

МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ФІЗИКО- ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1
з дисципліни
Криптографія

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав студент групи ФБ-91
Олександр Чернов

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.

Завдання

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.

Виконання роботи

Для обчислення усіх значень було використано текстовий файл розміром 1.4 МБ з текстом російською мовою.

Обчислення значень для тексту з пробілами

Частота літер:

$H = 4.3760216803315295$

$R = 0.12479566393369412$

м	0.02525591601253788
и	0.05757881565742934
х	0.006924827618239321
а	0.072695207653342
л	0.04421755781597693
	0.15850677070797917
б	0.013031906283999566
у	0.025292510630032638
г	0.0161185215211538
к	0.030777480798381954
о	0.09311781918599496
в	0.03963197074682577
с	0.04264539674899048
т	0.050744067097638665
е	0.06840659998001371
р	0.0399444324808195
ч	0.013383777606064573
ь	0.015226175848396944
п	0.023920212473979116
я	0.01604392480087602
ж	0.007580715762568492
ы	0.014559027821761694
н	0.05395454104015978
ц	0.0027910433266196285
й	0.009860841929549731
з	0.01496297609949232
ш	0.007349888175293848
ф	0.0018170635071436917
д	0.02365138278392145
ю	0.0044448385403251574
щ	0.0029782388699582116
э	0.002585550474533665

[illegible][illegible]

М	0.03001321574857412
И	0.06842457390402595
Х	0.08022921370866564
а	0.0638834529244589
Л	0.05254654021777309
Б	0.1548664425376754
У	0.03005671289578002
Г	0.09154560330008194
К	0.0365748046864429726
О	0.1106578351336956
В	0.047097195042409902
С	0.05067821408919553
Т	0.0630240687774934
Е	0.08129191964808044
Р	0.04746851322193788
О	0.015904795368605103
М	0.018094234532509698
П	0.02842517895755823
Я	0.019066017696154684
Ж	0.009008647364811748
И	0.017301420041146067
Н	0.06411761754272507
С	0.0033167745496512618
Ц	0.01171826651278706
И	0.017781457140848905
З	0.0087343402411789556
О	0.0021593322963185976
Д	0.0281064454480071
Ю	0.052820484734139529
Щ	0.0035392309364912105
Э	0.0030752309364912105

[illegible][illegible]

CoolPinkProgram

Лабораторная работа №1

Произвольная часть текста:
с_вами_чт

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

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:
2,73716774390203< H < 3,39945800008089

Двоичная таблица угаданных символов:
00000000000000000000000000000010
00000100000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000

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

Строка состояния:

Лабораторная работа №1

Произвольная часть текста:
апример_такие_вещи_

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

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:
2,0239154007163< H < 2,63173743342247

Двоичная таблица угаданных символов:
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1]=0,56
q[2]=0,08
q[3]=0,04
q[4]=0,02
q[5]=0
q[6]=0,02
q[7]=0,02
q[8]=0
q[9]=0,02
q[10]=0
q[11]=0
q[12]=0
q[13]=0
q[14]=0
q[15]=0
q[16]=0
q[17]=0,04
q[18]=0
q[19]=0
q[20]=0,02
q[21]=0,02
q[22]=0
q[23]=0,02
q[24]=0
q[25]=0,04
q[26]=0
q[27]=0,02
q[28]=0,02
q[29]=0,04
q[30]=0
q[31]=0,02
q[32]=0

Строка состояния:

Лабораторная работа №1



Произвольная часть текста:

и_зла_уже_в_следующий_момент_

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

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:

 $1,97368676517308 < H < 2,56943203625482$

Двоичная таблица угаданных символов:

00010000000000000000000000000000
00000000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

$q[1] = 0,52$
 $q[2] = 0,12$
 $q[3] = 0,04$
 $q[4] = 0,08$
 $q[5] = 0$
 $q[6] = 0$
 $q[7] = 0,02$
 $q[8] = 0$
 $q[9] = 0,02$
 $q[10] = 0$
 $q[11] = 0$
 $q[12] = 0$
 $q[13] = 0$
 $q[14] = 0,04$
 $q[15] = 0$
 $q[16] = 0$
 $q[17] = 0$
 $q[18] = 0$
 $q[19] = 0$
 $q[20] = 0$
 $q[21] = 0,02$
 $q[22] = 0$
 $q[23] = 0,04$
 $q[24] = 0$
 $q[25] = 0$
 $q[26] = 0$
 $q[27] = 0,04$
 $q[28] = 0$
 $q[29] = 0$
 $q[30] = 0,02$
 $q[31] = 0,02$
 $q[32] = 0,02$

Строка состояния: