# 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 Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 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 >1МВ, 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 Н1 на основі частот символів;
- частоти появи біграм двома методами: з перетином (крок 1 символ) та без перетину (крок 2 символи)
- ентропію Н2 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 символ:	ів:
Символ	Кількість	Ймовірність
_	122779	0.145642
0	73463	0.087143
e	60012	0.071187
a	59924	0.071083
Н	47699	0.056581
И	46785	0.055497
T	43940	0.052122
С	36680	0.043510
В	32767	0.038869
p	32606	0.038678
Л	32570	0.038635
Д	24672	0.029266
K	23715	0.028131
M	23344	0.027691
y	22256	0.026400
п	19054	0.022602
ь	16090	0.019086
Я	15453	0.018331
Ы	13743	0.016302
6	13402	0.015898
Γ	13320	0.015800
3	12345	0.014644
й	10931	0.012966
Ч	10770	0.012775
ж	7971	0.009455
X	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	Таблиця частот символів:		
Символ			
0	73463	0.101998	
e	60012	0.083322	
a	59924	0.083200	
Н	47699	0.066226	
И	46785	0.064957	
Т	43940	0.061007	
C	36680	0.050927	
В	32767	0.045494	
p	32606	0.045271	
Л	32570	0.045221	
Д	24672	0.034255	
K	23715	0.032926	
M	23344	0.032411	
y	22256	0.030901	
П	19054	0.026455	
Ь	16090	0.022340	
Я	15453	0.021455	
Ы	13743	0.019081	
6	13402	0.018608	
Γ	13320	0.018494	
3	12345	0.017140	
Й	10931	0.015177	
Ч	10770	0.014953	
ж	7971	0.011067	
X	7116	0.009880	
Ш	6673	0.009265	
Ю	5568	0.007731	
ц	353 <b>2</b>	0.004904	
Щ	1748	0.002427	
ф	1109	0.001540	
э	983	0.001365	

#### Підсумкові значення ентропій:

```
3 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%)

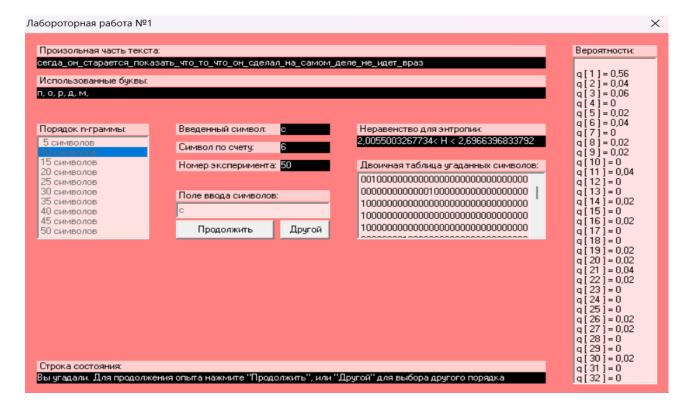
Модель H2 (з перетином):
    R (з пробілами) = 0.1934 (19.34%)
    R (без пробілів) = 0.1629 (16.29%)

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

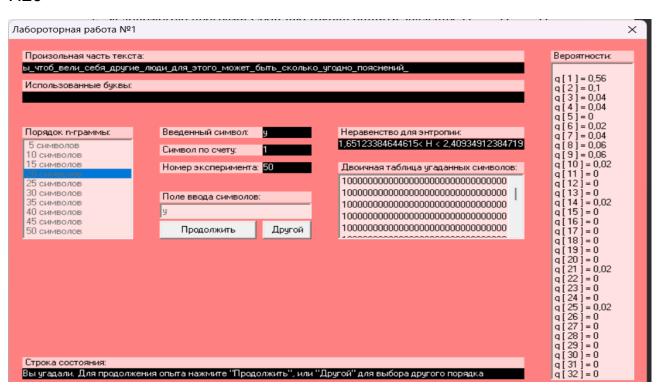
Далі було проведено експерименти з CoolPinkProgram.exe, для кожної прами по 50 експериментів. Під час експериментів навмисно підроблювати результати не намагалися, як виходило так і вгадували символи.

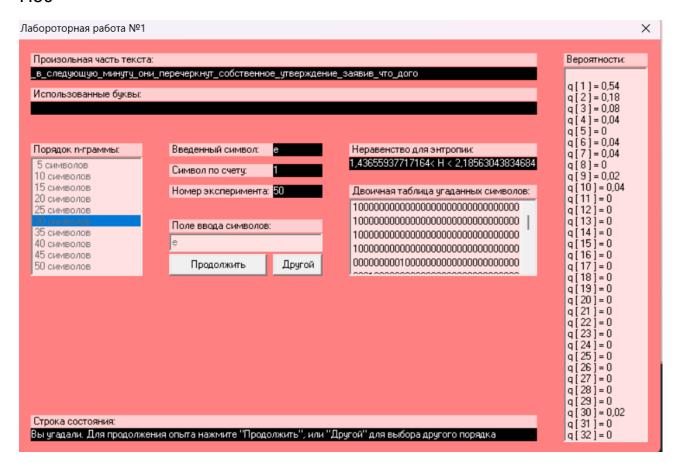
## Експерименти із CoolPinkProgram

#### H<sub>10</sub>



#### H<sub>2</sub>0





$$H_0 = \log_2 32 = 5$$

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

 $0,4606606332416 < R^{(10)} < 0,59889993464532$   $0,518130175230562 < R^{(20)} < 0,66975323071077$  $0,562873912330632 < R^{(30)} < 0,712688124565672$ 

#### 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 Н10 складає близько ~2.35 біт, тоді як Н30 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 (Н1, Н2) 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 (Н1 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.