КРИПТОГРАФІЯ

**КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1**

Експериментальна оцінка ентропії на символ джерела відкритого тексту

**Виконали:**

Винник Михайло та Кузнєцов Олексій ФБ-12

**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. Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.
2. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H(1) та H(2) за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1() та H(2) на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H(1) та H(2) на тому ж тексті, в якому вилучено всі пробіли.
3. За допомогою програми CoolPinkProgram оцінити значення H(10), H(20), H(30)
4. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела

**Хід роботи:**

1. Спочатку в коді програми ми відфільтрували текст та очистили його від зайвих символів, знаків переносу слова, подвійних пробілів, пробілів взагалі та чисел та інших символів, що не належать російському алфавіту. Таким чином отримуємо другу версію нашого текстового документу. Відповідний крок продемонстровано в коді методом *filter()*.
2. Далі починається виконання основного завдання:

В наступній частині програми ми виконуємо задачу, яка відповідає підрахунку частот букв, а потім і частот біграм в тексті. Щоб це здійснити, ми спочатку рахуємо кількість, яку трапляється кожна літера з алфавіту в нашому тексті. Також рахуємо кількість пробілів, тому що будемо порівнювати два випадки: коли в нашому алфавіті є пробіл, та коли його нема. Для цієї задачі використовуємо словники. Вони дуже зручні для того, щоб зберігати інформацію про елемент, та його кількість.

Після того, як ми підрахували, скільки в тексті трапляється кожна літера, ми можемо порахувати імовірність, або ж частоту: ділимо кількість кожного окремого символу алфавіту на загальну кількість символів в тексті, що відповідають символам алфавіту.

1. Тепер можна легко обчислити Н(1). Для цього скористаємось формулою:

Де n – кількість літер алфавіту, p(i) – імовірність (частота) появи літери в тексті.

Цю формулу ми теж реалізували в своєму коді.

Далі обчислюємо H(2). Для цього треба порахувати частоту біграм. Робимо це так само, як і з частотою символів, але з урахуванням того, що частота біграм – відношення кількості появ деякої біграми до загальної кількості біграм у тексті. Щоб порахувати біграми, ми розробили фрагмент коду, що їх формує з тексту та рахує в словнику.

1. H(2) обчислюємо за такою формулою:

*/2*

де p(i,j) – частота появи деякої біграми в тексті.

Таку формулу ми отримуємо з загальної формули для розрахунку ентропії n-грам:

Де в свою чергу, – ентропія n-грами відкритого тексту .

**Результати виконання програми:**

(на наступній сторінці)

Таблиця частот літер

З пробілом:

|  |  |  |
| --- | --- | --- |
| Буква | Кількість | Частота |
| а | 47917 | 0.06147 |
| б | 10629 | 0.01363 |
| в | 29047 | 0.03726 |
| г | 11416 | 0.01464 |
| д | 19428 | 0.02492 |
| е | 56840 | 0.07291 |
| ё | 1 | 0.0 |
| ж | 6235 | 0.008 |
| з | 12203 | 0.01565 |
| и | 49448 | 0.06343 |
| й | 7911 | 0.01015 |
| к | 20658 | 0.0265 |
| л | 31596 | 0.04053 |
| м | 23876 | 0.03063 |
| н | 45985 | 0.05899 |
| о | 73763 | 0.09462 |
| п | 18248 | 0.02341 |
| р | 29569 | 0.03793 |
| с | 36491 | 0.04681 |
| т | 37317 | 0.04787 |
| у | 17100 | 0.02194 |
| ф | 1448 | 0.00186 |
| х | 7790 | 0.00999 |
| ц | 1981 | 0.00254 |
| ч | 9675 | 0.01241 |
| ш | 5103 | 0.00655 |
| щ | 3072 | 0.00394 |
| ъ | 150 | 0.00019 |
| ы | 14333 | 0.01839 |
| ь | 11523 | 0.01478 |
| э | 2296 | 0.00295 |
| ю | 3876 | 0.00497 |
| я | 14067 | 0.01804 |
| \_ | 118574 | 0.1521 |

Частота "ё": 1.2827650256681283e-06

Таблиця частот літер

Без пробіла:

|  |  |  |
| --- | --- | --- |
| Буква | Кількість | Частота |
| а | 47917 | 0.07249 |
| б | 10629 | 0.01608 |
| в | 29047 | 0.04394 |
| г | 11416 | 0.01727 |
| д | 19428 | 0.02939 |
| е | 56840 | 0.08599 |
| ё | 1 | 0.0 |
| ж | 6235 | 0.00943 |
| з | 12203 | 0.01846 |
| и | 49448 | 0.07481 |
| й | 7911 | 0.01197 |
| к | 20658 | 0.03125 |
| л | 31596 | 0.0478 |
| м | 23876 | 0.03612 |
| н | 45985 | 0.06957 |
| о | 73763 | 0.11159 |
| п | 18248 | 0.02761 |
| р | 29569 | 0.04473 |
| с | 36491 | 0.05521 |
| т | 37317 | 0.05646 |
| у | 17100 | 0.02587 |
| ф | 1448 | 0.00219 |
| х | 7790 | 0.01179 |
| ц | 1981 | 0.003 |
| ч | 9675 | 0.01464 |
| ш | 5103 | 0.00772 |
| щ | 3072 | 0.00465 |
| ъ | 150 | 0.00023 |
| ы | 14333 | 0.02168 |
| ь | 11523 | 0.01743 |
| э | 2296 | 0.00347 |
| ю | 3876 | 0.00586 |
| я | 14067 | 0.02128 |

Частота "ё": 1.5128776142525175e-06

Таблиця частот біграм

З пробілом:

|  |  |  |
| --- | --- | --- |
| Біграма | Кількість | Частота |
| и\_ | 15616 | 0.02003 |
| о\_ | 14985 | 0.01922 |
| е\_ | 13857 | 0.01778 |
| \_п | 12465 | 0.01599 |
| \_с | 11913 | 0.01528 |
| \_в | 11166 | 0.01432 |
| \_н | 11063 | 0.01419 |
| а\_ | 9498 | 0.01218 |
| ст | 9300 | 0.01193 |
| \_и | 8901 | 0.01142 |
| я\_ | 8746 | 0.01122 |
| но | 8703 | 0.01116 |
| то | 8703 | 0.01116 |
| \_о | 8700 | 0.01116 |
| не | 7405 | 0.0095 |
| на | 7334 | 0.00941 |
| по | 7285 | 0.00934 |
| ни | 6967 | 0.00894 |
| ь\_ | 6875 | 0.00882 |
| м\_ | 6600 | 0.00847 |

Таблиця частот біграм

Без пробілу:

|  |  |  |
| --- | --- | --- |
| Біграма | Кількість | Частота |
| ст | 9436 | 0.01428 |
| то | 8912 | 0.01348 |
| но | 8871 | 0.01342 |
| ен | 7677 | 0.01161 |
| не | 7429 | 0.01124 |
| на | 7349 | 0.01112 |
| по | 7287 | 0.01102 |
| ни | 7127 | 0.01078 |
| ос | 6824 | 0.01032 |
| ов | 6672 | 0.01009 |
| ли | 6631 | 0.01003 |
| ра | 6469 | 0.00979 |
| ко | 6323 | 0.00957 |
| ал | 6183 | 0.00935 |
| ро | 5582 | 0.00844 |
| ес | 5487 | 0.0083 |
| го | 5480 | 0.00829 |
| ом | 5474 | 0.00828 |
| ер | 5451 | 0.00825 |
| он | 5316 | 0.00804 |

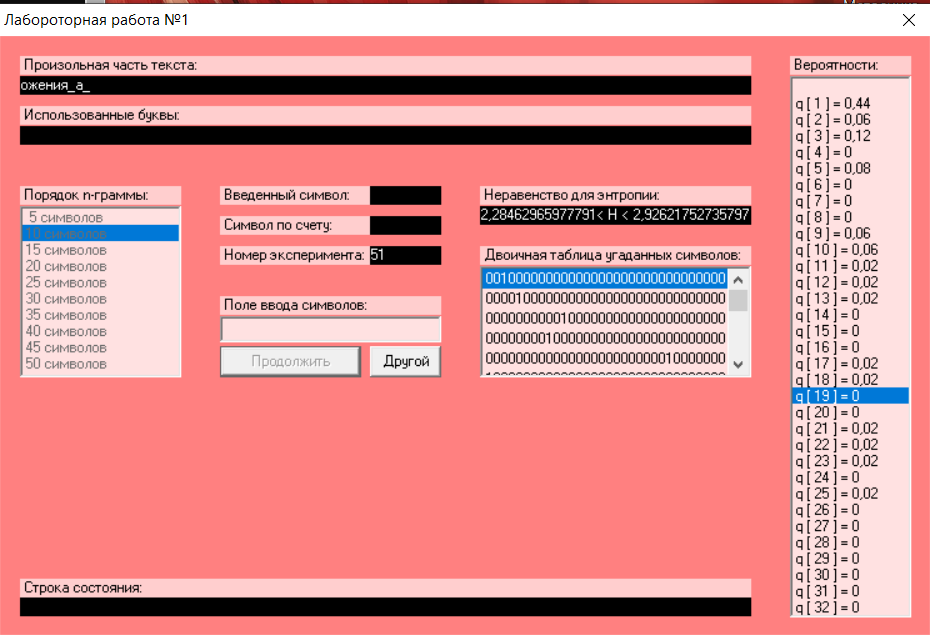
Надлишковість відкритого тексту обчислюється за формулою:

5.08746 для тексту з пробілами, 5.04439 для тексту без пробілів

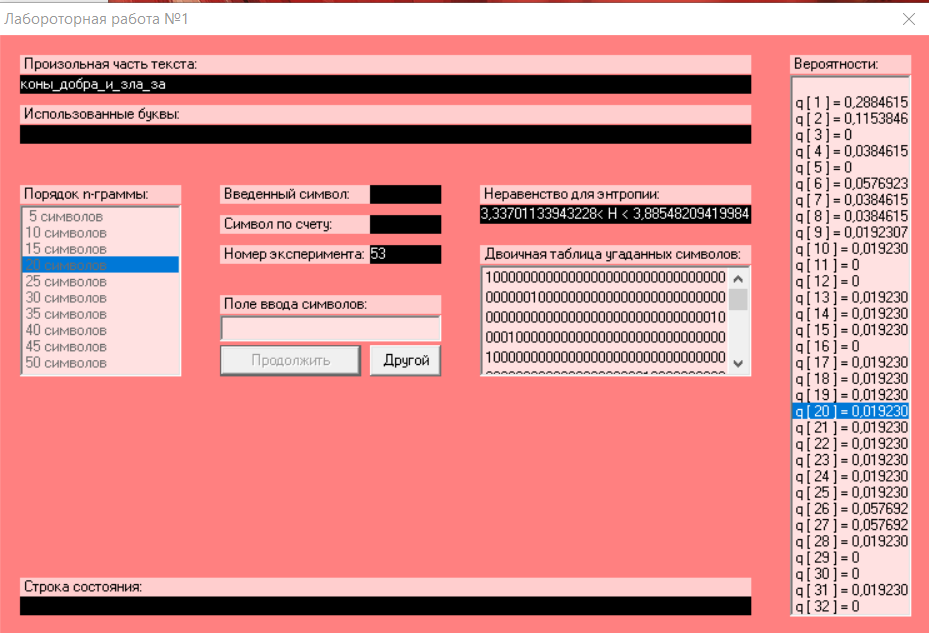
|  |  |  |
| --- | --- | --- |
|  | Ентропія | Надлишковість |
| H1 (з пробілом) | 4,4032187 | 0,134417397 |
| H1 (без пробіла) | 4,4676885 | 0,11425684 |
| H2 (з пробілом) | 4,0015652 | 0,213374248 |
| H2 (без пробіла) | 4,1504491 | 0,177151249 |

Eксперементи в програмі CoolPinkProgram

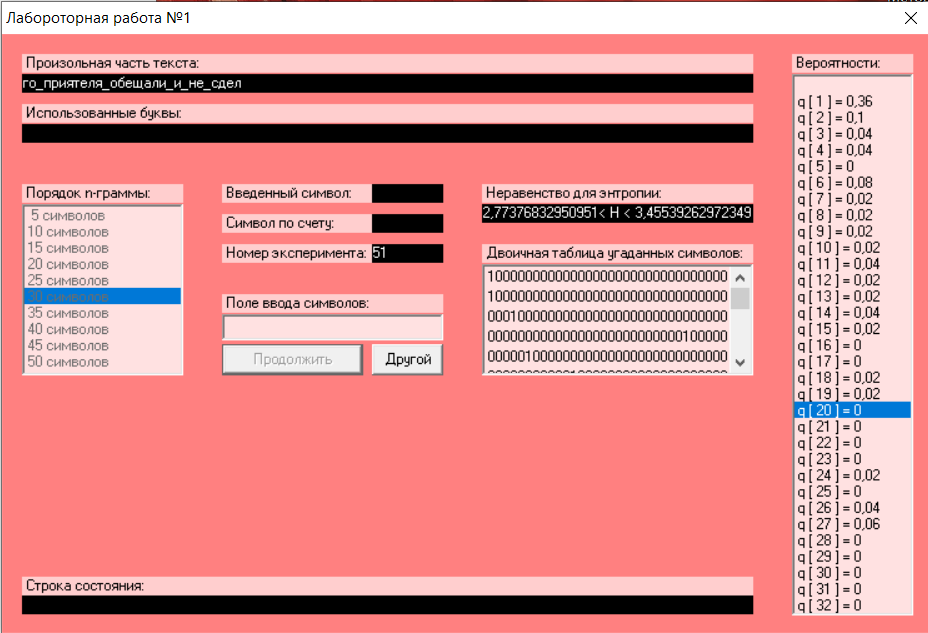
H(10)



H(20)



H(30)



**Висновок:**

При виконанні цього комп'ютерного практикуму ми навчились якісно очищувати текст для подальшої роботи над ним. Також ми навчились визначати експериментальним шляхом значення імовірностей(частот) літер та біграм(а за аналогіює й інших n-грам). Користуючись цими значиннями, ми тепер можемо вираховувати ентропію та надлишковість у різних моделях відкритого тексту.

Також ми навчилися працювати з програмою CoolPinkProgram, за допомогою якої змогли наближено обчислити значення Н(10), Н(20) та Н(30).