

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 групи ФБ-05
Акбаров Елшодбек, Олифіренко Іван
Варіант 7

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

Пишемо програму 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. Для коду використовуємо модулі: `re` – для модифікації тексту, `math` – отримати логарифм у формулі надлишковості, `collections` – використання Counter для підрахунків в тексті, `pandas` – запису в файл.

В результаті роботи було створено програму з функціями: очистки тексту, обчислення ентропії, частоту букв, підрахунок біграм, частота біграм, надлишковість.

а	0,10937
б	0,01922
в	0,03829
г	0,01498
д	0,02967
е	0,07376
ё	0
э	0,00099
ж	0,00933
з	0,0171
и	0,06458
ы	0,02006
й	0,00989

а	0,09315
б	0,01637
в	0,03262
г	0,01276
д	0,02527
е	0,06282
ё	0
э	0,00084
ж	0,00794
з	0,01456
и	0,055
ы	0,01709
й	0,00842

к	0,03928
л	0,05652
м	0,03066
н	0,06133
о	0,09368
п	0,0325
р	0,04621
с	0,04931
т	0,04974
у	0,0366
ф	0,00155
х	0,0089
ц	0,00509
ч	0,01357
ш	0,01342
щ	0,00311
ъ	0
ь	0,02317
ю	0,00749
я	0,02063

к	0,03346
л	0,04814
м	0,02612
н	0,05223
о	0,07979
п	0,02768
р	0,03935
с	0,042
т	0,04236
у	0,03117
ф	0,00132
х	0,00758
ц	0,00433
ч	0,01155
ш	0,01143
щ	0,00265
ъ	0
ь	0,01974
ю	0,00638
я	0,01757
	0,14827

Зліва частота літер в тексті без пробілів, справа – з.

Ентропія монограм без пробілами: $H_1 = 4.47548549103316$, $R = 0.11278036863575747$

Ентропія монограм з пробілів: $H_1 = 4.417222324314814$, $R = 0.13174356999741765$

Частота перехреснії біграм з пробілами:

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ь	ю	я					
а	0,00012	0,00313	0,00614	0,00024	0,00229	0,00048	0	0	0,00072	0,00217	0,00072	0	0,00036	0,00337	0,01481	0,00349	0,00289	0	0,00193	0,00628	0,00421	0,00614	0,00036	0,00048	0,00132	0	0,00108	0,00096	0,00048	0	0	0,00048	0,00385	0,02468	
б	0,00289	0	0	0	0	0,00181	0	0	0	0	0,00193	0	0,00012	0,00048	0	0,00036	0,00277	0	0,00006	0,00048	0	0,00253	0	0	0	0	0	0	0	0	0	0	0,00012		
в	0,00421	0	0,00012	0	0,00036	0,00578	0	0	0	0,00036	0,00193	0,00181	0	0	0,00241	0,00036	0,00144	0,00085	0	0,0006	0,00193	0	0,00048	0	0	0,00012	0,00012	0,00156	0	0	0,00048	0	0,00024	0,00445	
г	0,00132	0	0	0	0,00084	0	0	0	0	0,00072	0	0	0	0,00012	0,00578	0,00024	0,0012	0	0	0,00048	0	0	0,00048	0	0	0	0	0	0	0	0	0	0,00084		
д	0,00506	0	0,00072	0	0	0,00421	0	0	0	0	0	0,00096	0,00006	0,00024	0	0	0,00156	0,00397	0,00012	0,0006	0,00036	0,00012	0,00193	0	0	0,00048	0	0	0	0,00048	0	0,00024	0,0012		
е	0	0,00072	0,0012	0,00217	0,00253	0,00217	0	0	0,00024	0,00132	0,00012	0	0,00193	0,00084	0,00734	0,00385	0,00554	0,00024	0,00096	0,00614	0,00277	0,00481	0,00012	0	0,00024	0,00048	0,00132	0,00084	0,00024	0	0	0,00024	0,01444		
ж	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
з	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
и	0,00132	0	0	0	0,00072	0,00313	0	0	0	0	0,00144	0	0	0,00012	0	0,00036	0,00012	0	0,00036	0,00012	0	0,00012	0	0	0,00036	0	0	0,00012	0	0	0	0	0	0,00024	
й	0,00096	0,00096	0,00072	0,00012	0,00048	0,00036	0	0	0,00036	0	0,00072	0,00012	0	0,00024	0,00036	0,00012	0,0012	0,00024	0	0	0	0,00012	0	0	0	0,00012	0	0	0	0	0	0	0	0,00012	0,00281
к	0	0,00036	0,00084	0,00024	0,00048	0,00041	0	0	0,00012	0,00036	0	0	0,00132	0,00012	0,00181	0,00084	0,00024	0	0,0006	0,00024	0,00036	0,00048	0	0	0,00036	0	0	0	0	0	0	0	0	0,00045	
л	0	0	0,00012	0	0	0	0	0	0	0	0	0	0	0,00024	0	0,00024	0,0006	0	0	0,00012	0	0	0,00012	0	0	0,00012	0	0	0,00012	0,00012	0	0	0	0,00062	
м	0,00079	0	0,0006	0	0	0,00096	0	0	0	0	0	0,00061	0	0	0,00024	0,00072	0	0,00108	0,00746	0	0,00241	0	0	0	0	0	0	0	0	0	0	0	0	0,00409	
н	0,01095	0	0,00012	0,00036	0,00012	0,00481	0	0	0	0	0,0059	0,00072	0	0,00108	0	0	0,00012	0,00506	0	0	0,00193	0,00024	0,00325	0	0	0	0,00012	0	0	0	0,00542	0,0006	0,00084	0,0085	
о	0,00433	0,00024	0	0	0	0,00253	0	0	0	0,00361	0,00084	0,00036	0,00036	0,00012	0,00096	0,00349	0,00024	0	0,00012	0,00012	0,00205	0	0	0	0,00012	0	0	0	0	0	0	0,00048	0,0009		
п	0,01456	0	0	0	0	0,00088	0	0	0	0,00628	0,00445	0	0,00048	0	0,00012	0,00217	0,0071	0	0,00012	0,00036	0,00024	0,00385	0,00012	0,00012	0,00036	0	0	0,00048	0	0	0,00156	0,00084	0,0012		
р	0	0,00241	0,00409	0,00397	0,00506	0,0006	0	0	0,00144	0,00084	0,00084	0	0,00325	0,00349	0,00421	0,00734	0,00457	0	0,00156	0,00506	0,00457	0,00758	0	0,00024	0,00096	0	0,00277	0,00096	0,00012	0	0,00036	0,00108	0,0134		
с	0,00578	0	0	0	0	0,00241	0	0	0	0	0,00072	0,00048	0	0	0,00048	0,00144	0	0	0,00827	0,00012	0,00596	0,00012	0	0,00084	0	0	0	0,00012	0	0	0	0	0,00024	0	
т	0,00746	0	0,00012	0,00012	0,00034	0,00566	0	0	0,00024	0,00012	0,00181	0,00289	0	0,00036	0,00024	0,00036	0,00096	0,00086	0,00012	0,00012	0,00024	0,00132	0,00049	0	0,00024	0	0,00012	0,00072	0,00012	0	0,00181	0,00012	0,00169	0	
у	0,00096	0,00012	0,00156	0,00012	0,00012	0,00325	0	0	0	0,00156	0,00072	0	0,00205	0,0012	0,00169	0,00181	0,00301	0,00144	0,00012	0,0006	0,00086	0,00132	0	0,00024	0,00012	0	0,00012	0	0	0,00457	0,00036	0,00372	0,00313		
ф	0,00843	0	0,00084	0	0	0,00373	0	0	0	0	0,00337	0,00132	0	0,00006	0,0006	0	0,0012	0,00758	0,00024	0,00301	0,0006	0	0,00012	0	0	0,00048	0	0	0	0,00421	0	0,00072	0,00421		
х	0	0,00096	0,00036	0,00156	0,00096	0,00048	0	0	0,0012	0,00048	0,00012	0	0	0,00108	0,00132	0,0012	0,00048	0	0,0012	0,00132	0,0012	0,00132	0	0	0,00108	0	0,0012	0,00349	0,00036	0	0,00241	0,00212	0,00722		
ц	0,00024	0	0	0	0	0,00024	0	0	0	0	0,00036	0	0	0	0	0	0,00036	0	0	0,00036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ч	0,00072	0	0	0	0	0	0	0	0	0	0,00036	0	0	0,00012	0,00024	0,00012	0,0006	0,00253	0	0,00024	0,00012	0,00012	0,00024	0	0	0	0	0	0	0	0	0	0,00217		
ш	0,00132	0	0	0	0	0,00132	0	0	0	0	0,00012	0,00084	0	0	0	0	0	0	0	0	0	0,00012	0	0	0	0	0	0	0	0	0	0	0	0,0006	
щ	0,00241	0	0	0	0	0,00241	0	0	0	0	0,00169	0	0	0,00156	0	0	0,00036	0,00012	0	0	0	0,00169	0,00048	0	0	0	0	0	0	0	0	0	0,00072		
ъ	0,00156	0	0	0	0	0,00217	0	0	0	0	0,00277	0,00048	0	0	0,00072	0,00048	0	0	0	0	0,00012	0,00024	0	0	0	0	0	0	0	0	0	0,00036	0	0	
ь	0,00048	0	0	0	0	0,00072	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00012	
ю	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
я	0	0,00084	0	0,00012	0	0,00072	0	0	0,00012	0	0,00048	0	0	0,00108	0	0,00036	0,0012	0	0,00012	0	0,00084	0,00012	0	0	0	0,00048	0,00024	0,00036	0	0	0	0,00096	0,00229	0,00939	
а	0	0,00012	0	0,00012	0,00072	0	0	0	0,00012	0,00012	0	0	0	0	0	0	0,00012	0	0	0,00072	0,00012	0,00024	0	0	0	0	0	0,00012	0	0,00036	0	0	0,00012	0	0,00325
б	0	0	0,00012	0,00024	0,00072	0	0	0	0,00036	0,00012	0	0	0,00012	0,00024	0,00144	0,00048	0,00006	0,00048	0	0,00108	0,00012	0	0	0	0,00036	0,00024	0,00012	0	0,00036	0	0	0	0	0,00851	
в	0,00373	0,00614	0,01143	0,00265	0,00818	0,00217	0	0,00072	0,00289	0,00554	0,0077	0	0	0,00927	0,00205	0,00421	0,01613	0,00939	0,01769	0,00361	0,01709	0,00385	0,00469	0,00036	0,0012	0,00048	0,00301	0,00084	0,00012	0	0	0,00012	0,00036	0,00385	

10 символів:

Лабораторная работа №1

Произвольная часть текста:
_где_восх

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ:

Символ по счету:

Номер эксперимента: 51

Неравенство для энтропии:
 $2,35437963535664 < H < 3,04810207019148$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
00000000000000000000000000000000
00000000000000000000000000000000
00000100000000000000000000000000
00001000000000000000000000000000

Вероятности:

$q[1] = 0,36$
$q[2] = 0,2$
$q[3] = 0,04$
$q[4] = 0,06$
$q[5] = 0,06$
$q[6] = 0,04$
$q[7] = 0,04$
$q[8] = 0$
$q[9] = 0,04$
$q[10] = 0$
$q[11] = 0$
$q[12] = 0$
$q[13] = 0$
$q[14] = 0,04$
$q[15] = 0$
$q[16] = 0$
$q[17] = 0$
$q[18] = 0$
$q[19] = 0$
$q[20] = 0,02$
$q[21] = 0$
$q[22] = 0,02$
$q[23] = 0$
$q[24] = 0$
$q[25] = 0,02$
$q[26] = 0,04$
$q[27] = 0$
$q[28] = 0,02$
$q[29] = 0$
$q[30] = 0$
$q[31] = 0$
$q[32] = 0$

Поле ввода символов:

Продолжить Другой

Строка состояния:

$$2,3543796353 < H_{10} < 3,0481020701$$

$$0,39037958598 < R < 0,52912407294$$

20 символів:

Лабораторная работа №1

Произвольная часть текста:
ожих_взглядов_на_мо

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ:

Символ по счету:

Номер эксперимента: 51

Неравенство для энтропии:
 $1,86512527538893 < H < 2,54683034484409$

Двоичная таблица угаданных символов:

00010000000000000000000000000000
01000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

$q[1] = 0,5$
$q[2] = 0,16$
$q[3] = 0,02$
$q[4] = 0,08$
$q[5] = 0,04$
$q[6] = 0$
$q[7] = 0$
$q[8] = 0,02$
$q[9] = 0$
$q[10] = 0$
$q[11] = 0,06$
$q[12] = 0$
$q[13] = 0,02$
$q[14] = 0$
$q[15] = 0$
$q[16] = 0$
$q[17] = 0$
$q[18] = 0$
$q[19] = 0$
$q[20] = 0,02$
$q[21] = 0$
$q[22] = 0$
$q[23] = 0,02$
$q[24] = 0$
$q[25] = 0,02$
$q[26] = 0$
$q[27] = 0$
$q[28] = 0,02$
$q[29] = 0$
$q[30] = 0$
$q[31] = 0$
$q[32] = 0,02$

Поле ввода символов:

Продолжить Другой

Строка состояния:

