Криптографія

Лабораторна робота 1. Експериментальна оцінка ентропії на символ джерела відкритого тексту

ФБ-13 Ігнатенко Данило

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

Задача

0. Прочитати методичку. Двічі
1. Написати програму для розрахунку частот букв та біграм, питомої ентропії n-грам
2. Оцінити значення умовної ентропії за допомогою CoolPinkProgram
3. Оцінити надлишковість російської мови
4. Перечитати методичку ще раз і дописати пропущені функції

Хід роботи

В Інтернеті знайдено текст “Мастер и Маргарита” Булгакова як перший варіант того, що містить російський текст достатньої довжини. Добре, не перший, але варіанти з матами були відкинуті

Текст збережений у файлі raw.txt, відредагована версія у файлі format.txt і версія без пробілів у файлі no_spaces.txt

Першими були написані функції, що приводять текст до вигляду “малі літери без зайвих символів та одинарні пробіли” та “те ж саме без пробілів”. Потім йшли функції для обрахунку частот літер і біграм. Далі йшли функції для запису табличок та матриць частот букв та біграм. Функції обчислення питомої ентропії та надлишковості були написані в кінці, коли я перечитав методичку і зрозумів, як використовувати безпосереднє означення питомої ентропії

Труднощі

Серед дрібного: розібратися з прямим означенням питомої ентропії

Серед реальних труднощів: вивести частоти біграм у матриці

Шляхи розв’язання

Стосовно першого, почитав методу ще пару разів. Стосовно другого, писав, поки не запрацювало. Якщо конкретно, розібрався з csv, як врахувати відсутні біграми (наприклад,

“ыы”), як це все зберігати в змінних. Єдине, з чим не розібрався, – чому, наприклад, Google Docs нормально відображає csv файл, а встановлений Excel – ні. Із припущень тільки те, що файл записувався в кодуванні utf-8, а відкривався на російськомовній системі, де скоріш за все працює cp1251. Так чи інакше, якщо не працюватиме в Excel – заллю на диск і залишу посилання

Результати

Таблиці частот букв та біграм збережені у файлах відповідним чином:

Частоти букв у тексті з пробілами – letter_fq.csv

Частоти букв у тексті без пробілів – letter_fq_no_spaces.csv

Частоти біграм у тексті з пробілами та кроком 1 – bigram_fq_ovp.csv

Частоти біграм у тексті без пробілів та кроком 1 – bigram_fq_ovp_no_spaces.csv

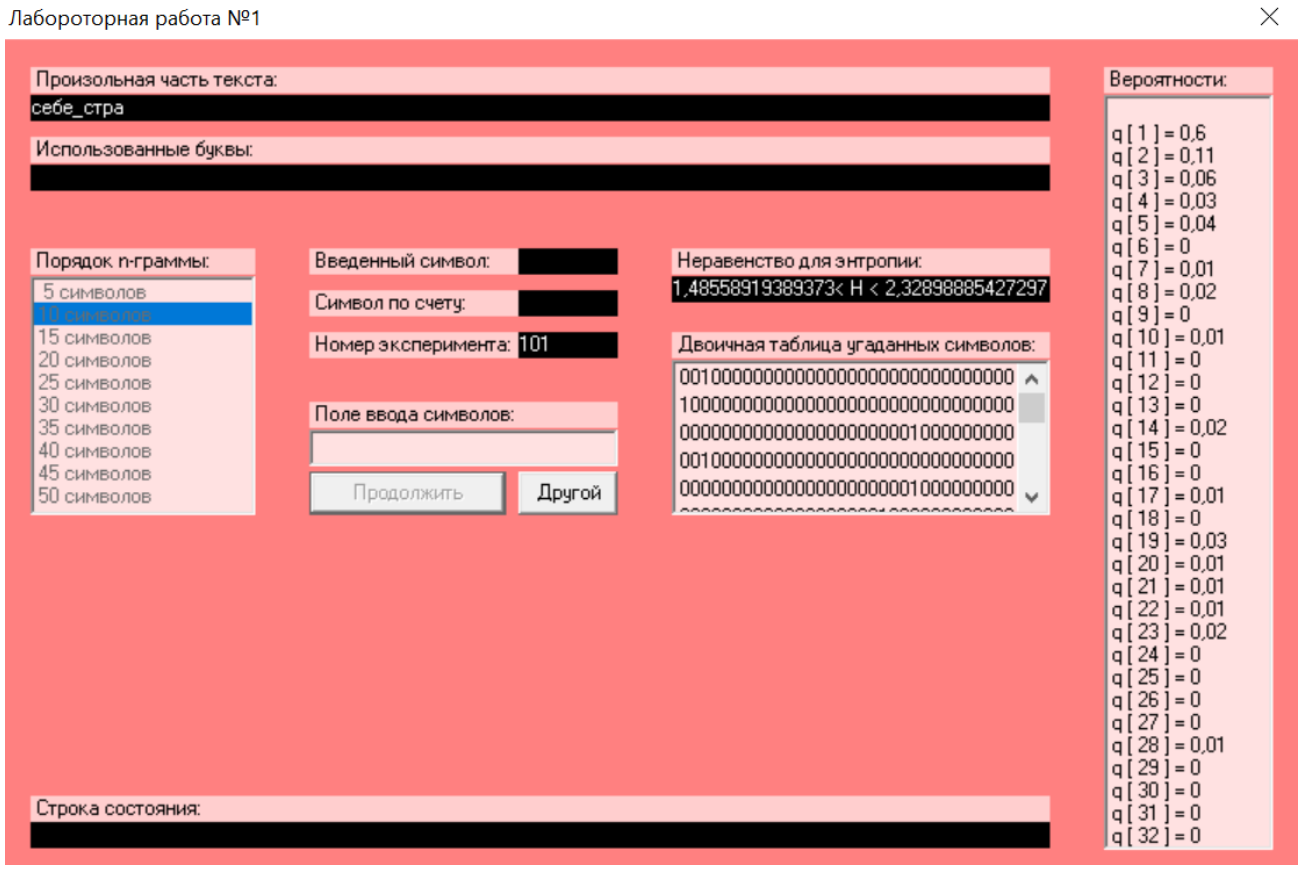
Частоти біграм у тексті з пробілами та кроком 2 – bigram_fq_no_ovp.csv

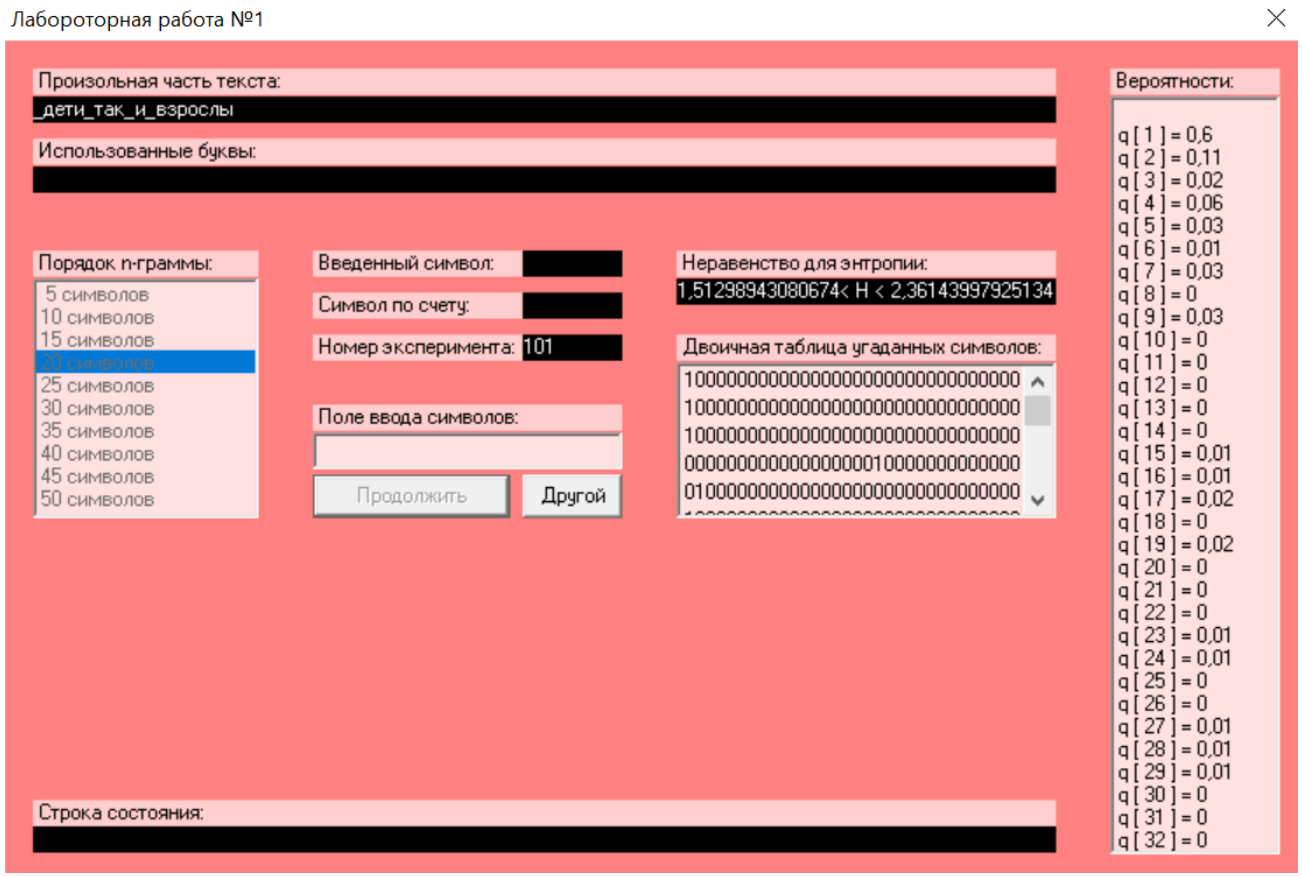
Частоти біграм у тексті без пробілів та кроком 2 – bigram_fq_ovp_no_ovp_no_spaces.csv

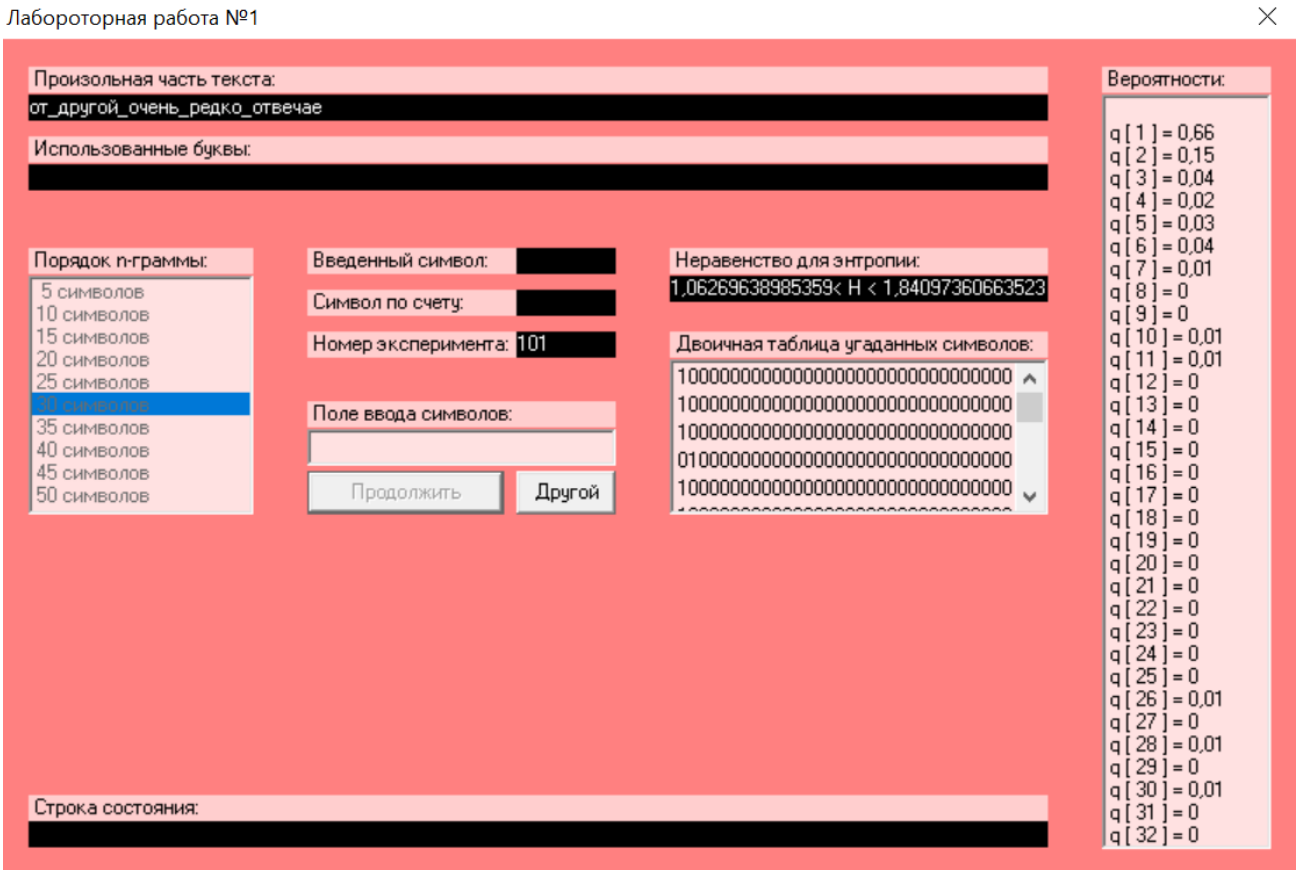
Значення ентропії та надлишковості вийшли наступні

```
(base) C:\Users\uranus\Desktop\Crypt\lab1>python qwe.py
Питома ентропія H1 тексту з пробілами:  4.37602610215994
Питома ентропія H1 тексту без пробілів:  4.450777424510994
Питома ентропія H2 тексту з пробілами та кроком 1:  3.9864385746232043
Питома ентропія H2 тексту без пробілів та кроком 1:  4.146723821647998
Питома ентропія H2 тексту з пробілами та кроком 2:  3.9861992406835998
Питома ентропія H2 тексту без пробілів та кроком 2:  4.146552910656274
Надлишковість при H(10) = 1.4856:  0.70288
Надлишковість при H(10) = 2.3290:  0.5342
Надлишковість при H(20) = 1.5130:  0.6974
Надлишковість при H(20) = 2.3614:  0.52772
Надлишковість при H(30) = 1.0627:  0.78746
Надлишковість при H(30) = 1.8410:  0.6318
```

Значення для оцінки надлишковості були отримані за допомогою CoolPinkProgram. Для оцінки надлишковості доцільним буде взяти значення $H^{(30)}$, тоді вона коливається від ~63% до ~79%. У середньому це 71%


Лабороторная работа №1
Произольная часть текста:
себе_стра
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: 101
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,48558919389373< H < 2,32898885427297
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0,6
q [2] = 0,11
q [3] = 0,06
q [4] = 0,03
q [5] = 0,04
q [6] = 0
q [7] = 0,01
q [8] = 0,02
q [9] = 0
q [10] = 0,01
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0,02
q [15] = 0
q [16] = 0
q [17] = 0,01
q [18] = 0
q [19] = 0,03
q [20] = 0,01
q [21] = 0,01
q [22] = 0,01
q [23] = 0,02
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0,01
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0
Строка состояния:


Лабороторная работа №1
Произольная часть текста:
_дети_так_и_взрослы
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: 101
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,51298943080674< H < 2,36143997925134
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0,6
q [2] = 0,11
q [3] = 0,02
q [4] = 0,06
q [5] = 0,03
q [6] = 0,01
q [7] = 0,03
q [8] = 0
q [9] = 0,03
q [10] = 0
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0,01
q [16] = 0,01
q [17] = 0,02
q [18] = 0
q [19] = 0,02
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0,01
q [24] = 0,01
q [25] = 0
q [26] = 0
q [27] = 0,01
q [28] = 0,01
q [29] = 0,01
q [30] = 0
q [31] = 0
q [32] = 0
Строка состояния:



Висновки

Були отримані значення ентропії на російському тексті з пробілами та без. Закономірним є більше значення ентропії для тексту без пробілів, з чого робимо припущення, що подібний текст складніше піддається аналізу

Також було оцінено значення надлишковості російської мови, а саме 71%