КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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:

- 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 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 H (10), H (20), H (30).
- 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

- 1. Відфільтруємо текст. Замінимо літри "ё" на "е" та "ъ" на "ь", також замінимо перенос рядка на пробіл. Видалимо всі символи крім літер російського алфавіту та пробілів. Видалимо зайві пробіли та збережемо текст в файл filtered.txt. Потім видалимо усі пробіли та збережемо текст в файл filtered по space.txt.
- 2. Порахуємо кількість та частоту кожної літри в обох файлах.
- 3. Порахуємо Н₁ за формулою:

$$H_1 = -\sum_{i=0}^n p_i \log_2 p_i$$

де рі – частота літри

4. Порахуємо надлишковість за формулою:

$$N = 1 - \frac{H_{\infty}}{H_0}$$

де $H_0 = \log_2 32$ для тексту з пробілами і $H_0 = \log_2 31$ для тексту без пробілів

- 5. Порахуємо кількість біграм з перетином та без в обох текстах.
- 6. Порахуємо Н₂ та надлишковість. Формула для Н₂:

$$H_2 = -\frac{1}{2} \sum_{i,j=0}^{n} p_{(i,j)} \log_2 p_{(i,j)}$$

де $p_{(i,j)}$ – частота біграми

- 7. Проведемо по 50 експериментів для $H_{10},\,H_{20},\,H_{30}.$
- 8. Порахуємо надлишковість.

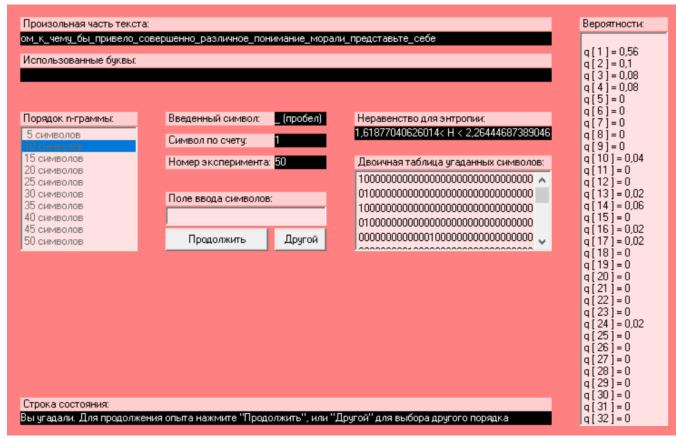
Отримані дані:

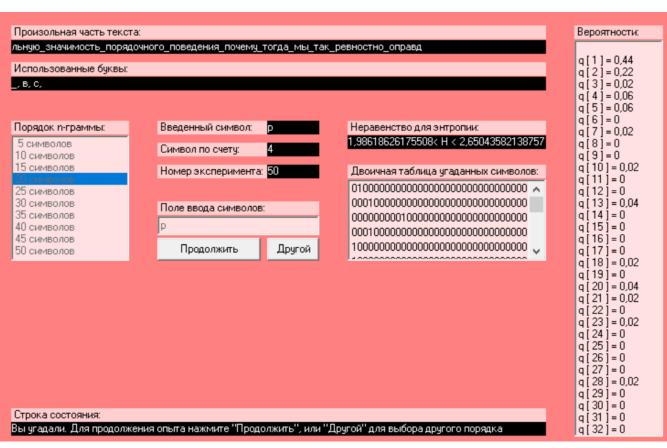
Символ	Кількість	Частота з	Частота без
		пробілами	пробілів
	120630	0,168341	-
Д	20099	0,028048	0,033726
a	43791	0,061111	0,073481
Н	36721	0,051245	0,061617
T	38309	0,053461	0,064282
e	51047	0,071237	0,085656
Л	27930	0,038977	0,046866
И	40346	0,056303	0,0677
Γ	11900	0,016607	0,019968
Ь	12680	0,017695	0,021277
p	28260	0,039437	0,04742
б	10643	0,014852	0,017859
0	60255	0,084087	0,101107
Ж	5457	0,007615	0,009157
С	32594	0,045485	0,054692
В	29020	0,040498	0,048695
R	12814	0,017882	0,021502
К	19786	0,027612	0,033201
M	20720	0,028915	0,034768
П	15025	0,020968	0,025212

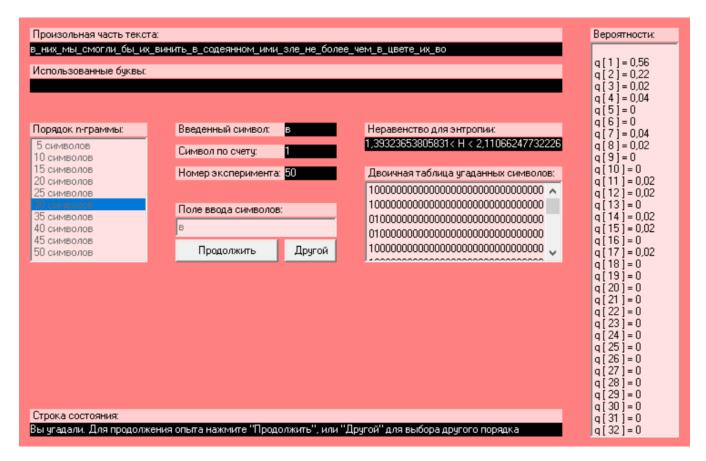
3	10764	0,015021	0,018062
ы	12220	0,017053	0,020505
У	14896	0,020788	0,024995
Ш	4588	0,006403	0,007699
X	5558	0,007756	0,009326
й	10775	0,015037	0,01808
Ч	8538	0,011915	0,014327
ц	3013	0,004205	0,005056
Щ	2084	0,002908	0,003497
Э	1532	0,002138	0,002571
ф	1264	0,001764	0,002121
Ю	3322	0,004636	0,005574

	Ентропія	Надлишковість
Н ₁ з пробілом	4.392184848990499	0.12156303020190029
Н ₁ без пробілу	4.494978570186723	0.09269268140169684
H ₂ з пробілом з перетином	3.999374726409234	0.20012505471815323
H ₂ з пробілом без перетину	3.998424417033625	0.20031511659327506
H ₂ без пробілу з перетином	4.197302636418452	0.15277829673029608
Н2 без пробілу без перетину	4.198665348127256	0.1525032346165992

CoolPinkProgram







	Ентропія	Надлишковість
H ₁₀	1,61877040626014< H	0.676245918747972> N
	<2,26444687389046	>0.547110625221908
H ₂₀	1,98618626175508< H <	0.602762747648984> N
	2,65043582138757	>0.469912835722486
H ₃₀	1,39323653805831< H	0.721352692388338> N
	<2,11066247732226	>0.577867504535548

Висновок:

В ході цієї лабораторної я навчився фільтрувати текст та аналізувати кількість та частоту літер та біграм у ньому. Також я дізнався про ентропію та надлишковість текстів.