НТУУ «Київський політехнічний інститут ім. Ігоря Сікорського»

Навчально-науковий Фізико-технічний інститут

Криптографія

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Варіант №6

Виконали:

Студенти 3 курсу НН ФТІ

групи ФБ-31

Гаврилюк Володимир

Гек Роман

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.

2. За допомогою програми CoolPinkProgram оцінити значення H10, H20, H30.

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

Виконання роботи

Було проаналізовано текст розміром >1MB, текст був попередньо відфільтрований як і було вказано у методичних вказівках

Був написаний скрипт мовою Python, який обчислює:

- частоти появи кожного символу в тексті;

- ентропію H1 на основі частот символів;

- частоти появи біграм двома методами: з перетином (крок 1 символ) та - без перетину (крок 2 символи)

- ентропію H2 для обох методів підрахунку біграм

- усі розрахунки були проведені для двох варіантів тексту: з пробілами та з видаленими пробілами

Результати програмного аналізу:
Таблиця частот символів (з пробілами)

```
Таблиця частот символів:
Символ     Кількість    Ймовірність
----------------------------------------
_          122779       0.145642
о          73463        0.087143
е          60012        0.071187
а          59924        0.071083
н          47699        0.056581
и          46785        0.055497
т          43940        0.052122
с          36680        0.043510
в          32767        0.038869
р          32606        0.038678
л          32570        0.038635
д          24672        0.029266
к          23715        0.028131
м          23344        0.027691
у          22256        0.026400
п          19054        0.022602
ь          16090        0.019086
я          15453        0.018331
ы          13743        0.016302
б          13402        0.015898
г          13320        0.015800
з          12345        0.014644
й          10931        0.012966
ч          10770        0.012775
ж          7971         0.009455
х          7116         0.008441
ш          6673         0.007916
ю          5568         0.006605
ц          3532         0.004190
щ          1748         0.002073
ф          1109         0.001316
э          983          0.001166
```

Таблиця частот символів (без пробілів)

```
Таблиця частот символів:
Символ     Кількість    Ймовірність
----------------------------------------
о          73463        0.101998
е          60012        0.083322
а          59924        0.083200
н          47699        0.066226
и          46785        0.064957
т          43940        0.061007
с          36680        0.050927
в          32767        0.045494
р          32606        0.045271
л          32570        0.045221
д          24672        0.034255
к          23715        0.032926
м          23344        0.032411
у          22256        0.030901
п          19054        0.026455
ь          16090        0.022340
я          15453        0.021455
ы          13743        0.019081
б          13402        0.018608
г          13320        0.018494
з          12345        0.017140
й          10931        0.015177
ч          10770        0.014953
ж          7971         0.011067
х          7116         0.009880
ш          6673         0.009265
ю          5568         0.007731
ц          3532         0.004904
щ          1748         0.002427
ф          1109         0.001540
э          983          0.001365
```

Підсумкові значення ентропій:

```
З ПРОБІЛАМИ:
  H1 = 4.4356 біт
  H2 (перетин) = 4.0686 біт
  H2 (без перетину) = 4.0690 біт

БЕЗ ПРОБІЛІВ:
  H1 = 4.4908 біт
  H2 (перетин) = 4.1854 біт
  H2 (без перетину) = 4.1841 біт
```

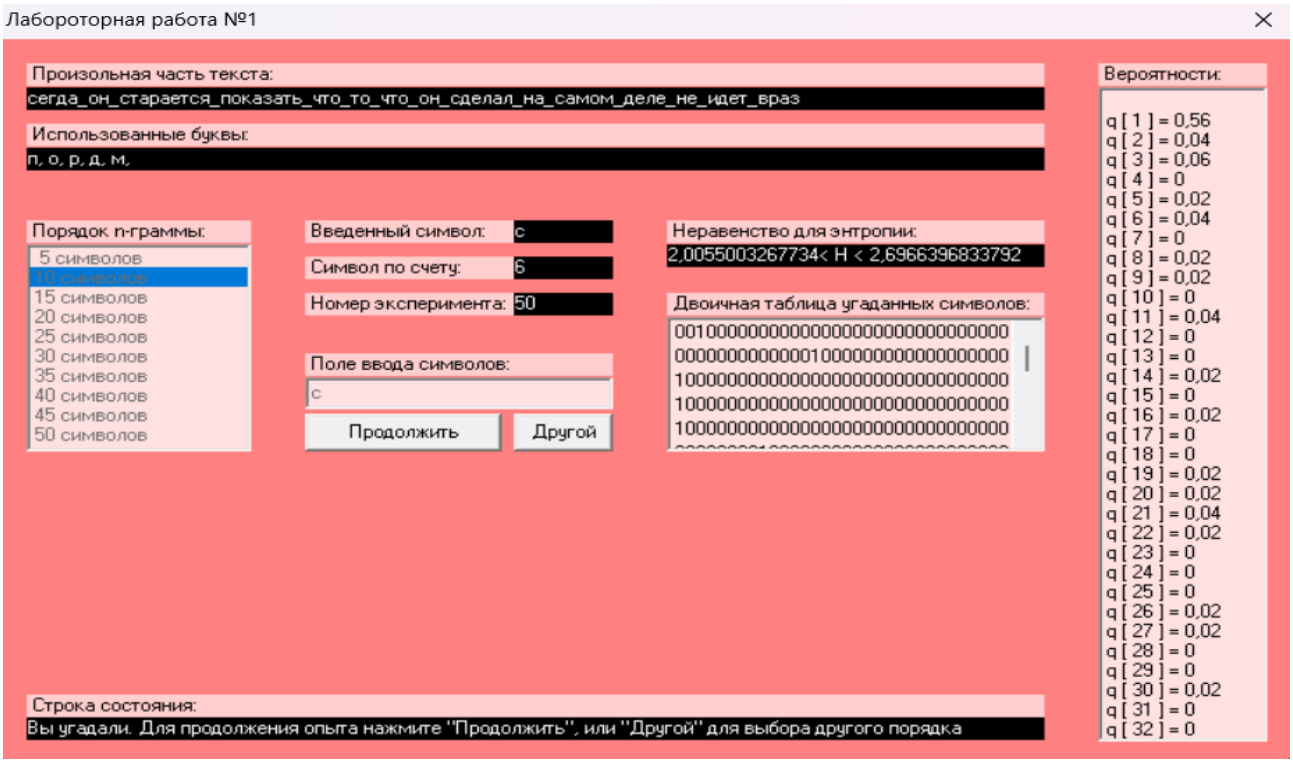
Надлишковість:

```
Модель H1:
  R (з пробілами) = 0.1207 (12.07%)
  R (без пробілів) = 0.1018 (10.18%)

Модель H2 (з перетином):
  R (з пробілами) = 0.1934 (19.34%)
  R (без пробілів) = 0.1629 (16.29%)

Модель H2 (без перетину):
  R (з пробілами) = 0.1934 (19.34%)
  R (без пробілів) = 0.1632 (16.32%)
```

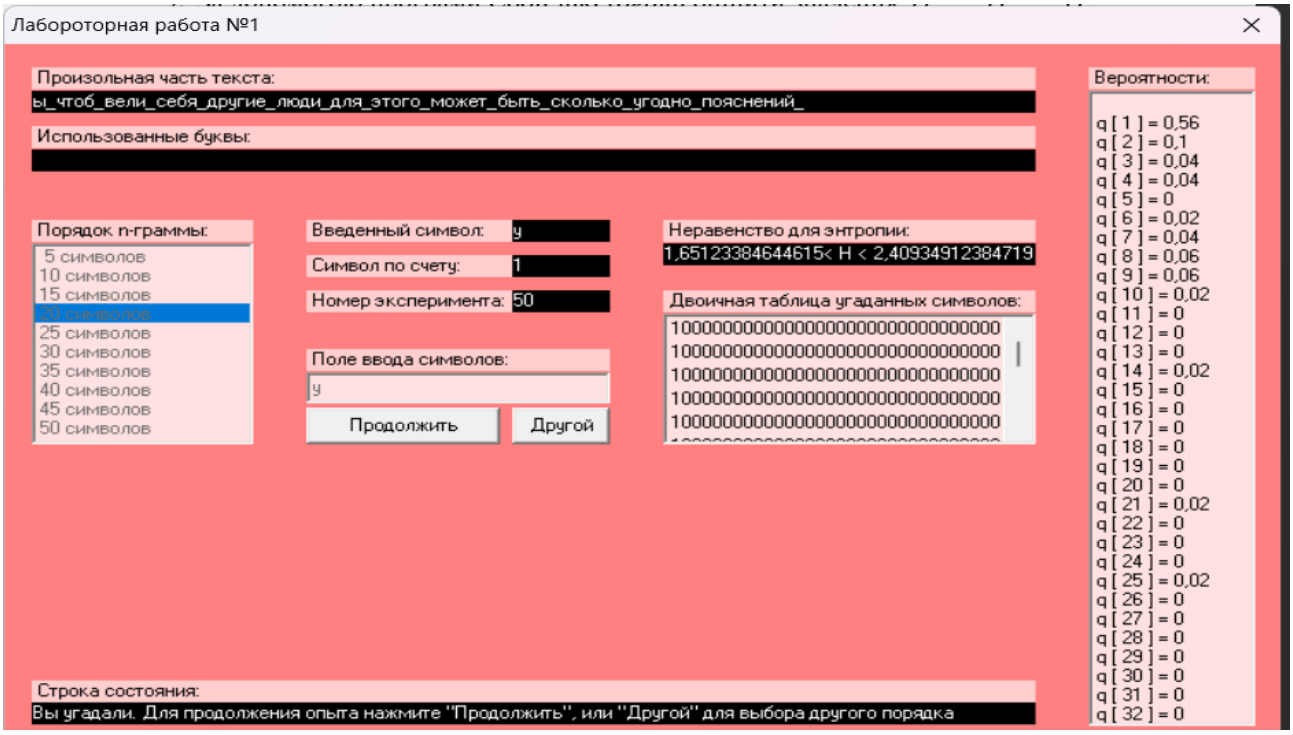
Далі було проведено експерименти з CoolPinkProgram.exe, для кожної n-грами по 50 експериментів. Під час експериментів навмисно підроблювати результати не намагалися, як виходило так і вгадували символи.

Експерименти із CoolPinkProgram

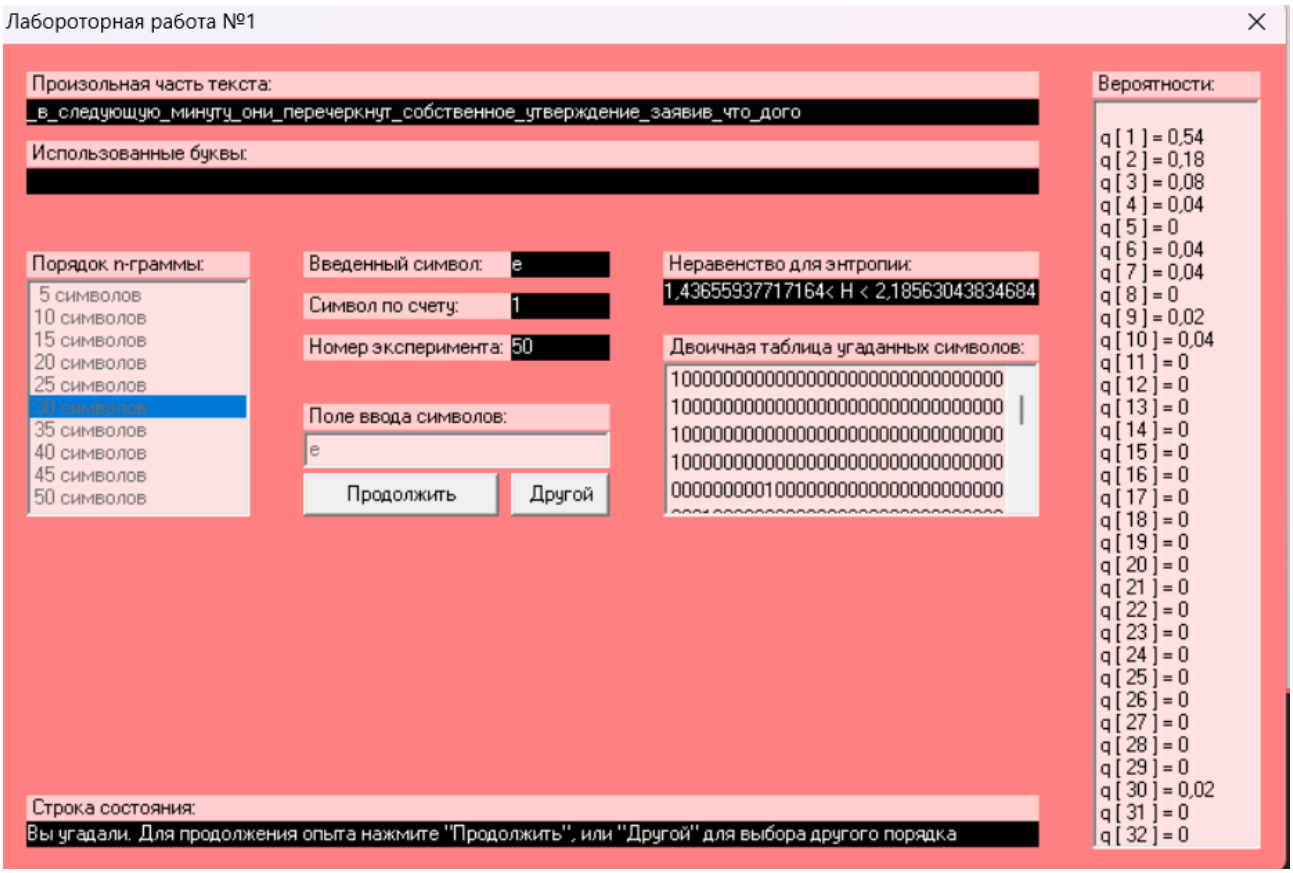
H10



H20



H30



$$H_0 = \log_2 32 = 5$$

$$R = 1 - \frac{H}{H_0}$$

$$0{,}4606606332416 < R^{(10)} < 0{,}59889993464532$$

$$0{,}518130175230562 < R^{(20)} < 0{,}66975323071077$$

$$0{,}562873912330632 < R^{(30)} < 0{,}712688124565672$$

Щодо того, наскільки точні результати експериментів

Було знайдено статті https://habr.com/ru/articles/791130/, https://habr.com/ru/articles/50643/ та дослідження https://cyberleninka.ru/article/n/analiz-izbytochnosti-russkoyazychnogo-teksta/viewer

С помощью теории информации можно без потери смысла сократить количество букв в словах. Тогда мозг, основываясь на лимитированном количестве информации «урезанного» текста, будет сам достраивать отсутствующие элементы для восстановления смысла. Наглядный пример – всем известная реклама курсов стенографии «если в мжт прчть здс сбщн». Выходит, что **все наши языки в той или иной степени избыточны**, т. е. содержат звуки или буквы, которые не являются строго необходимыми для передачи сообщения. Клод Шеннон оценил избыточность английского языка в 75%. Согласно учебнику «Теория информации» Ерохина, русский язык избыточен на 73%, французский – на 71%, немецкий – на 66%. Внутри языка наименьшей избыточностью обладает живая речь, чуть большей – литературный текст, а наиболее избыточными (80-95%) являются юридические документы и язык диспетчеров аэропорта. Без избыточности мы не смогли бы разгадывать кроссворды, играть в слова и писать стихи. Если **энтропия** языка характеризует **неопределённость и непредсказуемость** появления следующей буквы или слова, то **избыточность** – наоборот, **мера предопределённости, предсказуемое отклонение от случайного**.

	Русский язык	Французский язык
Язык в целом	72,6%	70,6%
Разговорная речь	72,0%	68,4%
Литературный текст	76,2%	71,0%
Деловой текст	83,4%	74,4%

Російська мова має 72.6 відсотків надлишковості, досить великий результат, але ще більше у англійської мови

В загальному надлишковість не є чимось поганим, наприклад як було сказано в одній із статтей – надлишковість допомагає диспетчерам літаків передавати усю потрібну інформацію попри загрозу втрати деякої її частини.

Також щодо ентропій і частот символів

№	Символ	$p(i)$
1		0,1612
2	о	0,0955
3	а	0,0707
4	е	0,0657
5	и	0,0557
6	н	0,0549
7	т	0,0473
8	с	0,0446
9	л	0,0423
10	в	0,0385
11	р	0,0380
12	к	0,0305
13	д	0,0254
14	м	0,0246
15	у	0,0240
16	п	0,0215
17	я	0,0193
18	г	0,0174
19	ь	0,0163
20	ы	0,0158
21	з	0,0149
22	б	0,0144
23	ч	0,0114
24	й	0,0096
25	ж	0,0085
26	ш	0,0079
27	х	0,0071
28	ю	0,0054
29	ц	0,0034
30	э	0,0025
31	щ	0,0023
32	ф	0,0019
33	ъ	0,0004
H_1		4,3832

Язык текста	Значение энтропии, бит				
	H_0	H_1	H_2	H_3	H
Русский	5,044	4,38	3,56	2,89	1,79(2,14)

Хотілось би прокоментувати, що наші результати більш менш збігаються із результатами іншого дослідження, і це також показує приблизну однозначність ентропії і надлишковості для різних текстів.

Підсумок та висновки

В ході виконання роботи було експериментально досліджено статистичні властивості російської мови. Також ми провели оцінку ентропії та надлишковості за допомогою різних моделей. Отримані результати дозволяють зробити декілька висновків

1. Підтверджено теоретичну залежність ентропій від моделі джерела. Чітко видно ієрархію H0 > H1 > H2 > Hn. Врахування зв'язків між сусідніми символами (модель біграм, або n-грам) значно зменшує невизначеність у порівнянні з моделлю, що враховує тільки частоти окремо одного символу.

2. Доведено вплив довжини контексту на точність оцінки ентропій. Оцінка умовної ентропії в експериментах показала, що зі збільшенням кількості символів (n) ентропія значно зменшується. Так середнє значення H10 складає близько ~2.35 біт, тоді як H30 складає ~1.81 біт. Це свідчить, що російська мова має залежності, які дозволяють людині ефективно передбачити наступні символи.

3. Встановлено значну надлишковість російської мови. Найбільш точні оцінки, які були отримані, показують, що надлишковість російської мови може сягати 60-70%. Це означає, що більше половини символів у тексті є передбачуваними і визначаються структурними закономірностями мови. Простіші статистичні моделі (H1, H2) недооцінюють реальну надлишковість, показуючи значення в діапазоні 10-20%.

4. Також проаналізовано вплив алфавіту на ентропію. Видалення пробілів з тексту призвело до незначного зростання ентропії (H1 з та без – 4.43 біт, 4.49 біт). Це пояснюється тим, що пробіл є дуже частим символом, структурно передбачуваним. Його видалення усуває частину передбачуваності з тексту, роблячи його в середньому трішки більш хаотичним для аналізу.