Міністерство освіти і науки україни
Національний технічний університет україни
"київський політехнічний інститут імені ігоря сікорського"
Фізико-технічний інститут

Криптографія
Комп’ютерний практикум №1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
Студенти гр. ФБ-11
Поліщук Олександра
Маленко Сергій

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.

Текст міститься у input.txt, звідки він зчитується та аналізується за допомогою класа та виклику методів класа. Ми вивели топ 10 частот для кожного з пунктів.

З пробілами:

Літери	
‘ ‘	0.14518372115092
о	0.09298623359369632
е	0.0726196748026008
и	0.07184087644265852
а	0.06920343325730778
т	0.060541755699462366
н	0.055885326849891524
с	0.047887918481408644
Р	0.040468064358849344
л	0.037208891390351405

```
Entropy for symbols with whitespaces H = 4.347184853762618
Redundancy for symbols with whitespaces R = 0.13056302924747631
```

Біграми(step 1)	
‘о ’	0.02148897982156326
‘ п’	0.018599567733480368
‘ н’	0.017976200799414917
‘и ’	0.016330250312092502
‘ст’	0.015301122224094686
‘ни’	0.014550136862582767
‘ с’	0.013877685917960984
‘е ’	0.013682986009420858
‘то’	0.013370484475545528
‘ в’	0.013205234973339202

```
Entropy for bigrams (step 1) with whitespaces H = 3.8994704009421124
Redundancy for bigrams (step 1) with whitespaces R = 0.22010591981157757
```

Біграми(step 2)	
‘о ’	0.021488944662777038
‘ п’	0.018583175992068036
‘ н’	0.017958173946904278
‘и ’	0.016348221034754695
‘ст’	0.015333819809619795
‘ни’	0.014679367406306957
‘ с’	0.01379912892385119
‘е ’	0.013609337726890467
‘то’	0.013311561883383127
‘о ’	0.013229755332969021

Entropy for bigrams (Step 2) with whitespaces H = 3.8997244913498395

Redundancy for bigrams (Step 2) with whitespaces R = 0.22005510173003207

Без пробілів:

Літери	
о	0.10877920308079822
е	0.08495354686082433
и	0.08404247581642302
а	0.08095708396017318
т	0.0708242896134073
н	0.06537700349499102
с	0.05602129915668508
р	0.047341242042483474
л	0.04352852456255192
в	0.04207387331518847

Entropy for symbols without whitespaces H = 4.386357390550359

Redundancy for symbols without whitespaces R = 0.11461776729476691

Біграми(step 2)	
‘ст’	0.018112356711792843
‘ни’	0.017274016625164368
‘то’	0.016313179356928078
‘пр’	0.013756050690864965
‘ос’	0.013361762887564814
‘ра’	0.012948334899638442
‘но’	0.012548305040950425

‘ко’	0.011679340658919997
‘по’	0.011024746344703242
‘ен’	0.010745299649160416

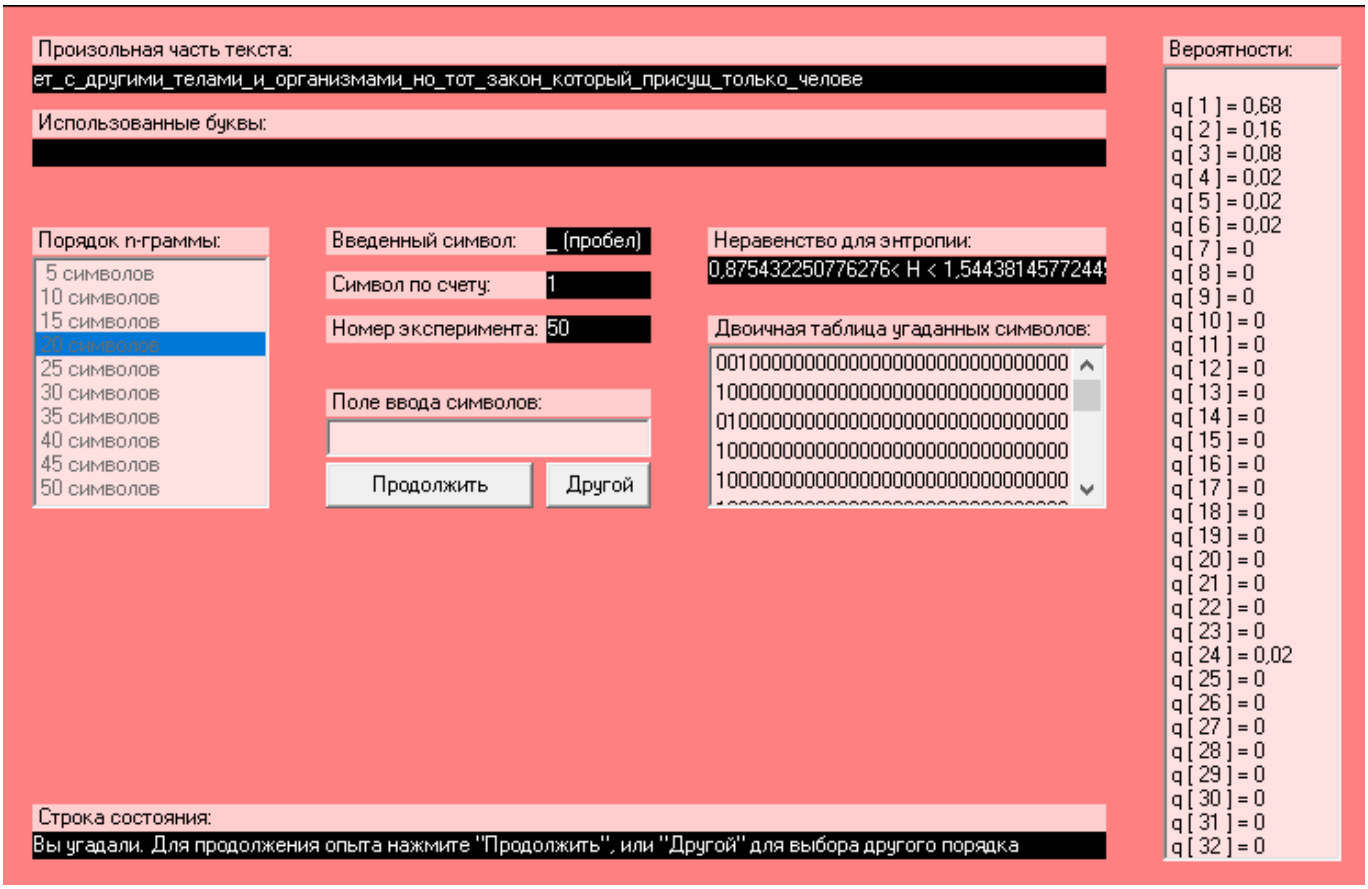
```
Entropy for bigrams (step 1) without whitespaces H = 4.022479821712499
Redundancy for bigrams (step 1) without whitespaces R = 0.18806612219240426
```

Біграми(step 2)	
‘ст’	0.01793431866815194
‘ни’	0.017241445310855145
‘то’	0.016575368160746617
‘пр’	0.013738798228387902
‘ос’	0.013509116452488411
‘ра’	0.012950224131132983
‘но’	0.012647809792865318
‘ко’	0.011824783429225475
‘по’	0.010806527556071064
‘ен’	0.010741451052899541

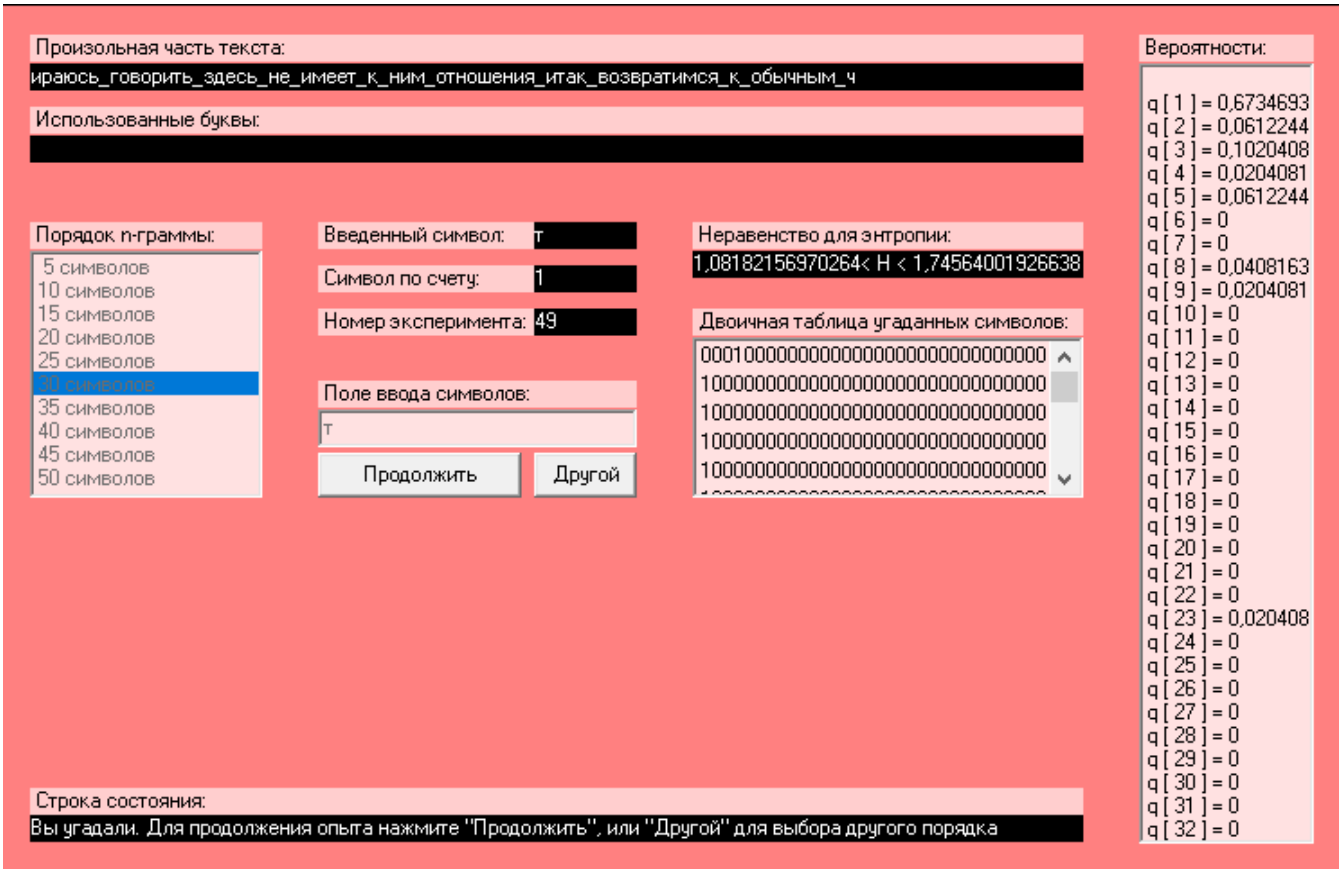
```
Entropy for bigrams (Step 2) without whitespaces H = 4.02217439945209
Redundancy for bigrams (Step 2) without whitespaces R = 0.18812777139668957
```

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

1,63473262855352 < H(10) < 2,35682542858739



$$0{,}875432250776276 < H(20) < 1{,}54438145772445$$



$$1{,}08182156970264 < H(30) < 1{,}74564001926638$$

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

H_0-ентропія відкритих текстів. Оскільки відкритий текст в нас рівно вірогідний звідси: $H_0 = \log_2 n$, де $n - 32$ літери

$$R = 1 - \frac{H_\infty}{H_0} = 1 - \frac{H_\infty}{\log_2 32} = \frac{H_\infty}{5}$$

$$H^{10}: R_l = 1 - \frac{1{,}63473262855352}{5} = 0.673053474289296 \approx 0.67$$

$$R_r = 1 - \frac{2{,}35682542858739}{5} = 0.528634914282522 \approx 0.53$$

$$0.67 < H^{10} < 0.53$$

$$H^{20}: R_l = 1 - \frac{0{,}875432250776276}{5} = 0.8249135498447448 \approx 0.82$$

$$R_r = 1 - \frac{1{,}54438145772445}{5} = 0.69112370845511 \approx 0.69$$

$$0.82 < H^{20} < 0.69$$

$$H^{30}: R_l = 1 - \frac{1{,}08182156970264}{5} = 0.783635686059472 \approx 0.78$$

$$R_r = 1 - \frac{1{,}74564001926638}{5} = 0.650871996146724 \approx 0.65$$

$$0.78 < H^{30} < 0.65$$

Висновки: під час роботи над даним комп'ютерним практикумом ми мали змогу дослідити поняття ентропії, біграм та надлишковості російської мови за допомогою власноруч написаного скрипту та програми CoolPinkProgram, що дало змогу оцінити різницю вихідних даних на практиці у порівнянні з отриманими внаслідок експериментального «вгадування» наступних літер(людського фактору у тому числі)