КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

ФБ-12 Приходько Юрій

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.

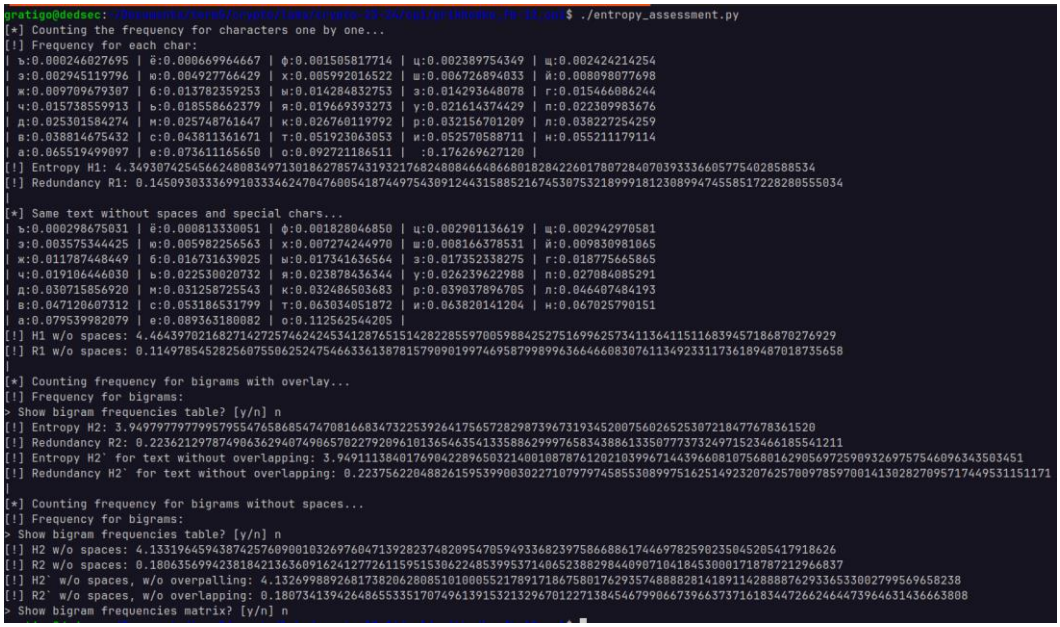
Порядок виконання роботи:

- Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.
- Написати програми для підрахунку частот букв і частот біграм в тексті, а такожпідрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, атакож 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.
- За допомогою програми CoolPinkProgram оцінити значення H (10) , H (20) , H (30) .
- Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

Напишемо програму мовою python що задовольнить поставлену задачу. Програма має підраховувати H1, H2, R1, R2 для текстів з пробілами і без. Для частотного аналізу біграм в тексті візьмемо їх з перетинами і без. Текст довільний, тому перед виконанням треба його спарсити. Крім функцій безпосереднього обрахування заданих значеннь з тексту, програма має мати презентабельний аутпут, напишемо відповідні функції.

Результати виконання програми:



Таким чином

*	H	R
Монограми з пробілом	4.349	0.145
Монограми без пробілу	4.464	0.114
Біграми з перетином з пробілом	3.950	0.223
Біграми з перетином без пробілу	3.949	0.224
Біграми без перетину з пробілом	4.133	0.181
Бігрми без перетину без пробілу	4.132	0.181

Програма створює таблиці частот для монограм, та матрицю для біграм. Для зручності запису частоти біграм в матриці вказані в %

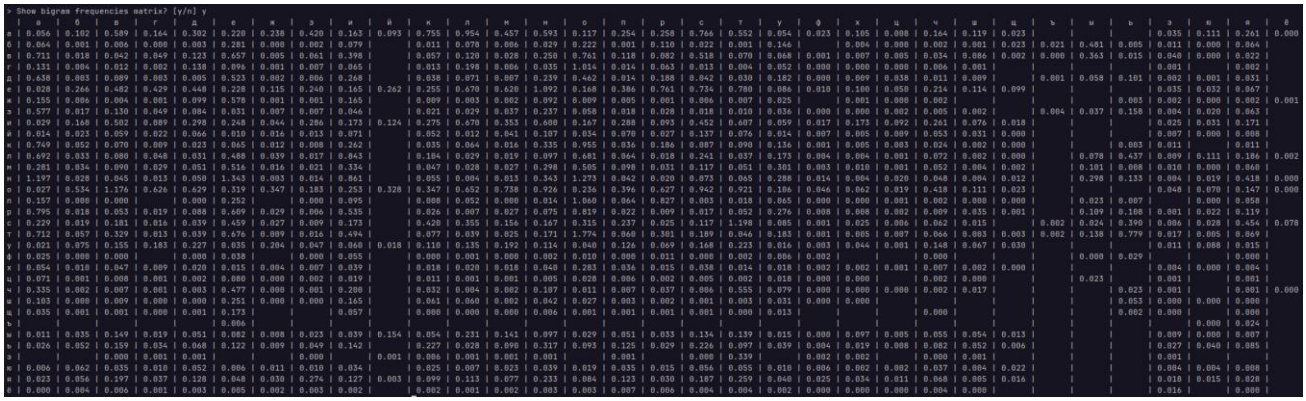
Частота символів з пробілом:

```
[!] Frequency for each char:
| ъ:0.000246027695 | ё:0.000669964667 | ф:0.001505817714 | ц:0.002389754349 | щ:0.002424214254
| э:0.002945119796 | ю:0.004927766429 | х:0.005992016522 | ш:0.006726894033 | й:0.008098077698
| ж:0.009709679307 | б:0.013782359253 | ы:0.014284832753 | з:0.014293648078 | г:0.015466086244
| ч:0.015738559913 | ь:0.018558662379 | я:0.019669393273 | у:0.021614374429 | п:0.022309983676
| д:0.025301584274 | м:0.025748761647 | к:0.026760119792 | р:0.032156701209 | л:0.038227254259
| в:0.038814675432 | с:0.043811361671 | т:0.051923063053 | и:0.052570588711 | н:0.055211179114
| а:0.065519499097 | е:0.073611165650 | о:0.092721186511 |  :0.176269627120 |
```

Частота символів без пробілу:

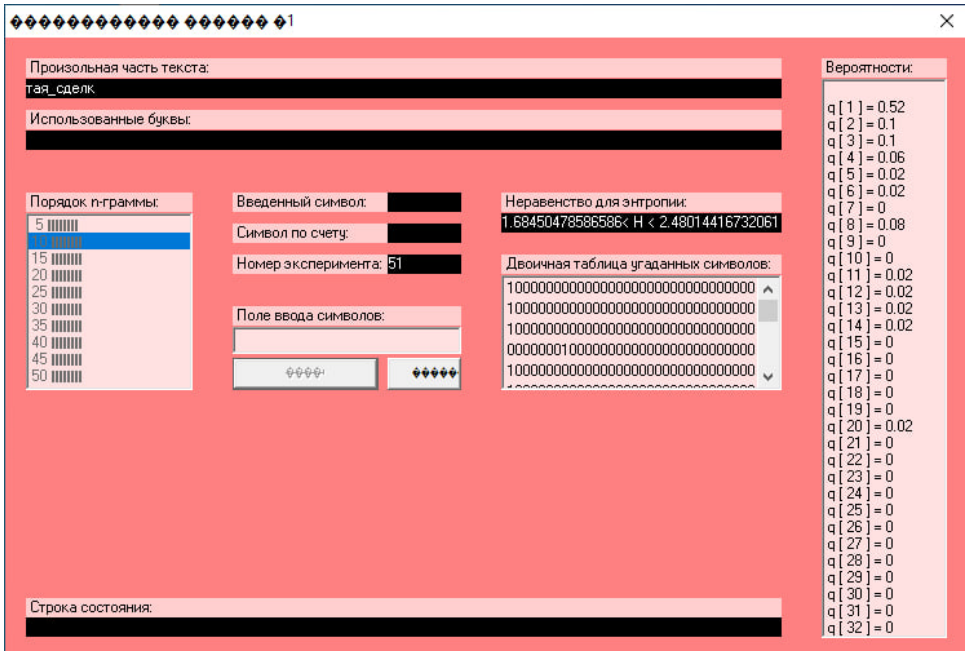
```
[*] Same text without spaces and special chars...
| ъ:0.000298675031 | ё:0.000813330051 | ф:0.001828046850 | ц:0.002901136619 | щ:0.002942970581
| э:0.003575344425 | ю:0.005982256563 | х:0.007274244970 | ш:0.008166378531 | й:0.009830981065
| ж:0.011787448449 | б:0.016731639025 | ы:0.017341636564 | з:0.017352338275 | г:0.018775665865
| ч:0.019106446030 | ь:0.022530020732 | я:0.023878436344 | у:0.026239622988 | п:0.027084085291
| д:0.030715856920 | м:0.031258725543 | к:0.032486503683 | р:0.039037896705 | л:0.046407484193
| в:0.047120607312 | с:0.053186531799 | т:0.063034051872 | и:0.063820141204 | н:0.067025790151
| а:0.079539982079 | е:0.089363180082 | о:0.112562544205 |
```

Матриця частот в % для біграм з перетином:



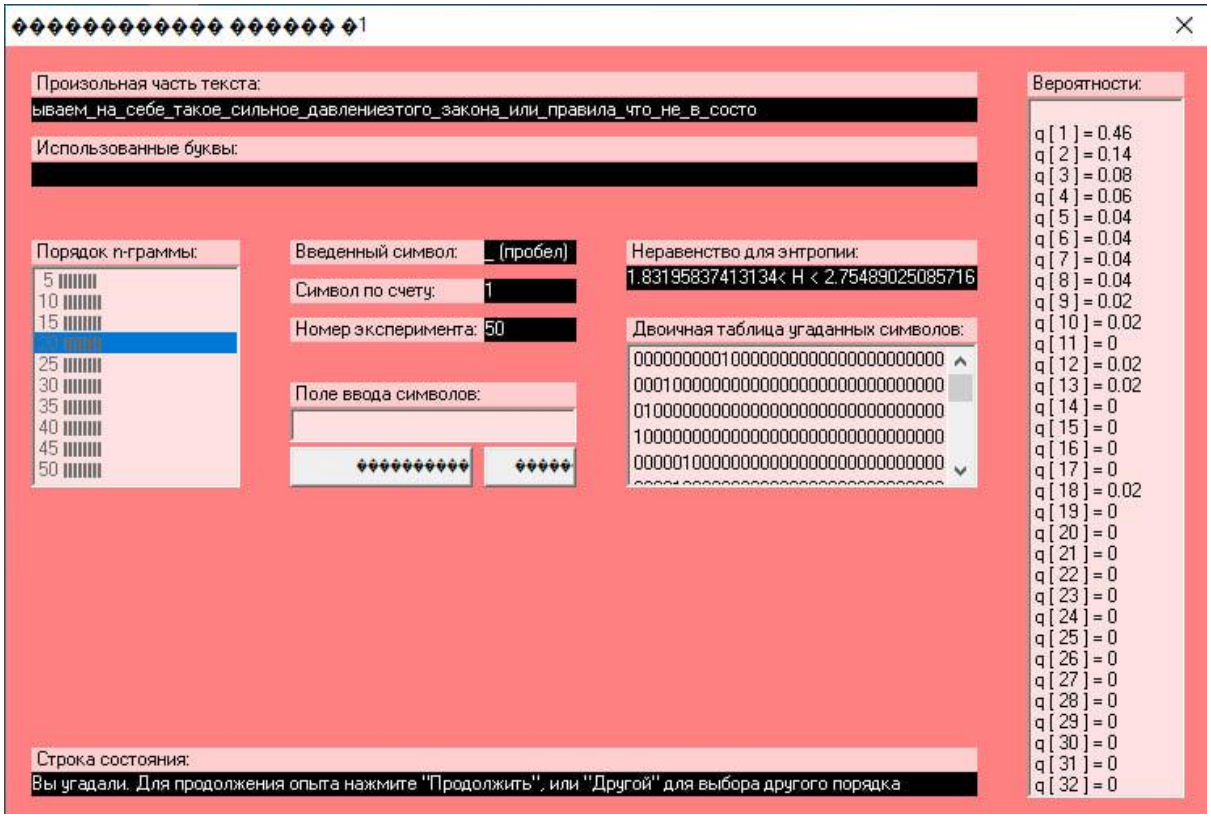
Оцінка значеннь H (10) , H (20) , H (30) за допомогою CoolPinkProgram

H(10)



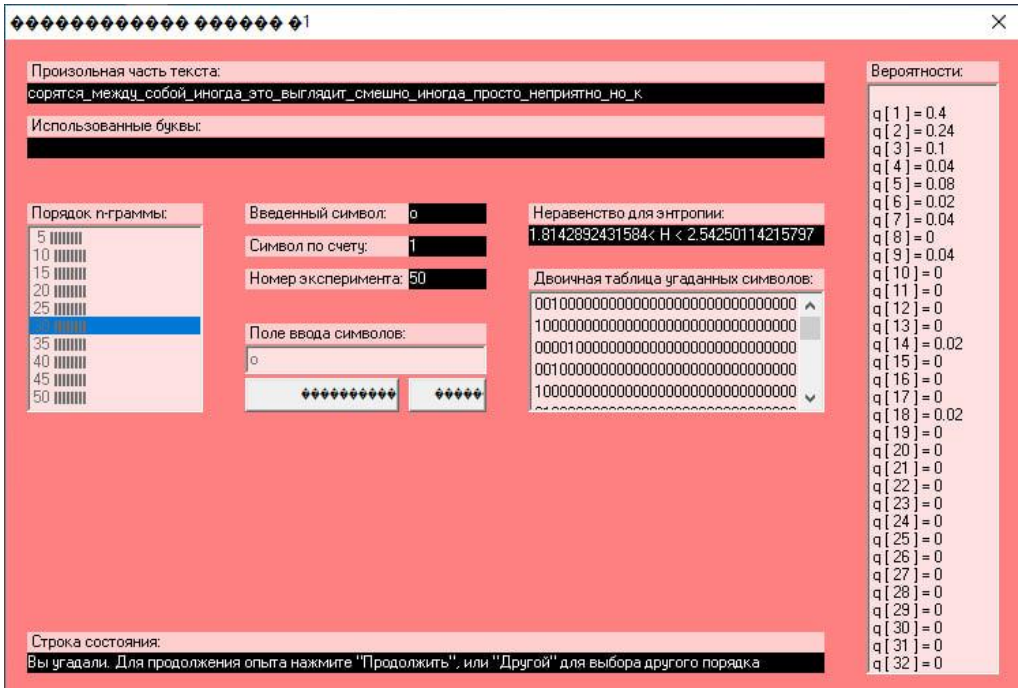
0.5124988142987461 < R < 0.6688909897862052

H (20)



0.4584942756692265 < R < 0.6399072717981559

H (30)



0.500241833830253 < R < 0.6433803928179445

Висновки:

При виконанні комп’ютерного практикуму я отримав навички аналізу частот монограм на біграм. Закріпив на практиці формули обрахування ентропії та надлишковості. Зміг порівняти та оцінити отримані значення для різних наборів вхідних алфавітів, спробував з двума різними за змістом текстами. При роботі з CoolPinkProgram застосував отримані дані для передбачення наступного символу, як найбільш ймовірного, у ситуаціях де знання слів російської мови не допомогло б (наприклад нове слово після пробілу).