

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
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:

студенти гр. ФБ-11

Цема В.В.

Ципун Р.Г.

Перевірила

Селюх П. В.

Київ 2023

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.

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.

Обраний текст - Лев Товстий «Війна і Мир» Том 1 (1.23Мб)

Букви:

З пробілами

H_1 in text: 4.385949914023147

R_1 in text: 0.11470001605958835

Буква, частота

пробіл	0,16095
а	0,07038
б	0,01449
в	0,03862
г	0,0174
д	0,02552
е	0,06683
ж	0,0085
з	0,01494
и	0,05575
й	0,00966
к	0,03009
л	0,04247
м	0,0248
н	0,05467
о	0,09539
п	0,02156
р	0,03824
с	0,04378
т	0,04765
у	0,02406
ф	0,00188
х	0,00716
ц	0,00339
ч	0,01143
ш	0,00792
щ	0,00235
ы	0,01593
ь	0,01678
э	0,00254
ю	0,00544
я	0,01942

Без пробілів

H1 in text: 4.468590618019441

R1 in text: 0.09801906544343475

Буква, частота

а	0,083875
б	0,01727
в	0,046033
г	0,020735
д	0,030412
е	0,079655
ж	0,01013
з	0,017809
и	0,066447
й	0,011515
к	0,035861
л	0,050614
м	0,029559
н	0,065162
о	0,113683
п	0,025692
р	0,045571
с	0,052175
т	0,056794
у	0,028672
ф	0,002239
х	0,008531
ц	0,004046
ч	0,013628
ш	0,009445
щ	0,002805
ы	0,01898
ь	0,020003
э	0,003023
ю	0,006487
я	0,02315

Біграми, що перетинаються:

З пробілами

H2 in text, with interceptions: 3.9784674005199263

R2 in text, with interceptions: 0.19694998920839157

*	а	б	в	г	д	е	ж
а	1.557807106092895e-05	0.0006620680200894805	0.00341159756234344	0.0009487045276105732	0.002112386435861966	0.0008505626799267207	0.001188606821948879
б	0.0011403148016599991	2.4924913697486324e-05	3.738737054622948e-05	1.869368527311474e-05	1.4020263954836056e-05	0.0019690681821014195	4.673421318278685e-06
в	0.005767001906755898	9.34684263655737e-06	2.1809299485300533e-05	2.1809299485300533e-05	0.0002726162435662567	0.004167134008798494	-
г	0.0010281526900213108	-	7.633254819855186e-05	-	0.000861467329669371	0.000563926172405628	-
д	0.004012911105295298	3.427175633404369e-05	0.0009128749641704366	9.34684263655737e-06	2.0251492379207637e-05	0.004263718049376254	1.2462456848743162e-05
е	9.34684263655737e-06	0.0010421729539761468	0.0013147891975424035	0.00333059159282661	0.002406811978913523	0.0014394137660298351	0.0009284530352313656
ж	0.0013350406899216112	4.984982739497265e-05	-	2.336710659139343e-05	0.000623122842437158	0.0036966762627584403	1.2462456848743162e-05

Повний файл *lab1_bigram_with_interceptions_in_text_with_spaces.csv*

Без пробілів

H2 in text, with interceptions: 4.150657749319352

R2 in text, with interceptions: 0.16219352458497438

*	а	б	в	г	д	е	ж
а	0.0003861792391526336	0.001583706206717291	0.005820538051651473	0.0017192402666122055	0.003267670485137669	0.0015317205399082824	0.001533577170865747
б	0.0013683370156513989	3.527598819182711e-05	5.941219063886671e-05	2.2279571489575018e-05	2.042294053211043e-05	0.002359777946937487	9.283154787322924e-06
в	0.006986502292939232	0.0002487885483002544	0.0004641577393661462	0.0004121720725571378	0.0007073763947940067	0.005126158073559719	2.9706095319433356e-05
г	0.0012402294795863427	4.270251202168545e-05	0.00016524015521434803	1.2996416702252093e-05	0.0010991255268190342	0.0006980932400066839	-
д	0.0047993910250459515	9.468817883069382e-05	0.0011752473960750822	9.283154787322924e-05	5.5698928723937545e-05	0.005111305025900002	2.2279571489575018e-05
е	0.0001949462505337814	0.0021963944226806036	0.0034421937951393403	0.00454032583873304	0.0037244017006739572	0.002047863946083437	0.0013367742893745011
ж	0.0015911327305471491	6.683871446872505e-05	2.9706095319433356e-05	3.341935723436253e-05	0.000755648799688086	0.004418781678765712	1.4853047659716678e-05

Повний файл *lab1_bigram_with_interceptions_in_text_without_spaces.csv*

Біграми, що не перетинаються:

Без пробілів

H2 in text, without interceptions: 4.1497924670210695

R2 in text, without interceptions: 0.1623681810265184

*	а	б	в	г	д	е	ж
а	0.0004530179536213587	0.0015892760995896845	0.005710996825161063	0.001667254599803197	0.003197118508754015	0.001552143480440393	0.0015447169566105345
б	0.0013961864800133678	2.9706095319433356e-05	7.055197638365422e-05	1.8566309574645847e-05	7.426523829858339e-06	0.00245817938768311	7.426523829858339e-06
в	0.007062624162195281	0.0002487885483002544	0.00047901078702586287	0.0004270251202168545	0.0006721004066021797	0.005020330108984237	2.5992833404504185e-05
г	0.0012847886225654927	4.084588106422086e-05	0.0001559570004270251	1.1139785744787509e-05	0.0011325448840533968	0.0006758136685171089	-
д	0.004764115036854124	0.00010025807170308757	0.001180817288947476	0.00010397133361801674	7.426523829858339e-05	0.005168860585581404	2.2279571489575018e-05
е	0.0001819498338315293	0.0022242438870425725	0.0034533335808841275	0.0046341508698316035	0.0035610181764170737	0.0020311542674662556	0.0012736488368207052
ж	0.0017155270046972763	6.683871446872505e-05	2.2279571489575018e-05	2.2279571489575018e-05	0.0007760717402201965	0.00438164905961642	1.8566309574645847e-05

Повний файл *lab1_bigram_without_interceptions_in_text_without_spaces.csv*

З пробілами

H2 in text, without interceptions: 3.9784288918899122

R2 in text, without interceptions: 0.19695776214018546

*	а	б	в	г	д	е	ж
а	1.869368527311474e-05	0.0007321693398636607	0.0034988347602846427	0.0008848344362607644	0.002078114679527922	0.0008661407509876497	0.001218205156964644
б	0.0011808177864184146	2.1809299485300533e-05	2.8040527909672112e-05	6.231228424371581e-06	1.869368527311474e-05	0.0019784150247379767	6.231228424371581e-06
в	0.005611221196146608	3.1156142121857905e-06	2.4924913697486324e-05	2.1809299485300533e-05	0.0002804052790967211	0.004181154272753331	-
г	0.0010343839184456824	-	6.542789845590159e-05	-	0.000831868994653606	0.0006324696850737154	-
д	0.0040908014605999425	3.11561421218579e-05	0.0009222218068069939	6.231228424371581e-06	1.557807106092895e-05	0.0043026632270285765	1.869368527311474e-05
е	3.1156142121857905e-06	0.001025037075809125	0.0012649393701474308	0.0033087822933413093	0.0023554043444124575	0.001473685522363879	0.0009471467205044802
ж	0.0012431300706621303	4.673421318278686e-05	-	3.427175633404369e-05	0.0006418165277102729	0.003623459328772074	1.2462456848743162e-05

Повний файл *lab1_bigram_without_interceptions_in_text_with_spaces.csv*

2. За допомогою програми CoolPinkProgram оцінити значення H10, H20, H30.

H10:

Произвольная часть текста:
ак_же_как_их_не_могут_нарушит

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: 52

Неравенство для энтропии:
 $1,60199746514508 < H < 2,406148748468$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000
00010000000000000000000000000000
00100000000000000000000000000000
10000000000000000000000000000000

Вероятности:

q[1] = 0,549019607943137
q[2] = 0,0980392156862745
q[3] = 0,0980392156862745
q[4] = 0,0586235294117647
q[5] = 0,0196079431372549
q[6] = 0,0196079431372549
q[7] = 0,0196079431372549
q[8] = 0
q[9] = 0,0392156862745098
q[10] = 0
q[11] = 0,019607943137254
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0,019607943137254
q[16] = 0
q[17] = 0,0392156862745098
q[18] = 0
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0,019607943137254
q[32] = 0

Строка состояния:

$$1,60199746514508 < H(30) < 2,406148748468$$

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

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

$$H(0) = \log_2(32) = 5$$

H(10):

$$0,709001541 < R < 0,547588205$$

$$71\% < R < 54,7\%$$

H(20):

$$0,597605532 < R < 0,450416012$$

$$60\% < R < 45\%$$

H(30):

$$0,679600507 < R < 0,51877025$$

$$68\% < R < 52\%$$

Висновки:

Ми засвоїли теоретичні поняття ентропії на символ джерела та його надлишковості. Навчилися на практиці визначати частоти монограм та біграм відкритого тексту, а також розраховувати відповідні значення ентропій (H_1 , H_2). Попрацювавши з CoolPinkProgram, ми отримали значення ентропій для H_{10} , H_{20} , H_{30} . У підсумку, ми набули практичних навичок оцінки ентропії.