

Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Фізико-технічний інститут

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1
з дисципліни «Криптографія»

Виконав:

студент 3 курсу

НН ФТІ групи ФБ-25

Черняк Денис

Тема: «Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Мета: 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.

Хід роботи

Завдання 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.

Виконання

Частоти біграм з перетином:

Lab1.py bigram_matrix_overlap_with_spaces.csv									
	Перша літера	а	б	в	г	д	е		
1	а	1.2356075482314655e-05	0.0006985934984231747	0.002823838481381295	0.0011633720300271643	0.0019959814240662135	0.0011291552056146006		
2	б	0.0005731318089104413	8.554206103140915e-05	4.6572899894878315e-05	9.50467344793435e-06	1.2356075482314655e-05	0.0016490608432166095		
3	в	0.005610608736315646	1.0455140792727785e-05	5.98794427219864e-05	0.00011025421199603845	0.00032696076660894164	0.004433930163461374		
4	г	0.004319874082086162	1.9009346895868698e-06	0.00014827290578777586	0.0002879916054724108	0.0010322075364456703	0.0014038402682599034		
5	д	0.00470671429141709	4.6572899894878315e-05	0.0007945907002473116	2.946448768859648e-05	0.00012070935278876623	0.004403515208427984		
6	е	7.223551820430106e-05	0.0015986860739425576	0.001549261772813299	0.0034739581452200045	0.0024769179005316912	0.0006482187291491226		
7	ж	0.0012508150257481603	1.140560813752122e-05	2.851402034380305e-06	2.091028158545557e-05	0.0006434663924251554	0.0029797151259274186		
8	з	0.005811157346067061	0.00012070935278876623	0.0007299589208013581	0.00048188694381027153	0.000817401916522354	0.00047998600912068463		
9	и	0.00017868786082116578	0.0004790355417758912	0.0021223935809237403	0.0005085000294644877	0.0017830767388324839	0.0016310019636655343		
10	й	9.504673447934349e-07	9.504673447934349e-07	1.0455140792727785e-05	3.421682441256366e-05	0.0001929448709930673	4.752336723967175e-06		
11	к	0.006525908789351724	1.9009346895868698e-06	0.00017488599144199204	7.413645289388793e-05	9.504673447934349e-07	0.0004847383458446518		
12	л	0.0067359620725510735	5.417663865322579e-05	3.3266357067770225e-05	0.00012546168951273342	0.0006225561108396999	0.0032173319621257774		
13	м	0.0028105319385541873	0.0006102000353573852	3.0414955033389916e-05	0.00011785795075438593	7.508692023868136e-05	0.0036849618957641474		
14	н	0.010974095962985	2.851402034380305e-06	8.554206103140915e-06	0.00015967851392529708	0.000869677620485993	0.008722438823169353		
15	о	5.70280406876061e-06	0.0035347880552867843	0.005626766681177135	0.004218174076193264	0.00411742453764516	0.001411444007018251		
16	п	0.002163263676749858	0.0	0.0	6.653271413554044e-06	0.0	0.0018515103876576113		
17	р	0.007079080784021504	4.562243255008480e-05	0.00040775049091638357	0.00011500654872000562	0.00045337292346646845	0.004853086262515279		
18	с	0.001328753348021222	7.413645289388793e-05	0.0011700253014407184	9.124486510016976e-05	0.00024046823823273904	0.0026346954797674014		
19	т	0.004987102158131153	3.611775910215053e-05	0.0022041337725759755	4.2771030515704575e-05	0.00011405608137521219	0.004791305885103706		
20	у	2.6613085654216176e-05	0.00047238227036233715	0.0008972411734850025	0.0009580710835517824	0.0015464103699789186	0.0001843906648899264		
21	ф	0.00011120467934083188	0.0	0.0	0.0	0.0	0.0002423691729223259		
22	х	0.0004514719887768816	1.9009346895868698e-06	0.00038208787260696084	1.3306542827108088e-05	4.752336723967175e-06	5.607757334281266e-05		
23	ц	0.00042295796843307853	0.0	4.8473834584465185e-05	1.9009346895868698e-06	2.851402034380305e-06	0.0006624757393210241		
24	ч	0.0016186458881832196	0.0	9.504673447934349e-07	2.851402034380305e-06	0.0	0.0022516571398156476		
25	ш	0.0009542692141726087	0.0	1.8058879551075265e-05	9.504673447934349e-07	0.0	0.002158511340025891		

Lab1.py

bigram_matrix_overlap_without_spaces.csv

Частоти біграм без перетину:

Lab1.py

bigram_matrix_non_overlap_with_spaces.csv

CSV

	Перша літера	а	б	в	г	д	е	ж
1	а	1.140560813752122e-05	0.0006824355535616863	0.0027677609080384827	0.0011576692259584037	0.0019826748812391054	0.001193786980605542	
2	б	0.0005664785374968872	8.554206103140915e-05	3.991962848132427e-05	9.50467344793435e-06	9.50467344793435e-06	0.0017146430900073567	
3	в	0.005655280701520938	1.140560813752122e-05	7.793832227306167e-05	9.694766916893037e-05	0.0003421682441256366	0.004360744177912279	
4	г	0.004309418941293434	1.9009346895868698e-06	0.00014447103640800212	0.0002946448768859648	0.000994188842653933	0.0014485122334651948	
5	д	0.004765643266794283	4.752336723967175e-05	0.0008059963083848329	3.611775910215053e-05	0.00014066916702942838	0.004404465675772778	
6	е	8.744299572099602e-05	0.0015682711189091677	0.001562568314840407	0.0034406917881522345	0.002509233790254668	0.00065772302597057	
7	ж	0.0012261028747835312	7.603738758347479e-06	1.9009346895868698e-06	1.900934689586867e-05	0.0006843364882512732	0.0029502506382388222	
8	з	0.005814959215446235	0.00011785795075438593	0.0007679776145930955	0.0004714318030175437	0.0007546710717659873	0.0004904411499134124	
9	и	0.00015587664454612334	0.00046953086832795686	0.002039702921926711	0.0004980448886717599	0.001792581422804183	0.0016576150493197506	
10	й	0.0	0.0	1.140560813752122e-05	2.281121627504244e-05	0.000190093468958687	3.8018693791737395e-06	
11	к	0.006463177944595357	3.8018693791737395e-06	0.0001653813179940577	8.174019165223541e-05	1.9009346895868698e-06	0.00048663298053423866	
12	л	0.00679584151527306	5.702804068760609e-05	3.80186937917374e-05	0.00013116449358149403	0.0006406149903907751	0.0032962207517436325	
13	м	0.002803878667140633	0.0006216056434949065	2.6613085654216170e-05	0.00012165982013355967	7.793832227306167e-05	0.003590865628629597	
14	н	0.010924671661055742	1.9009346895868698e-06	9.50467344793435e-06	0.00017488599144199204	0.0008687271531411995	0.008750002376168362	
15	о	7.603738758347479e-06	0.003520531045114883	0.005653379766831351	0.004286607725018391	0.004054693692888793	0.0013382580214691565	
16	п	0.0020967309626143175	0.0	0.0	9.50467344793435e-06	0.0	0.001826798236692982	
17	р	0.007094288261538198	5.512710599801922e-05	0.0003668803950902659	0.0001785795075438593	0.0004923420846029993	0.0048093647646547805	
18	с	0.0012812299807815502	7.033458351471419e-05	0.001170975768785512	0.00010455140792727785	0.00024331964026711933	0.002598577720665251	
19	т	0.005016566645819749	3.421682441256366e-05	0.002279220692814657	4.562243255008488e-05	0.00012546168951273342	0.004856888131894452	
20	у	3.80186937917374e-05	0.0004942430192925862	0.0008744299572099601	0.0010150991242393885	0.00151416580332549515	0.0001520747751669496	
21	ф	0.00011405608137521219	0.0	0.0	0.0	0.0	0.00021670655461290315	
22	х	0.0004828374111550649	0.0	0.00037638506853820026	1.3306542827108088e-05	1.9009346895868698e-06	5.512710599801922e-05	
23	ц	0.0004277103051570457	0.0	4.182056317091114e-05	0.0	1.9009346895868698e-06	0.000627308447563667	
24	ч	0.0015644692495299939	0.0	1.9009346895868698e-06	3.8018693791737395e-06	0.0	0.002178471154266553	
25	ш	0.0009561701488621956	0.0	1.900934689586867e-05	1.9009346895868698e-06	0.0	0.002144254329853989	

Lab1.pybigram_matrix_non_overlap_without_spaces.csv

	Перша літера	а	б	в	г	д	е	ж
1	а	0.0002300141391044332	0.001468031417225353	0.004927263666109672	0.002667713005495534	0.0032179429068826095	0.0018130526258820028	0.0002300141391044332
2	б	0.000703572660790031	0.00012177219129058228	6.990625796311204e-05	1.578528405618659e-05	2.48054463740075e-05	0.0019731605070233238	0.000703572660790031
3	в	0.006920719538348092	0.0001894234086742391	0.0004329677912554036	0.0005637601448638068	0.0006584718492009264	0.005518084297926941	0.006920719538348092
4	г	0.005175318129849746	6.314113622474637e-05	0.00023677926084279886	0.0003517863303950155	0.0013079235360840319	0.0016935354751708757	0.005175318129849746
5	д	0.005587990555890053	9.245666375766432e-05	0.001062124112923412	0.00014432259708513454	0.0001646179623002316	0.005233949184915583	0.005587990555890053
6	е	0.00025030950431953025	0.0024647593533445636	0.003364520544547199	0.004776175947286171	0.0036576758198763786	0.001071144275241233	0.00025030950431953025
7	ж	0.0015153872693939128	3.157056811237318e-05	3.382560869182841e-05	3.833568985073886e-05	0.000773478918753143	0.003549433872062528	0.0015153872693939128
8	з	0.0068485582398055255	0.0001961885304126048	0.001012513220175397	0.0006674920115187473	0.0010260434636521283	0.0005727803071816277	0.0068485582398055255
9	и	0.0004713034811061425	0.001231252156382554	0.0047378402574354325	0.0016664749882174129	0.003073620309797475	0.002521135367830944	0.0004713034811061425
10	й	8.794658259875387e-05	0.0002300141391044332	0.0006697470520982025	0.0004036522637224857	0.0005524849419665307	0.00011275202897276136	8.794658259875387e-05
11	к	0.00784303113534528	0.0002300141391044332	0.0005569950231254411	0.00029766535648809	0.00021197381446879136	0.0007080827419489414	0.00784303113534528
12	л	0.008113636004879908	0.0003044304782264557	0.000983197692642479	0.0012876281708689347	0.00109143964045633	0.004239476289375827	0.008113636004879908
13	м	0.003393836072080117	0.0009110363940999118	0.0007915192433887848	0.0005434647796487098	0.0004938538869006947	0.004496550915433723	0.003393836072080117
14	н	0.013279933972411834	0.0002232490173660675	0.0005412097390692545	0.0002954103159086348	0.0012041916694290914	0.010298770326372024	0.013279933972411834
15	о	0.00013079235360840318	0.00515953284579356	0.008742792326547916	0.0057120177877600905	0.005793199248620479	0.0022370002548195854	0.00013079235360840318
16	п	0.002640652518542071	0.0	0.0	9.020162317820909e-06	2.255040579455227e-06	0.0022257250519223093	0.002640652518542071
17	р	0.008478952578751654	9.471170433711955e-05	0.000644941605724195	0.00018716836809478386	0.0006133710376118218	0.005763883721087561	0.008478952578751654
18	с	0.0016551997853201368	0.0002300141391044332	0.0016957905157503309	0.0003811018579279334	0.0004938538869006947	0.003242748353256617	0.0016551997853201368
19	т	0.005894676074695964	0.0002886451941702691	0.0030871505532742063	0.0002074637333098809	0.0003901220202457543	0.005842810141368494	0.005894676074695964
20	у	0.0001240272318700375	0.0007216129854256728	0.0017611866925545325	0.00158754856793648	0.002200919605548302	0.0003157058611237318	0.0001240272318700375
21	ф	0.00013079235360840318	2.255040579455227e-06	6.765121738365682e-06	0.0	0.0	0.0002954103159086348	0.00013079235360840318
22	х	0.0005682702260227173	8.794658259875387e-05	0.0007802440404915086	0.0001894234086742391	0.00019844357099206	0.0001240272318700375	0.0005682702260227173
23	ц	0.0005276794955925232	1.3530243476731363e-05	7.44163391220225e-05	2.0295365215097047e-05	1.8040324635641818e-05	0.0007464184317996802	0.0005276794955925232
24	ч	0.0018987441679013013	2.255040579455227e-06	2.48054463740075e-05	4.7355852168559774e-05	4.510081158910454e-06	0.002699283573607907	0.0018987441679013013
25	ш	0.0011523257361016211	9.020162317820909e-06	1.3530243476731363e-05	2.255040579455227e-06	9.020162317820909e-06	0.002559471057681683	0.0011523257361016211

Частоти букв:

Lab1.pyletter_frequencies_with_spaces.csv

	Літера	Частота
1		0.1570284617175879
2	о	0.09288623391929589
3	а	0.07072610883791221
4	е	0.06176130936256968
5	и	0.05827880032125766
6	н	0.05418894322388712
7	т	0.048463333380856656
8	л	0.045465562224661756
9	р	0.045352456718134426
10	с	0.043892540264134625
11	в	0.035166307865585035
12	м	0.028336256017640657
13	к	0.027700393968339964
14	д	0.025730077035305075
15	у	0.02461422943309429
16	п	0.023390028656563207
17	г	0.020132780161864434
18	я	0.01671205143924381
19	ь	0.01553822538410725
20	з	0.015374745156185398
21	ы	0.01479401016048626
22	б	0.013583115914134861
23	ч	0.011703093293033557
24	й	0.008225336583928563
25	ж	0.0081844665269481
26	ш	0.007447855034858356
27	х	0.006898485431725619

Lab1.py letter_frequencies_without_spaces.csv

	Літера	Частота
1	о	0.11018905111381966
2	а	0.08390094519919315
3	е	0.07326618581738928
4	и	0.06913495613387259
5	н	0.06428324179758102
6	т	0.05749106723057651
7	л	0.05393487224645762
8	р	0.0538006974832648
9	с	0.05206882827096086
10	в	0.041717076162782175
11	м	0.03361472449636544
12	к	0.03286041427303775
13	д	0.030523067347838488
14	у	0.02919936002020514
15	п	0.027747115524471108
16	г	0.023883107848321632
17	я	0.0198251669010027
18	ь	0.018432680913245306
19	з	0.018238747642075854
20	ы	0.01754983352181693
21	б	0.016113374292340877
22	ч	0.013883141673892185
23	й	0.009757549585467633
24	ж	0.00970906626767527
25	ш	0.008835239028394311
26	х	0.008183533035743481
27	ё	0.005546266051642626
28	ю	0.005544011013605772
29	щ	0.0035043291092712505

Ентропія та надлишковість:

З пробілами:

Тексти	H	R
Текст з пробілами монограми	4.406696524751244	0.12641708389912865
Текст з пробілами біграми(з перетином)	4.0188120068200375	0.210055752302588
Текст з пробілами біграми(без перетину)	4.018162669281308	0.2101833871490707

Без пробілів:

Тексти	H	R
Текст без пробілів монограми	4.48359558232663	0.1111726252474392
Текст без пробілів біграми(з перетином)	4.178746528108722	0.17160586004327194
Текст без пробілів біграми(без перетину)	4.178337155941605	0.1716870139256671

Результат у програмі:

```
Ентропія та надлишковість для тексту з пробілами:  
H1 (ентропія букв): 4.406696524751244, R: 0.13381253834018836  
H2 (ентропія біграм з перетином): 4.0188120068200375, R: 0.210055752302588  
H2 (ентропія біграм без перетину): 4.018162669281308, R: 0.2101833871490707  
  
Ентропія та надлишковість для тексту без пробілів:  
H1 (ентропія букв): 4.48359558232663, R: 0.1111726252474392  
H2 (ентропія біграм з перетином): 4.178746528108722, R: 0.17160586004327194  
H2 (ентропія біграм без перетину): 4.178337155941605, R: 0.1716870139256671
```

Завдання 2

За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}, H^{(20)}, H^{(30)}$.

Під час виконання були проблеми з ієрогліфами, і на жаль, повністю прибрати її не вдалось, але залишок символів не заважав виконання роботи, оскільки одна кнопка відповідала за продовження дії, а інша до перехід до наступного значення

Виконання

$$H^{(10)}$$

$$2,1984629707794 < H < 2,92598127152669$$

Лабораторная работа №1

Произвольная часть текста:
едставления_о_том_что_мы_называем_добром_то_хотя_нам_и_пришлось_бы_воевать_

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: и

Символ по счету: 1

Номер эксперимента: 53

Неравенство для энтропии:
 $2,1984629707794 < H < 2,92598127152669$

Двоичная таблица угаданных символов:
00001000000000000000000000000000
00000100000000000000000000000000
00000000000010000000000000000000
10000000000000000000000000000000
00000000000000001000000000000000

Поле ввода символов:
и

Продолжить Другой

Вероятности:
q[1] = 0,4716981
q[2] = 0,1320754
q[3] = 0
q[4] = 0,0566037
q[5] = 0,0377358
q[6] = 0,0377358
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0,018867
q[12] = 0
q[13] = 0,037735
q[14] = 0,018867
q[15] = 0,018867
q[16] = 0,018867
q[17] = 0,018867
q[18] = 0,037735
q[19] = 0
q[20] = 0
q[21] = 0,018867
q[22] = 0
q[23] = 0,018867
q[24] = 0
q[25] = 0,018867
q[26] = 0,018867
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0,018867
q[31] = 0
q[32] = 0

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

$H^{(20)}$

$$2,62138549320994 < H < 3,24023073433794$$

[illegible]

$H^{(30)}$

$$2,33839612073776 < H < 2,92891056032524$$

Лабораторная работа №1

Произвольная часть текста:
ди_для_этого_может_быть_сколько_угодно_пояснений_и_извинений_например_вы_ст

Использованные буквы:

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

Введенный символ: **к**

Символ по счету: **1**

Номер эксперимента: **51**

Неравенство для энтропии:
 $2,33839612073776 < H < 2,92891056032524$

Двоичная таблица угаданных символов:

00000100000000000000000000000000
10000000000000000000000000000000
00000000000000000000000000000000
01000000000000000000000000000000
00010000000000000000000000000000

Поле ввода символов:
к

Продолжить Другой

Вероятности:

q[1] = 0,3529411
q[2] = 0,2549019
q[3] = 0,0392156
q[4] = 0,0196078
q[5] = 0
q[6] = 0,0588235
q[7] = 0
q[8] = 0,0784313
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0,039215
q[14] = 0,019607
q[15] = 0,019607
q[16] = 0,019607
q[17] = 0
q[18] = 0,019607
q[19] = 0,019607
q[20] = 0,019607
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0,019607
q[27] = 0
q[28] = 0
q[29] = 0,019607
q[30] = 0
q[31] = 0
q[32] = 0

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

Завдання 3

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

Виконання

Для обрахунків використаємо цю формулу

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

$H_{(0)}$ можна знайти за наступною формулою, 32 це кількість літер у нашому алфавіті, це вказано у методичці, що у наданому тексті використовується лише 32 символи

$$H_{(0)} = \log_2 32 = 5$$

$$H^{(10)}$$

$$R = 1 - \frac{2,1984629707794}{5} \approx 0,56030740584412$$

$$R = 1 - \frac{2,92598127152669}{5} \approx 0,414803745694662$$

$$H^{(20)}$$

$$R = 1 - \frac{2,62138549320994}{5} \approx 0,475722901358012$$

$$R = 1 - \frac{3,24023073433794}{5} \approx 0,351953853132412$$

$H^{(30)}$

$$R = 1 - \frac{2,33839612073776}{5} \approx 0,532320775852448$$

$$R = 1 - \frac{2,92891056032524}{5} \approx 0,414217887934952$$

Висновки:

У ході виконання роботи я дослідив та оцінив ентропію для символів джерела відкритого тексту російською мовою. Було розроблено програму для підрахунку частот символів та біграм у тексті, а також обчислено ентропії H_1 та H_2 (з перетином та без) як з урахуванням пробілів, так і без них. Це дало можливість більш детально вивчити вплив пробілів на загальну ентропію тексту. Зокрема, результати показали, що включення пробілів впливає на розподіл частот окремих символів та біграм, проте не призводить до суттєвих змін значень H_1 та H_2 , що свідчить про низький внесок пробілів у загальну ентропію при обробці тексту значного обсягу.

Додатково, за допомогою програми CoolPinkProgram було обчислено ентропії H_{10} , H_{20} , H_{30} для тексту, що включає лише літери та пробіли, без урахування розділових знаків. Це дозволило оцінити надлишковість російської мови, яка, за нашими підрахунками, варіюється від 47% до 65% залежно від значення ентропії H_{NN} . Високий рівень надлишковості свідчить про значну передбачуваність структури російської мови, що можна враховувати при проєктуванні систем стиснення тексту або криптографічного аналізу.

Результати також вказали на помітну різницю у значеннях H між текстом із пробілами та без них, а також між різними завданнями, що підтверджує залежність ентропії від обраного тексту та характеру символів, які він містить.