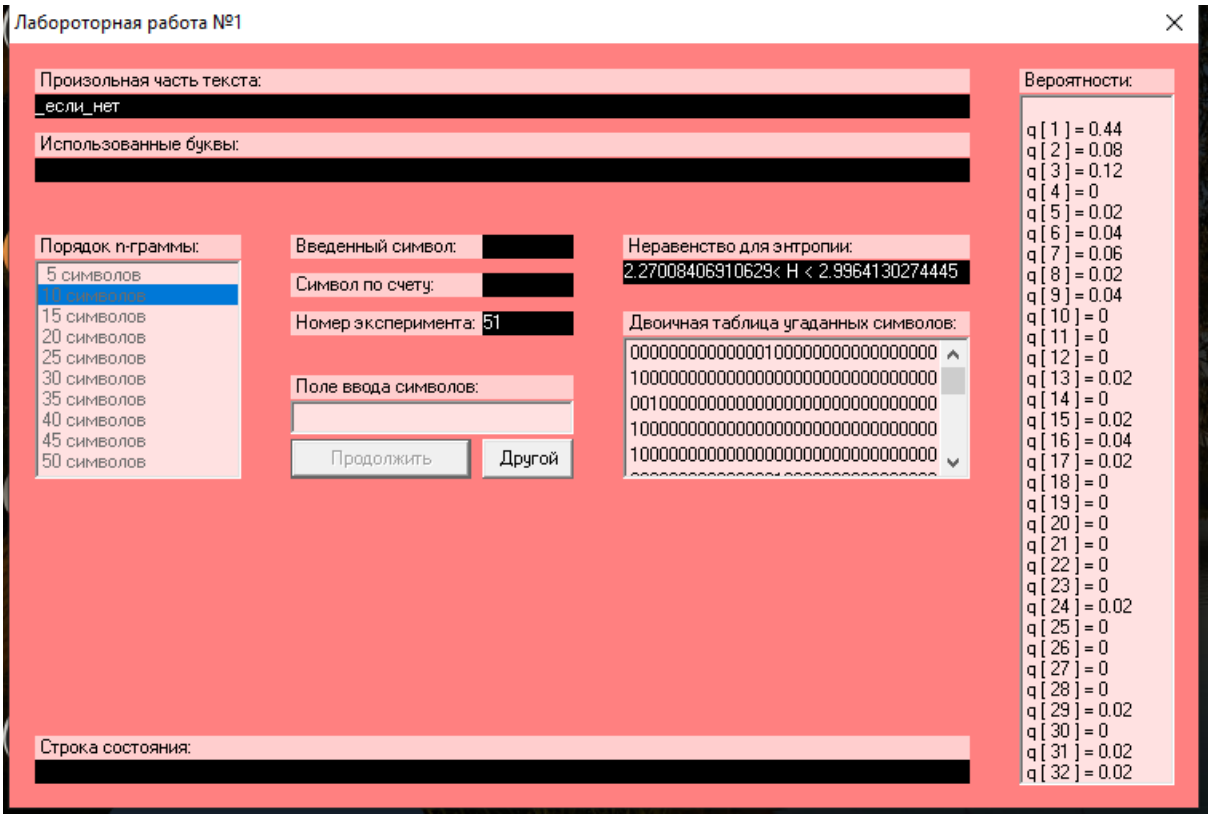
```
Entropy H.1 = 4.371725294760028
Redundancy R.1 = 0.12565494104799446

Entropy H.2 = 3.9306710395230273
Redundancy R.2 = 0.21386579209539458

Entropy H.2 = 3.9307352773977775
Redundancy R.2 = 0.21385294452044445
```

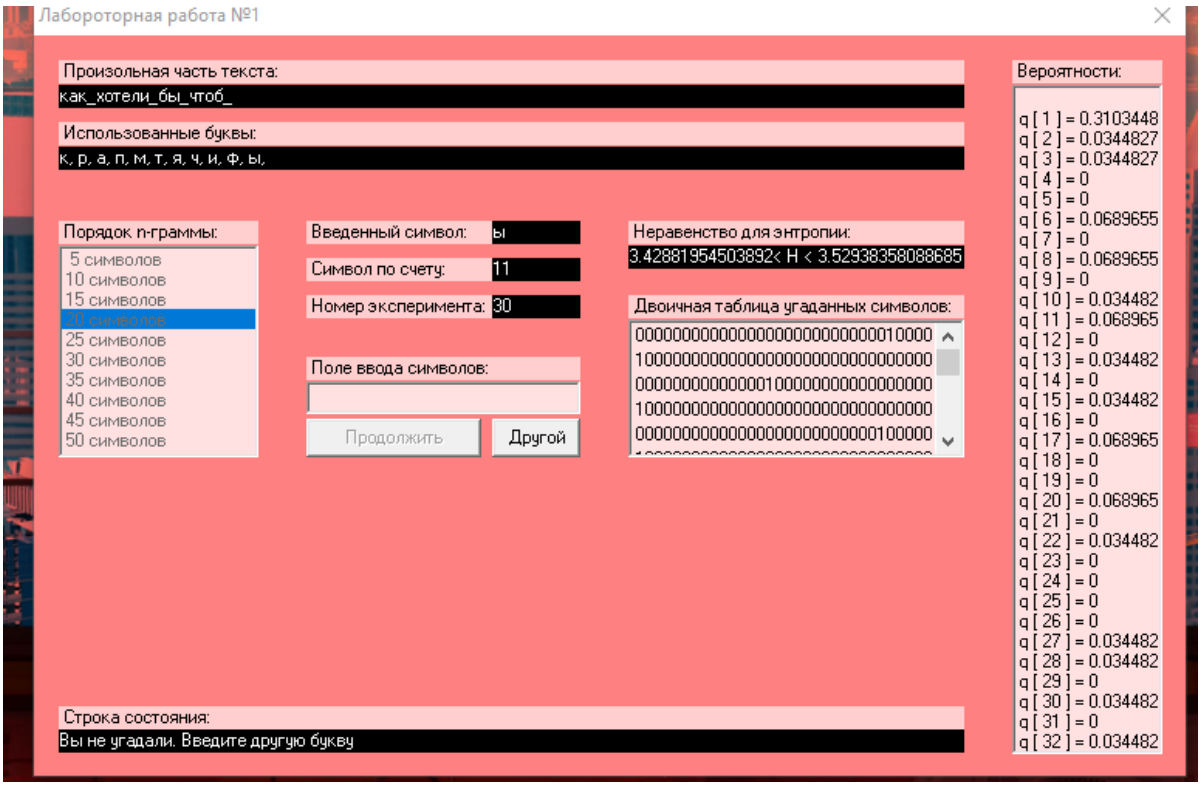
Без пробілів

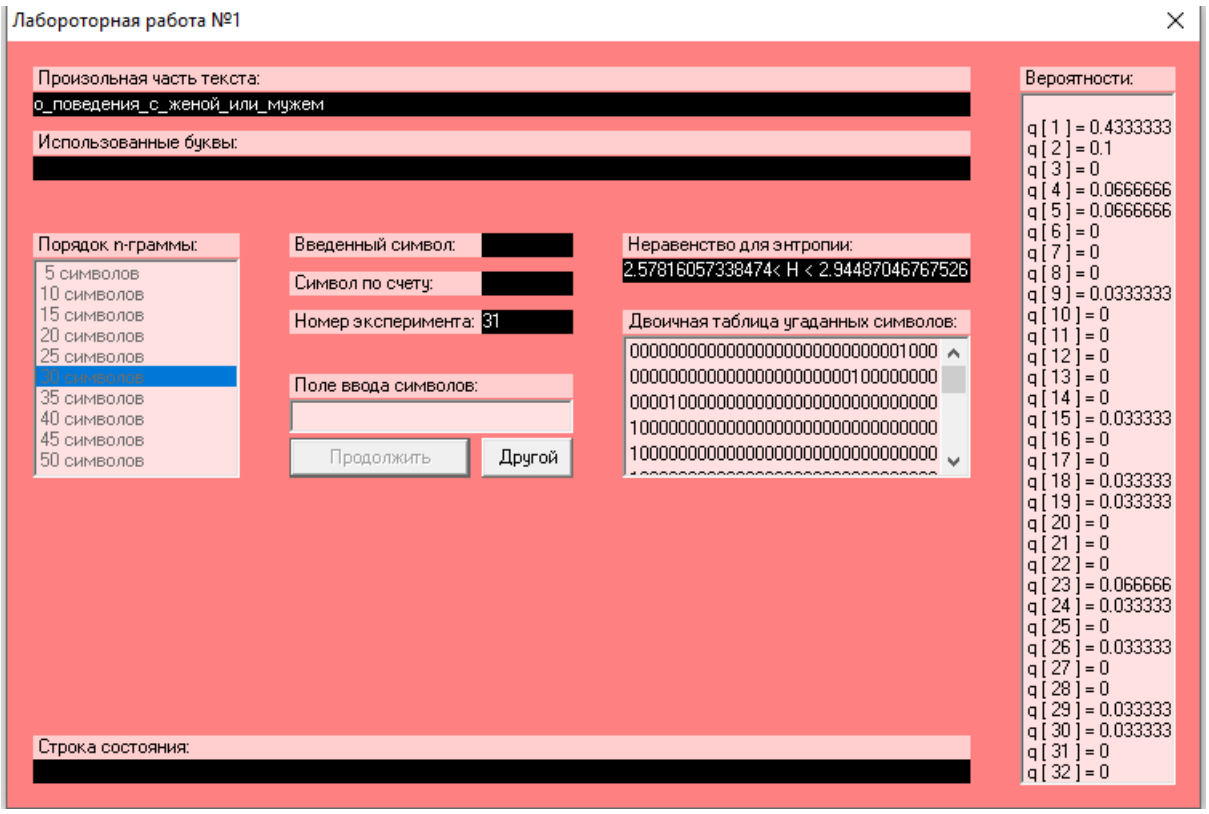
```
"о"    -->   0.1018578247065359
"е"    -->   0.07848984052400255
"т"    -->   0.06627971896850639
"и"    -->   0.061060402740910624
"а"    -->   0.060190050276148
"н"    -->   0.05349729169261166
"с"    -->   0.04978919839418571
"р"    -->   0.043814270702005026
"в"    -->   0.039622026354048746
"л"    -->   0.03165312396365789
"м"    -->   0.028401196104962827
"д"    -->   0.026867584687953455
"к"    -->   0.024908592001992576
"п"    -->   0.02300696981612055
"у"    -->   0.018171056442777125
"я"    -->   0.01663744502576775
"б"    -->   0.014632276083637794
"ы"    -->   0.014174711363416932
"ь"    -->   0.013907448789832268
"ч"    -->   0.012967132300571046
"з"    -->   0.012401822982622336
"г"    -->   0.01219472947010647
"й"    -->   0.009694215234526406
"ж"    -->   0.0075001434262502225
"х"    -->   0.007441373645671396
"ю"    -->   0.005830801801713559
"щ"    -->   0.004662402592586891
"ш"    -->   0.0043363702384234025
"ц"    -->   0.003500999785910085
"ф"    -->   0.0031693703097867075
"э"    -->   0.003057427870588943
```

```
Entropy H.1 = 3.966059345058023
Redundancy R.1 = 0.1994545438696368

Entropy H.2 = 4.059183620300855
Redundancy R.2 = 0.18065749397325115

Entropy H.2 = 4.057968090491926
Redundancy R.2 = 0.18090284755489683
```

За допомогою програми CoolPinkProgram оцінити значення H(10) , H(20) , H(30)







H(10)
2.2708406150629 < H < 2.9964130274445
H(20)
3.42881547503682 < H < 3.5238350888865
H(30)
2.57816507338474 < H < 2.94487046767526

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела

Надлишковість джерела відкритого тексту (мови) дорівнює $R = 1 - \frac{H_\infty}{H_0}$

H0 = log2(32) = 5 (32 букви алфавіту, враховуючи пробіл)

Для H(10)

- $R_{\min} = 1 - \frac{2.2708406150629}{5} = 1 - 0.45416812301258 = 0.54583187698742$
- $R_{\max} = 1 - \frac{2.9964130274445}{5} = 1 - 0.5992826054889 = 0.4007173945111$

Для H(20)

- $R_{\min} = 1 - \frac{3.42881547503682}{5} = 1 - 0.685763095007364 = 0.314236904992636$
- $R_{\max} = 1 - \frac{3.5238350888865}{5} = 1 - 0.7047670177773 = 0.2952329822227$

Для H(30)

- $R_{\min} = 1 - \frac{2.57816507338474}{5} = 1 - 0.515633014676948 = 0.484366985323052$
- $R_{\max} = 1 - \frac{2.94487046767526}{5} = 1 - 0.588974093535052 = 0.411025906464948$

Висновки:

Під час виконання лабораторної роботи ознайомився із наступними термінами:
Ентропія - міра невизначеності, показує кількість необхідної інформації для однозначного опису ансамблю (пара з множини символів та множини ймовірностей).
Вирахувавши частоту появу символів алфавіту можна визначити ентропію тексту, тобто скільки інформації містить текст
Сукупна ентропія говорить нам, скільки інформації містять два ансамблі (із урахуванням взаємних залежностей)
Умовна ентропія говорить нам, скільки інформації залишиться в ансамблі X, якщо поведінку ансамблю Y буде однозначно визначено

Для Аналізу був взят уривок з Капіталу Карла Макса

Було встановлено, що у російській мові (без пробілу) найчастіше трапляються літери “о”, “а”, “е”, “и”, “н”, “т”, “с”, “л”, “в”, “р”. Також було встановлено частоту різних біграм.

Під час виконання лабораторної роботи, працюючи з програмою CoolPinkProgram, я також зіткнулися з поняттям умовної ентропії джерела, яке визначає, скільки інформації про наступний символ ми отримаємо з значень попередніх. Було оцінено умовну ентропію для n=10, 20, 30. Це дозволило розрахувати надлишковість російської мови (величину можливого стиснення тексту деякою схемою кодування символів без втрати його змісту).