

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**  
**“КІЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”**  
**ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ**

**Криптографія**

**КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1**  
**«Експериментальна оцінка ентропії на символ джерела**  
**відкритого тексту»**

**ФБ-32 Дорошенко Ілля**

**Мета:** 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

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
1. 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  $H_1$  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.
2. За допомогою програми CoolPinkProgram оцінити значення (10)  $H_1$ , (20)  $H_2$ , (30)  $H$ .
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

### Хід роботи:

Для дослідження було обрано текст російською мовою «TEXT.txt». Згідно з методичними вказівками, текст пройшов попередню фільтрацію:

- Усі символи, крім літер, були вилучені або замінені на пробіли.
- Прописні літери замінені на відповідні рядкові.
- Послідовності пробілів трактуються як один пробіл.
- Буква «ё» замінена на «е», а «ъ» на «ъ».
- Алфавіт дослідження склав 32 літери (без пробілу) або 33 символи (з пробілом).

При підрахунку біграм було реалізовано два підходи:

1. **Крок 1:** пари букв, що перетинаються (більш точна статистика).
2. **Крок 2:** пари букв, що не перетинаються.

На основі роботи програми було отримано такі значення ентропії та надлишковості:

A	B	C
Параметр	З пробілами	Без пробілів
$H_1$	4,383006841	4,468568723
$H_2$ (step 1)	3,975230577	4,150663708
$H_2$ (step 2)	3,974254627	4,149794329
$R(H_1)$	0,123398632	0,098023485
$R(H_2$ step 1)	0,204953885	0,162192322

Таблиці частот символів:

(3 пробілом)

(Без пробілу)

	А	В
1	Символ	Частота
2		0,162346
3	о	0,095222
4	а	0,070264
5	е	0,066722
6	и	0,055668
7	н	0,054587
8	т	0,047573
9	с	0,043704
10	л	0,042398
11	в	0,03856
12	р	0,038175
13	к	0,030044
14	д	0,025471
15	м	0,024757
16	у	0,024013
17	п	0,021519
18	я	0,019391
19	г	0,017368
20	ь	0,016753
21	ы	0,015897
22	з	0,014918
23	б	0,014467
24	ч	0,011416
25	й	0,009648
26	ж	0,008485
27	ш	0,00791
28	х	0,007145
29	ю	0,005433
30	ц	0,003388
31	э	0,002533
32	щ	0,002349
33	ф	0,001875

  

	А	В
1	Символ	Частота
2	о	0,113677
3	а	0,083882
4	е	0,079653
5	и	0,066457
6	н	0,065167
7	т	0,056793
8	с	0,052174
9	л	0,050615
10	в	0,046034
11	р	0,045574
12	к	0,035867
13	д	0,030408
14	м	0,029556
15	у	0,028667
16	п	0,025689
17	я	0,023135
18	г	0,020735
19	ь	0,02
20	ы	0,018979
21	з	0,017809
22	б	0,017271
23	ч	0,013629
24	й	0,011518
25	ж	0,01013
26	ш	0,009443
27	х	0,00853
28	ю	0,006486
29	ц	0,004045
30	э	0,003024
31	щ	0,002805
32	ф	0,002239

Частота біграм з перекриванням з пробілами:

Без пробілів:

