

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

Хід роботи:

1. Відфільтруємо текст. Замінімо літри "ё" на "е" та "ь" на "ъ", також замінімо 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_no_space.txt`.
2. Порахуємо кількість та частоту кожної літри в обох файлах.
3. Порахуємо H_1 за формулою:

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

де p_i – 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 H_2 та надлишковість. Формула для H_2 :

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

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 пробілами	Частота без пробілів
	120630	0,168341	-
д	20099	0,028048	0,033726
а	43791	0,061111	0,073481
н	36721	0,051245	0,061617
т	38309	0,053461	0,064282
е	51047	0,071237	0,085656
л	27930	0,038977	0,046866
и	40346	0,056303	0,0677
г	11900	0,016607	0,019968
ь	12680	0,017695	0,021277
р	28260	0,039437	0,04742
б	10643	0,014852	0,017859
о	60255	0,084087	0,101107
ж	5457	0,007615	0,009157
с	32594	0,045485	0,054692
в	29020	0,040498	0,048695
я	12814	0,017882	0,021502
к	19786	0,027612	0,033201
м	20720	0,028915	0,034768
п	15025	0,020968	0,025212

з	10764	0,015021	0,018062
ы	12220	0,017053	0,020505
у	14896	0,020788	0,024995
ш	4588	0,006403	0,007699
х	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
H ₁ з пробілом	4.392184848990499	0.12156303020190029
H 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
H ₂ без пробілу без перетину	4.198665348127256	0.1525032346165992

CoolPinkProgram

[illegible]

Произвольная часть текста:	
<code>льную_значимость_порядочного_поведения_почему_тогда_мы_так_ревностно_оправд</code>	
Использованные буквы:	
<code>_ , в , с ,</code>	

Порядок n-граммы:	Введенный символ: <code>p</code>	Неравенство для энтропии:
<code>5 символов</code>	Символ по счету: <code>4</code>	<code>1,98618626175508 < H < 2,65043582138757</code>
<code>10 символов</code>	Номер эксперимента: <code>50</code>	Двоичная таблица угаданных символов:
<code>15 символов</code>		<code>01000000000000000000000000000000 ^</code>
<code>20 символов</code>		<code>00010000000000000000000000000000</code>
<code>25 символов</code>		<code>00000000010000000000000000000000</code>
<code>30 символов</code>		<code>00010000000000000000000000000000</code>
<code>35 символов</code>		<code>10000000000000000000000000000000 v</code>
<code>40 символов</code>		
<code>45 символов</code>		
<code>50 символов</code>		

Поле ввода символов:
<code>p</code>
Продолжить Другой

Строка состояния:
<code>Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка</code>

[illegible]

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

Висновок: