

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:
Дудченко И.В і Терпило С.Е

Київ – 2022

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

Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку $1 H$ та $2 H$ за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення $1 H$ та $2 H$ на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення $1 H$ та $2 H$ 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

Перед виконанням роботи були розглянуті теоретичні відомості в методичних вказівках. В якості експериментального тексту була взята книга «Good Omens» в перекладі на російську мову. Оригінал тексту можна знайти у файлі badtxt.txt. Відредагований текст з пробілами міститься у файлі spaces.txt, а без пробілів у nospaces.txt. В ході виконання роботи було прийнято рішення пробіли замінити на «_» для кращого сприйняття. Усі таблиці також наведені у файлах з відповідними назвами. Єдина відмінність - таблиці біграм були виконанні у двох варіантах(таблиці у вигляді матриці і звичайні).

Аналіз результатів для тексту з пробілами

Текст без пробілів

а	0,081296
б	0,016728
в	0,046564
г	0,018979
д	0,031136
е	0,087972
ё	0
э	0,003554
ж	0,011568
з	0,017309
и	0,063784
ы	0,017417
й	0,010371
к	0,0328
л	0,046495
м	0,03076
н	0,067019
о	0,111932
п	0,027224
р	0,039487
с	0,053293
т	0,063036
у	0,026203
ф	0,002211
х	0,007079
ц	0,003261
ч	0,018446
ш	0,008342
щ	0,002955
ъ	0,00022
ь	0,022638
ю	0,005954
я	0,023967

Текст з пробілами

1	а	0,067572
2	б	0,013904
3	в	0,038703
4	г	0,015775
5	д	0,02588
6	е	0,073121
7	ё	0
8	э	0,002954
9	ж	0,009615
10	з	0,014387
11	и	0,053017
12	ы	0,014476
13	й	0,00862
14	к	0,027263
15	л	0,038646
16	м	0,025567
17	н	0,055705
18	о	0,093036
19	п	0,022628
20	р	0,032821
21	с	0,044296
22	т	0,052395
23	у	0,021779
24	ф	0,001838
25	х	0,005884
26	ц	0,002711
27	ч	0,015332
28	ш	0,006934
29	щ	0,002456
30	ъ	0,000183
31	ь	0,018817
32	ю	0,004949
33	я	0,019921
34		0,168816

Біграми

Текст з пробілами

1	аа	0
2	аб	0,00040969
3	ав	0,00288247
4	аг	0,00075354
5	ад	0,00175216
6	ае	0,00127663
7	ає	0
8	аэ	0
9	аж	0,0017229
10	аз	0,0032446
11	аи	0,00022314
12	аы	0
13	ай	0,00077549
14	ак	0,00540645
15	ал	0,00805481
16	ам	0,003219
17	ан	0,00340555
18	ао	0,00001097
19	ап	0,00069135
20	ар	0,00210332
21	ас	0,00544303
22	ат	0,0037933
23	ау	0,00006584
24	аф	0,00021216
25	ах	0,00069867
26	ац	0,00002561
27	ач	0,00089254
28	аш	0,00079743
29	ащ	0,00016827
30	аъ	0
31	аь	0
32	аю	0,00076451
33	ая	0,00187287
34	а	0,01660345
35	ба	0,00062185
36	бб	0
37	бв	0,00002561
38	бг	0,00001097

Текст без пробілів

1	аа	0,00068654
2	аб	0,00105182
3	ав	0,00584
4	аг	0,00149191
5	ад	0,00308944
6	ае	0,00206843
7	ає	0
8	аэ	0,00030806
9	аж	0,00221806
10	аз	0,00419406
11	аи	0,00170315
12	аы	0
13	ай	0,00092419
14	ак	0,00758716
15	ал	0,00991964
16	ам	0,00468696
17	ан	0,00657935
18	ао	0,00120145
19	ап	0,00238969
20	ар	0,0029266
21	ас	0,0082473
22	ат	0,00560675
23	ау	0,00061173
24	аф	0,00033447
25	ах	0,00091979
26	ац	0,00009242
27	ач	0,00169875
28	аш	0,00105622
29	ащ	0,00024645
30	аъ	0
31	аь	0
32	аю	0,0009638
33	ая	0,00250411
34	ба	0,00071295
35	бб	0
36	бв	0,00004841
37	бг	0,0000088
38	бд	0,00002641

Перехрестні біграми

Текст з пробілами

1	аа	0
2	аб	0,00040786
3	ав	0,00303793
4	аг	0,00077732
5	ад	0,00173936
6	ае	0,00129492
7	аё	0
8	аз	0
9	аж	0,00166803
10	аз	0,00313853
11	аи	0,00025057
12	аы	0
13	ай	0,00081572
14	ак	0,00534428
15	ал	0,00800727
16	ам	0,00327936
17	ан	0,00350067
18	ао	0,00000732
19	ап	0,00063648
20	ар	0,00202285
21	ас	0,00538817
22	ат	0,00381891
23	ау	0,00006036
24	аф	0,00025606
25	ах	0,00068953
26	ац	0,00002926
27	ач	0,0008962
28	аш	0,00081024
29	ащ	0,00018107
30	аъ	0
31	аь	0
32	аю	0,00083219
33	ая	0,00186373
34	а	0,01681747
35	ба	0,0005871
36	бб	0
37	бв	0,00003658
38	бг	0,00000732

Текст без пробілів

1	аа	0,00066894
2	аб	0,00112443
3	ав	0,00573218
4	аг	0,00156012
5	ад	0,00306303
6	ае	0,00211684
7	аё	0
8	аз	0,00038728
9	аж	0,00219385
10	аз	0,00419626
11	аи	0,00165474
12	аы	0
13	ай	0,0009814
14	ак	0,00753655
15	ал	0,00989763
16	ам	0,00479038
17	ан	0,00627569
18	ао	0,00123445
19	ап	0,0024161
20	ар	0,00284959
21	ас	0,00815708
22	ат	0,00580699
23	ау	0,00058752
24	аф	0,00035647
25	ах	0,0009792
26	ац	0,00007261
27	ач	0,00168115
28	аш	0,00104962
29	ащ	0,00022665
30	аъ	0
31	аь	0
32	аю	0,00100341
33	ая	0,00269555
34	ба	0,00071295
35	бб	0,0000066
36	бв	0,00005061
37	бг	0,0000088
38	бд	0,000022

Значення Н і R

	E	R
letters including space	4.363926750671057	0.1422194349436191
letters not including space	4.4622350062498795	0.11540714292613852
bigrams including spaces	3.9464790335576905	0.224273639591288
crossed bigrams including spaces	3.946960015449667	0.22417909700552663
bigrams not including spaces	4.133355455939298	0.18060417997930378
crossed bigrams not including spaces	4.134752310896859	0.1803272676436477

CoolPinkProgram

[illegible]

Лабораторная работа №1

Произвольная часть текста:

сподствовавшие_скажем_в_древнем_египте_вавилоне_индии_китае_греции_и_риме_т

Использованные буквы:

и,

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:

е

Символ по счету:

2

Номер эксперимента:

53

Неравенство для энтропии:

$1,91946409282983 < H < 2,68554353314245$

Двоичная таблица угаданных символов:

00000000001000000000000000000000

10000000000000000000000000000000

10000000000000000000000000000000

00000000000001000000000000000000

10000000000000000000000000000000

Поле ввода символов:

е

Продолжить

Другой

Вероятности:

$q[1] = 0,5471698$

$q[2] = 0,0943396$

$q[3] = 0,0188679$

$q[4] = 0,0188679$

$q[5] = 0,0188679$

$q[6] = 0,0188679$

$q[7] = 0,0566037$

$q[8] = 0$

$q[9] = 0,0188679$

$q[10] = 0,018867$

$q[11] = 0,018867$

$q[12] = 0,018867$

$q[13] = 0$

$q[14] = 0,037735$

$q[15] = 0$

$q[16] = 0,018867$

$q[17] = 0$

$q[18] = 0$

$q[19] = 0$

$q[20] = 0$

$q[21] = 0$

$q[22] = 0$

$q[23] = 0$

$q[24] = 0$

$q[25] = 0,018867$

$q[26] = 0,037735$

$q[27] = 0$

$q[28] = 0$

$q[29] = 0$

$q[30] = 0,018867$

$q[31] = 0,018867$

$q[32] = 0$

Строка состояния:

Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

Лабораторная работа №1

Произвольная часть текста:

е_могут_нарушить_животные_то_есть_человек_не_может_не_подчиняться_тем_закон

Использованные буквы:

п, т, б, й, ц, у,

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:

е

Символ по счету:

7

Номер эксперимента:

53

Неравенство для энтропии:

$1,7736818696663 < H < 2,55325210980818$

Двоичная таблица угаданных символов:

10000000000000000000000000000000

10000000000000000000000000000000

10000000000000000000000000000000

10000000000000000000000000000000

000000000000000000000000100000000

Поле ввода символов:

е

Продолжить

Другой

Вероятности:

$q[1] = 0,5283018$

$q[2] = 0,1320754$

$q[3] = 0,0188679$

$q[4] = 0,0188679$

$q[5] = 0,0754716$

$q[6] = 0,0377358$

$q[7] = 0,0377358$

$q[8] = 0,0188679$

$q[9] = 0,0377358$

$q[10] = 0$

$q[11] = 0$

$q[12] = 0$

$q[13] = 0$

$q[14] = 0$

$q[15] = 0$

$q[16] = 0$

$q[17] = 0$

$q[18] = 0$

$q[19] = 0,018867$

$q[20] = 0,018867$

$q[21] = 0$

$q[22] = 0$

$q[23] = 0$

$q[24] = 0,018867$

$q[25] = 0$

$q[26] = 0$

$q[27] = 0$

$q[28] = 0$

$q[29] = 0$

$q[30] = 0,018867$

$q[31] = 0,018867$

$q[32] = 0$

Строка состояния:

Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

Отримані результати

$$2,0465 < H(10) < 2,7066$$

$$1,9194 < H(20) < 2,6855$$

$$1,7736 < H(30) < 2,5532$$

Висновки

Під час виконання лабораторної роботи, ми отримали змогу ознайомитись з такими поняттями як ентропія, надлишковість та обрахувати їх на практиці. Успішно були проведені експерименти на різних видах тексту(з пробілом та без), прорахована частота монограм та біграм, проведено знайомство з невеличкою програмою. Набуті навички знадобляться у майбутніх лаб.роботах та у професійній діяльності.