

Міністерство освіти і науки України Національний технічний
університет України “Київський політехнічний інститут ім. Ігоря
Сікорського” Фізико-технічний інститут

КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1
Експериментальна оцінка ентропії на символ джерела
відкритого тексту

Виконали студенти:
ФБ-23 Лишиленко Ангеліна
ФБ-23 Тіщенко Олександр

Київ-2024

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

Хід роботи:

1. Для того, аби виконати підрахунки, ми беремо російській текст. У нашому випадку це “Вечера на хуторі близ Диканьки”, Гоголь Н. та починаємо писати код:

Функція для підрахунку частот букв, в якій ми вибираємо у тексті лише букви російської мови

```
def let_freq(text):
    letter_counts = collections.Counter(text)
    total_letters = sum(letter_counts.values())
    letter_frequencies = {char: count / total_letters for char, count in letter_counts.items()}
    return letter_counts, letter_frequencies
```

Функція обчислює кількість і частоту біграм у тексті, з можливістю обрати варіант із перетином або без нього

```
def bi_freq(text, overlap=False):
    if overlap:
        bigrams = [text[i:i+2] for i in range(len(text)-1)]
    else:
        bigrams = [text[i:i+2] for i in range(0, len(text)-1, 2)]
    bigram_counts = collections.Counter(bigrams)
    total_bigrams = sum(bigram_counts.values())
    bigram_frequencies = {bigram: count / total_bigrams for bigram, count in bigram_counts.items()}
    return bigram_counts, bigram_frequencies
```

Функція обчислює ентропію H1 на основі частот букв, використовуючи формулу

$$H_1 = - \sum_{i=1}^n p(i) \log_2 p(i)$$

```
def calculate_H1(letter_frequencies):  
    return -sum(frequency * math.log2(frequency) for frequency in letter_frequencies.values())
```

Функція обчислює ентропію H2 на основі частот біграм, формула якої

$$H_2 = - \frac{1}{2} \sum_{i,j} p(i,j) \log_2 p(i,j)$$

```
def calculate_H2(bigram_frequencies):  
    return -sum(frequency * math.log2(frequency) for frequency in bigram_frequencies.values()) / 2
```

Далі в нас присутні функції для зберігання результату у csv форматі, функція для обробки тексту, що повертає всі підрахунки, завантаження тексту та обробка з і без пробілів та виведення значень ентропії

В результаті ми отримуємо 6 таблиць та ось такий текст у консолі

```
bigram_bez_peretuny_with_spaces  
bigram_bez_peretuny_without_spaces  
bigram_z_peretunom_with_spaces  
bigram_z_peretunom_without_spaces  
letter_frequencies_with_spaces  
letter_frequencies_without_spaces  
Обробка тексту з пробілами...  
Обробка тексту без пробілів...  
Результат записано у CSV файли  
Ентропія H1 з пробілами: 4.376771855603328  
Ентропія H2 без перетину з пробілами: 3.9738468135581932  
Ентропія H2 з перетином з пробілами: 3.975113332378612  
Ентропія H1 без пробілів: 4.480412919199987  
Ентропія H2 без перетину без пробілів: 4.155916668456753  
Ентропія H2 з перетином без пробілів: 4.1557080546217025  
Press any key to continue . . . |
```

letter_frequencies_with_spaces

ФайлЗмінитиВиглядВставитиФорматДаніІнструменти

100%грн. %0.00123За ум...

A1	Symbol				
	A	B	C	D	E
1	Symbol	Count	Frequency		
2		4377	0.011002186863735766		
3	В	574	0.0014428273383103336		
4	е	25693	0.06458286202649373		
5	ч	4692	0.011793982354271926		
6	р	12791	0.03215192418872383		
7	а	25897	0.06509564386798382		
8		63445	0.15947766633989394		
9	н	17576	0.04417967473543976		
10	х	2944	0.007400145790915718		
11	у	10333	0.025973405726063898		
12	т	18162	0.045652665711484804		
13	о	33618	0.08450343111379233		
14	б	5482	0.013779755171807054		
15	л	14706	0.03696538043888094		
16	и	19306	0.0485282658421939		
17	з	5316	0.013362491516476888		
18	д	376	0.0009451273156876052		
19	к	12532	0.03150089234094965		
20	ь	6305	0.015848477993112637		
21	п	627	0.0015760500716386395		
22	в	13389	0.03365507880250358		

letter_frequencies_without_spaces

ФайлЗмінитиВиглядВставитиФорматДані

100%грн. %0.00123

A1	Symbol				
	A	B	C	D	
1	Symbol	Count	Frequency		
2		4377	0.013089701990220854		
3	В	574	0.0017165841769218118		
4	е	25693	0.0768365805882441		
5	ч	4692	0.014031729892190141		
6	р	12791	0.03825231394948936		
7	а	25897	0.0774466558000948		
8	н	17576	0.052562166365118054		
9	х	2944	0.008804222677452636		
10	у	10333	0.030901505749360767		
11	т	18162	0.054314637319257744		
12	о	33618	0.10053680637588408		
13	б	5482	0.01639427605903375		
14	л	14706	0.043979245480508994		
15	и	19306	0.05773584341402874		
16	з	5316	0.015897842307519776		
17	д	376	0.0011244523528268313		
18	к	12532	0.03747775767453684		
19	ь	6305	0.01885551086322652		
20	п	627	0.0018750841096341045		
21	в	13389	0.04004067168084693		
22					

bigram_z_peretunom_with_spaces

ФайлЗмінитиВиглядВставитиФорматДаніІнс

100%грн. %0.001233

Symbol	Count	Frequency	
	2643	0.006643557910559562	
В	76	0.00019103685251703622	
Ве	52	0.0001307094254063932	
еч	396	0.0009954025473256098	
че	1145	0.002878121001736927	
ер	2127	0.005346518227680737	
ра	2118	0.005323895442514246	
а	4866	0.01223138584668287	
н	5641	0.014179459013797385	
на	3626	0.00911446877929965	
х	796	0.002000859665836327	
ху	103	0.0002589052080165096	
ут	574	0.0014428309650628788	
то	4455	0.01198278657413109	
ор	2245	0.005643128077641398	
ре	1854	0.004660293744297173	
е	5282	0.013277061249934018	
б	2426	0.006098097423767498	
бл	290	0.0007289564109202697	
ли	2152	0.005409359297587657	
из	569	0.0014302627510814948	

bigram_z_peretunom_without_spaces

ФайлЗмінитиВиглядВставитиФорматДаніІнс

100%грн. %0.00123

Symbol	Count	Frequency	
	2643	0.007904086319919613	
В	76	0.00022728360208622422	
Ве	52	0.00015550983300636394	
еч	497	0.00148631513469544	
че	1146	0.003427197473563328	
ер	2254	0.006740753146083545	
ра	2118	0.00633403512129767	
ан	1931	0.005774797837217092	
на	3628	0.010849801425905546	
ах	376	0.0011244557155844776	
ху	123	0.0003678405665342839	
ут	665	0.001988731518254462	
то	4580	0.01369682759940667	
ор	2395	0.007162424039427724	
ре	1856	0.005550504808842528	
еб	931	0.002784224125556247	
бл	291	0.0008702569500933059	
ли	2264	0.006770658883200153	
из	763	0.002281807741997225	
зД	7	2.0934015981625914e-05	
н..	20	9.67362276394845e-06	

bigram_bez_peretuny_with_spaces

A	B	C	D
Symbol	Count	Frequency	
	1316	0.006615891209813237	
Be	27	0.00013573636980619862	
че	552	0.002775054671593394	
ра	1058	0.005318854787220672	
н	2808	0.014116582459844658	
а	2432	0.012226327828469446	
ху	53	0.00026644546665661214	
то	2296	0.011542618706482669	
ре	951	0.004780936580951663	
б	1252	0.006294145740642988	
ли	1074	0.005399291154513234	
з	264	0.0013272000603272755	
Ди	14	7.038182138099188e-05	
ка	1508	0.007581127617323983	
нь	238	0.001196490963476862	
ки	635	0.0031923183269235603	
По	135	0.0006786818490309932	

bigram_bez_peretuny_without_spaces

A	B	C	D
Symbol	Count	Frequency	
	1281	0.007661849849275085	
Be	33	0.00019737786496961578	
че	541	0.003235800756017034	
ра	1068	0.006387865448107565	
на	1786	0.010682329298052538	
ху	58	0.000346906550552658	
то	2289	0.013690846451983348	
ре	937	0.0056043351356524235	
бл	144	0.0008612852289583234	
из	378	0.002260873726015599	
Ди	14	8.373606392650366e-05	
ка	1565	0.009360495717498444	
нь	244	0.0014593999712904924	
ки	742	0.004438011388104694	
По	132	0.0007895114598784631	
ве	886	0.005299296617063017	
ст	1638	0.009797119479400929	
и,	425	0.002541987654911718	
да	750	0.004485860567491268	
нн	373	0.0022309679888989904	
ье	233	0.0013936073496339538	

Отже наші значення ентропії

Ентропія	з пробілами	без пробілів
H1	4.376771855603328	4.48041291919998
H2(з перетином)	3.97511333237861	4.15570805462170
H2(без перетину)	3.973846813558193	4.15591666845675

2. Оцінка значень (10) Н , (20) Н , (30) Н за допомогою програми CoolPinkProgram:

(10)H:

(20)H:

[illegible]

(30)H:

Произвольная часть текста:

м_не_будут_попадаться_индивид

Использованные буквы:

Порядок n-граммы:

5

|||||

10

|||||

15

|||||

20

|||||

25

|||||

30

|||||

35

|||||

40

|||||

45

|||||

50

|||||

Введенный символ:

Символ по счету:

Номер эксперимента:

74

Неравенство для энтропии:

$1,39303373156825 < H < 2,22770755574721$

Двоичная таблица угаданных символов:

10000000000000000000000000000000

10000000000000000000000000000000

10000000000000000000000000000000

10000000000000000000000000000000

00001000000000000000000000000000

Поле ввода символов:

⬅⬅⬅⬅

➡➡➡➡

Вероятности:

q[1] = 0,6027397

q[2] = 0,1643835

q[3] = 0,0273972

q[4] = 0,0136986

q[5] = 0,0273972

q[6] = 0,0136986

q[7] = 0,0136986

q[8] = 0,0136986

q[9] = 0,0136986

q[10] = 0

q[11] = 0,013698

q[12] = 0

q[13] = 0

q[14] = 0

q[15] = 0,013698

q[16] = 0,013698

q[17] = 0

q[18] = 0,027397

q[19] = 0

q[20] = 0

q[21] = 0,013698

q[22] = 0

q[23] = 0

q[24] = 0,013698

q[25] = 0

q[26] = 0

q[27] = 0

q[28] = 0

q[29] = 0

q[30] = 0,013698

q[31] = 0

q[32] = 0

Строка состояния:

	Ентропія	Надлишковість
(10)H	$1,7609216260059 < H < 2,52110339978607$	$0.4957793200427859 < R < 0,64781567479882$
(20)H	$1,14153173720053 < H < 1,82107547724955$	$0.63578490455009 < R < 0.771693652559894$
(30)H	$1,39303373156825 < H < 2,22770755574721$	$0.5544584888505579 < R < 0.72139325368635$

Висновки:

У ході виконання лабораторної роботи ми закріпили знання з ентропії на символ джерела та надлишковості, також набули практичних навичок в експериментальній оцінці ентропії на символ джерела. Також дізнались про програму CoolPinkProgram яку в подальшому використали для знаходження ентропії.