Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» Фізико-технічний інститут

КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали: ФБ-31 Аль-Фітурі Асія ФБ-31 Гриб Вероніка

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 а) частот букв і b) частот біграм в тексті, а також підрахунку с) Н1 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.

А) частота букв

1. Частота літер

Частота окремої літери обчислюється як відношення кількості появ цієї літери до загальної кількості літер у тексті

$$P(l) = \frac{f(1)}{n}$$

Дані збереглись а табличці

Створено файл 'crypto_5tables.xlsx' з 7 аркушами.

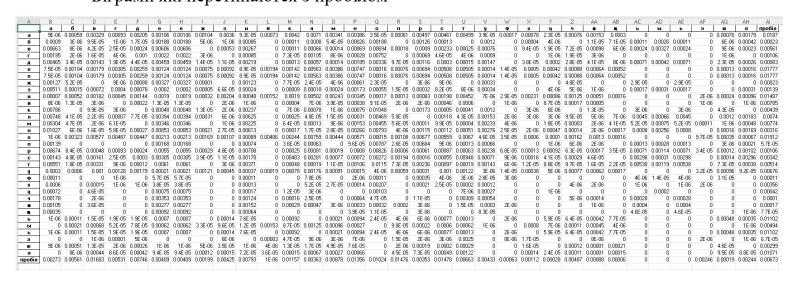
Частота літери з пробілами

	Α	В	С	
1	Літера	Кількість	Частота	
2	пробіл	173144	0.16767	
3	0	95774	0.09275	
4	a	74368	0.07202	
5	e	73361	0.07104	
6	e	73361	0.07104	
7	И	53873	0.05217	
8	н	52690	0.05102	
9	Т	51049	0.04944	
10	С	44734	0.04332	
11	Л	43373	0.042	
12	p	43076	0.04171	
13	В	41001	0.0397	
14	К	31695	0.03069	
15	M	27873	0.02699	
16	Д	25559	0.02475	
17	у	24698	0.02392	
18	п	21472	0.02079	
19	Г	18252	0.01768	
20	ь	17094	0.01655	
21	ь	17094	0.01655	
22	Я	16841	0.01631	
23	ы	14150	0.0137	
24	3	13902	0.01346	
25	б	13350	0.01293	
26	ч	11986	0.01161	
27	й	10369	0.01004	
28	ж	8371	0.00811	
29	Ш	8121	0.00786	
30	X	7379	0.00715	
31	ю	5223	0.00506	
32	ц	3753	0.00363	
33	3	2827	0.00274	
34	щ	2233	0.00216	
35	Φ	1050	0.00102	
20	I			

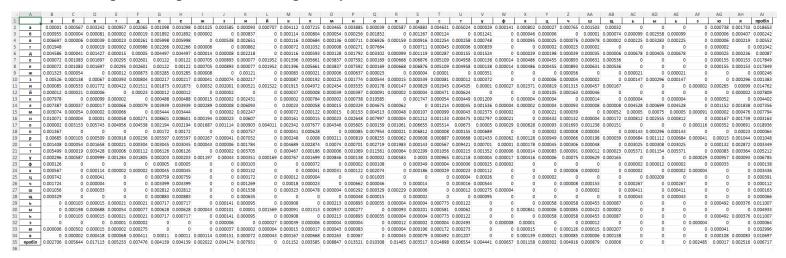
Частота літерів без пробілів

	Α	В	С	
1	Літера	Кількість	Частота	
2	0	95774	0.11143	
3	a	74368	0.08653	
4	e	73361	0.08535	
5	и	53873	0.06268	
6	н	52690	0.0613	
7	т	51049	0.05939	
8	С	44734	0.05205	
9	л	43373	0.05046	
LO	р	43076	0.05012	
11	В	41001	0.0477	
12	к	31695	0.03688	
L3	M	27873	0.03243	
L4	д	25559	0.02974	
15	у	24698	0.02874	
16	п	21472	0.02498	
L7	г	18252	0.02124	
18	ь	17094	0.01989	
19	я	16841	0.01959	
20	ы	14150	0.01646	
21	3	13902	0.01617	
22	6	13350	0.01553	
23	ч	11986	0.01395	
24	й	10369	0.01206	
25	ж	8371	0.00974	
26	ш	8121	0.00945	
27	x	7379	0.00859	
28	ю	5223	0.00608	
29	ц	3753	0.00437	
30	3	2827	0.00329	
31	щ	2233	0.0026	
32	ф	1050	0.00122	
33				

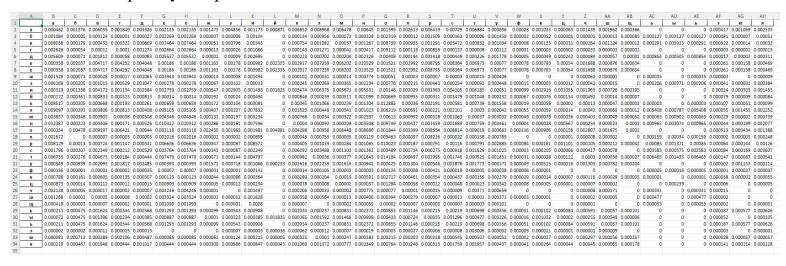
Біграми які перетинаются з пробілом



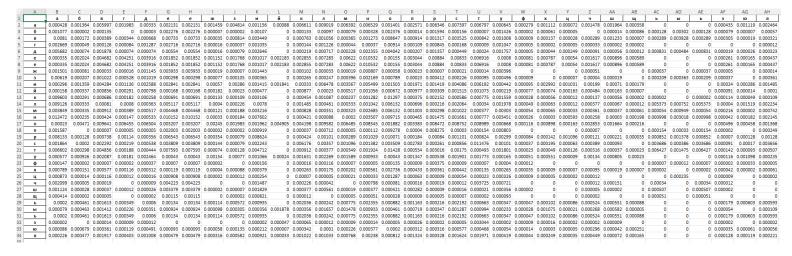
Біграми з пробілом без перетину



Без пробілу з перетином



Без пробілу без перетину



$$H(Z) = -\sum_{i=1}^{n} p_i \log p_i.$$

Обчислюємо ентропію символів за формулою:

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

Надлишковість:

	Α	В	С	
1	Показник	3 пробілами	Без пробілів	
2	H1	4.357953	4.452091	
3	R1	0.13608	0.109582	
4	Н2 (з перє	3.967175	4.138175	
5	R2 (з пере	0.213548	0.172365	
6	Н2 (без пе	3.967294	4.138428	
7	R2 (без пе	0.213524	0.172314	
8				

2. За допомогою програми CoolPinkProgram оцінити значення $\rm H~(10), H~(20)$, $\rm H~(30)$

Hадлишковість джерела відкритого тексту (мови) дорівнює $R=1-\frac{H_{\infty}}{H_0}$ і характеризує величину можливого ущільнення тексту деякою схемою кодування символів без втрати його змісту.

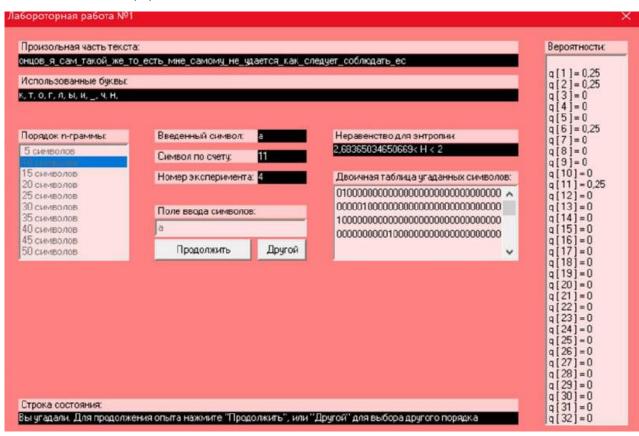
 $H\infty$ — це ентропія «нескінченного порядку» (умовна ентропія, яку оцінюють через експеримент з n-грамами, наприклад H(10), H(20), H(30)

H0=log2m, де m — кількість символів в алфавіті (для російської мови без пробілу ≈ 32 , з пробілом ≈ 33).

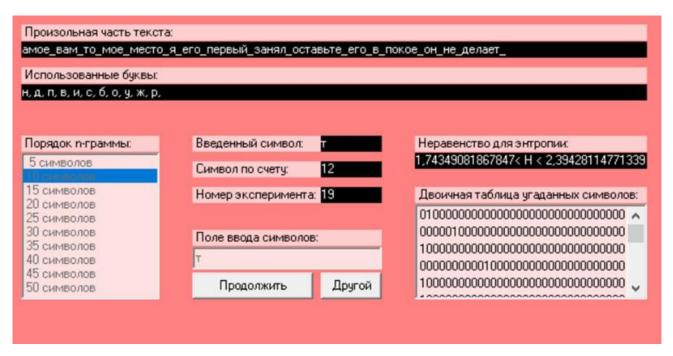
H(10)

Ентропія для H10: 2.028< H < 2.894

Надлишковість (R) 0.426< R < 0.598



тут ми багато разів вгадували що може бути після с...

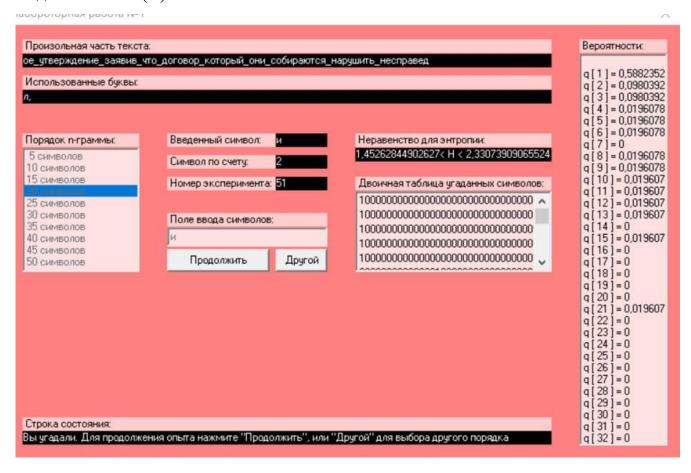


Найважче було вгадати перше слово, бо там було «амое вам_» і треба було вгадати, що буде після пробілу

H(20)

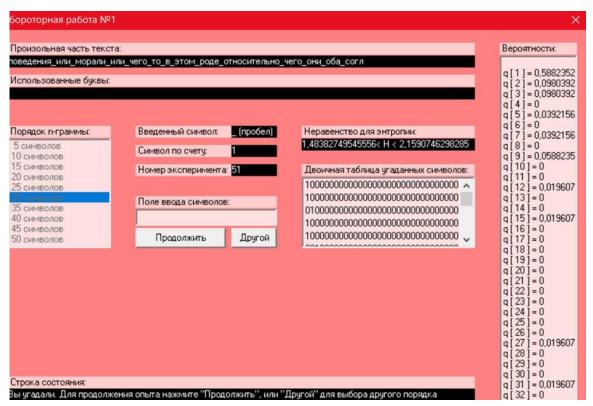
Ентропія 1.453 < H < 2.331

Надлишковість (R) 0.538< R < 0.712



H(30) Ентропія 1.484 < H < 2.159

Надлишковість (R) 0.572 < R < 0.706



Порядок п-грам	Ентропія (Н)	Надлишковість (R)
10 символів	2.028 <h<2.894< td=""><td>0.426<r<0.598< td=""></r<0.598<></td></h<2.894<>	0.426 <r<0.598< td=""></r<0.598<>
20 символів	1.453 <h<2.331< td=""><td>0.538<<i>R</i><0.712</td></h<2.331<>	0.538< <i>R</i> <0.712
30 символів	1.484 <h<2.159< td=""><td>0.572<r<0.706< td=""></r<0.706<></td></h<2.159<>	0.572 <r<0.706< td=""></r<0.706<>

Висновок:

У роботі на основі книги російською мовою підраховано частоти літер і біграм для двох варіантів (з пробілом/без пробілу) та побудовано матриці біграм з перетином і без, що загалом дало 7 таблиць результатів.

За програмою CoolPinkProgram було досліджено ентропію та надлишковість відкритого тексту за допомогою n-грамного аналізу. Метою було визначити міру невпорядкованості тексту (ентропію H) та частку повторюваної інформації (надлишковість R).

Отримані результати:

H10:2.028<H<2.894

R10: 0.426<R<0.598

H20:1.452<H<2.330

R20: 0.538<R<0.712

H30:1.483<H<2.159

*R*30: 0.572<*R*<0.706

Аналіз показав, що зі збільшенням порядку п-грам ентропія зменшується, а надлишковість зростає.