

НТУУ "КПІ ім Ігоря Сікорського"
Фізико-технічний інститут

КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1
Експериментальна оцінка ентропії на символ джерела
відкритого тексту

Виконали:
студенти групи ФБ-14
Сергєєв Олег
Деркач Семен
Перевірила:
Селюх П.В.

Експериментальна оцінка ентропії на символ джерела
відкритого тексту

Варіант № 3

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

Частоти для монограм, без пробілів:

```
1 а ~ 0.07822784712620215
2 б ~ 0.02208340921708328
3 в ~ 0.04577069370494087
4 г ~ 0.017297093271812017
5 д ~ 0.030778999192835325
6 е ~ 0.08640953675790507
7 ж ~ 0.011073066835105666
8 з ~ 0.01614604453546215
9 и ~ 0.06732754764444919
10 й ~ 0.009965485044344538
11 к ~ 0.04191424026656792
12 л ~ 0.04452830936140454
13 м ~ 0.029992742670730183
14 н ~ 0.06301276552664056
15 о ~ 0.11294748257608228
16 п ~ 0.029731390782696633
17 р ~ 0.04145701201613449
18 с ~ 0.050144348774369685
19 т ~ 0.06188097429783001
20 у ~ 0.02808624942433808
21 ф ~ 0.001060813558165651
22 х ~ 0.009348144374084237
23 ц ~ 0.0034393908465215166
24 ч ~ 0.01933343714046921
25 ш ~ 0.009278817346942705
26 щ ~ 0.003066345414759944
27 ы ~ 0.018758462986795402
28 ь ~ 0.02029411165943043
29 э ~ 0.0025320871341692557
30 ю ~ 0.0057200299536774414
31 я ~ 0.018393120558049557
```

```
*** Ентропія монограми без пробілів ***
4.463308520322652
```

```
*** Надлишковість монограми без пробілів ***
0.09908525203876895
```

Частоти для монограм, з пробілами:

```
1 ~ 0.2173327941785059
2 а ~ 0.06122637052769563
3 б ~ 0.0172839601869472
4 в ~ 0.035823220950557524
5 г ~ 0.013537867659882876
6 д ~ 0.02408971329623845
7 е ~ 0.06762991069063924
8 ж ~ 0.008666526279706807
9 з ~ 0.012636979561639564
10 и ~ 0.052695063589694564
11 й ~ 0.007799658334313028
12 к ~ 0.03280490131356547
13 л ~ 0.03485084746784557
14 м ~ 0.02347433610102349
15 н ~ 0.04931802512582074
16 о ~ 0.08840029059239421
17 п ~ 0.0232697845490801
18 р ~ 0.03244704375637609
19 с ~ 0.039246337342974384
20 т ~ 0.048432209247194304
21 у ~ 0.02198218635895223
22 ф ~ 0.000830263983467067
23 х ~ 0.00731648603688043
24 ц ~ 0.0026918984235750187
25 ч ~ 0.015131647225656535
26 ш ~ 0.007262226046259657
27 щ ~ 0.002399927997853716
28 ы ~ 0.014681633811381076
29 ь ~ 0.015883535667115818
30 э ~ 0.001981781562196806
31 ю ~ 0.004476879861059973
32 я ~ 0.014395692273506526
```

```
*** Ентропія монограми з пробілами ***
4.248552479753668
```

```
*** Надлишковість монограми з пробілами ***
0.15028950404926644
```

Частоти для біграм, без пробілів(перетинаються):

```
1 зл ~ 0.0002090815
2 ал ~ 0.0084072776
3 оу ~ 0.0008198196
4 ля ~ 0.0017584855
5 лм ~ 0.0001892738
6 ьи ~ 0.0009546222
7 фо ~ 9.13356e-05
8 вф ~ 1.59562e-05
9 мл ~ 0.0003268274
10 шэ ~ 4.4017e-06
11 сс ~ 0.0010360539
12 ож ~ 0.0030118742
13 еу ~ 0.0007048248
14 ип ~ 0.0027890373
15 кж ~ 0.0001903742
16 жф ~ 1.6506e-06
17 лб ~ 0.0003130721
18 сц ~ 3.57639e-05
19 зю ~ 8.2532e-06
20 ро ~ 0.0085283248
21 бв ~ 0.0001760686
22 за ~ 0.0055967819
23 лв ~ 0.0007097767
24 жд ~ 0.0009359149
25 ос ~ 0.0090812903
26 ся ~ 0.0046647185
27 вг ~ 0.0004176128
28 дк ~ 0.0004407218
29 цц ~ 2.64103e-05
30 уф ~ 3.24627e-05
31 хн ~ 0.0005139003
32 хм ~ 0.0002305399
33 шж ~ 2.7511e-06
34 ек ~ 0.0030476381
35 ьи ~ 0.0004731845
36 ин ~ 0.0057530428
37 шр ~ 7.703e-06
38 хю ~ 3.3013e-06
39 да ~ 0.0064193526
40 цм ~ 0.000212933
41 ох ~ 0.0008627363
42 ре ~ 0.006679604
43 уа ~ 0.0001276498
44 ен ~ 0.0107511914
45 ез ~ 0.0020616537
46 зе ~ 0.0004445733
47 йж ~ 0.0001138944
48 юж ~ 0.0001039905
49 сч ~ 0.0004220145
50 сх ~ 0.0002547493
51 ве ~ 0.006323065
52 ды ~ 0.0005579175
53 мн ~ 0.0023422631
```

```
*** Ентропія біграми без пробілів з перетином ***
4.134798718566226
```

```
*** Надлишковість біграми без пробілів з перетином ***
0.16539465545657028
```

Всього 874 біграми*

Частоти для біграм, без пробілів(не перетинаються):

```
1 |л ~ 0.0001980771
2 ал ~ 0.0085206171
3 оу ~ 0.0008242209
4 ля ~ 0.0017155679
5 лм ~ 0.0001903741
6 ьи ~ 0.0009958877
7 фо ~ 9.02351e-05
8 вф ~ 1.65064e-05
9 мл ~ 0.0003411328
10 шэ ~ 5.5021e-06
11 сс ~ 0.0010828215
12 ож ~ 0.0030107721
13 еу ~ 0.0006415498
14 ип ~ 0.0027334641
15 кж ~ 0.0001969767
16 жф ~ 1.1004e-06
17 лб ~ 0.0003158229
18 сц ~ 3.63141e-05
19 зю ~ 8.8034e-06
20 ро ~ 0.0085613329
21 бв ~ 0.0001815707
22 за ~ 0.0055846741
23 лв ~ 0.0006800647
24 жд ~ 0.000952971
25 ос ~ 0.0091258527
26 ся ~ 0.00464931
27 вг ~ 0.0003994555
28 дк ~ 0.0004148615
29 цч ~ 2.86111e-05
30 уф ~ 3.0812e-05
31 хн ~ 0.0004610795
32 хм ~ 0.0002178848
33 шж ~ 2.2009e-06
34 ек ~ 0.0030272785
35 ьи ~ 0.0004621799
36 ин ~ 0.0057266294
37 шр ~ 7.703e-06
38 хю ~ 3.3013e-06
39 да ~ 0.0064826237
40 цм ~ 0.000223387
41 ох ~ 0.000875941
42 ре ~ 0.0067126132
43 уа ~ 0.0001265493
44 ен ~ 0.010665352
45 ез ~ 0.0020280895
46 зе ~ 0.0004500752
47 йж ~ 0.0001089424
48 юж ~ 0.0001133441
49 сч ~ 0.0004247654
50 сх ~ 0.0002475964
51 ве ~ 0.0062834461
52 ды ~ 0.0005766245
53 мн ~ 0.0023670215
```

```
*** Ентропія біграми без пробілів без перетину ***
4.134046283477035
```

```
*** Надлишковість біграми без пробілів без перетину ***
0.16554653379203577
```

Частоти для біграм, з пробілами(перетинаються):

1	зл	~	0.0001489997
2	ал	~	0.0063665056
3	оу	~	7.6653e-05
4	ля	~	0.0013483177
5	лм	~	3.8757e-06
6	о	~	0.0212694857
7	ьи	~	8.6127e-05
8	фо	~	7.01935e-05
9	мл	~	0.0001589043
10	сс	~	0.0006252818
11	ож	~	0.0018599119
12	еу	~	9.47397e-05
13	ип	~	0.0002110111
14	кж	~	5.16762e-05
15	лб	~	4.65086e-05
16	сц	~	1.93786e-05
17	р	~	0.0038688235
18	зю	~	5.5983e-06
19	ро	~	0.006631346
20	бв	~	5.46906e-05
21	за	~	0.0043696518
22	лв	~	1.72254e-05
23	жд	~	0.0006877238
24	ос	~	0.0053489154
25	ся	~	0.0036384338
26	вг	~	9.474e-06
27	дк	~	0.0002295284
28	уф	~	6.0289e-06
29	хн	~	0.0001235922
30	хм	~	1.7656e-05
31	ек	~	0.0013681269
32	ьи	~	2.2393e-05
33	ин	~	0.0028305628
34	шр	~	4.306e-07
35	да	~	0.0050108671
36	цм	~	1.72254e-05
37	у	~	0.0036595349
38	ох	~	0.0004732677
39	ре	~	0.0052106816
40	уа	~	1.67948e-05
41	ен	~	0.0065107682
42	ез	~	0.0010107
43	зе	~	0.0003358952
44	юж	~	3.1867e-05
45	сч	~	0.0002657017
46	сх	~	0.0001774216
47	ве	~	0.0048571304
48	ды	~	0.0004366637
49	мн	~	0.0011480725
50	ты	~	0.0011984568
51	га	~	0.0006416459

*** Ентропія біграми з пробілами з перетином ***
3.8776470172305193

*** Надлишковість біграми з пробілами з перетином ***
0.2244705965538959

Частоти для біграм, без пробілів(не перетинаються):

```
1 |ны ~ 0.0031918641
2 чу ~ 0.0006028885
3 лб ~ 3.96184e-05
4 лл ~ 9.21558e-05
5 вн ~ 0.0011179276
6 нд ~ 0.0002230688
7 шу ~ 0.0002712998
8 ли ~ 0.0053992974
9 ом ~ 0.0047550679
10 гу ~ 0.0006666225
11 ыз ~ 3.01444e-05
12 юк ~ 1.63641e-05
13 рф ~ 1.72254e-05
14 ью ~ 0.0002721611
15 би ~ 0.0005003975
16 жж ~ 1.63641e-05
17 ыч ~ 0.0001016298
18 бж ~ 1.55028e-05
19 иг ~ 0.0004392474
20 зл ~ 0.0001610574
21 ев ~ 0.0016019609
22 кф ~ 2.5838e-06
23 сз ~ 1.20578e-05
24 фт ~ 3.4451e-06
25 ам ~ 0.0031350203
26 тк ~ 0.0004366635
27 ын ~ 0.0001507221
28 фе ~ 0.0001421094
29 ши ~ 0.0020954683
30 ьи ~ 9.21558e-05
31 що ~ 2.5838e-06
32 ах ~ 0.0008216509
33 их ~ 0.0016889491
34 сл ~ 0.0026613222
35 лп ~ 2.49768e-05
36 ид ~ 0.0010929508
37 ь ~ 5.1676e-06
38 вк ~ 0.0002489068
39 ха ~ 0.000546906
40 ы ~ 0.0039153303
41 у ~ 0.0036758975
42 ри ~ 0.0042012718
43 яж ~ 0.0001067974
44 ик ~ 0.0028964487
45 фь ~ 1.63641e-05
46 кб ~ 1.7225e-06
47 йр ~ 8.613e-07
48 бя ~ 0.0003987677
49 ых ~ 0.001071419
50 еу ~ 8.61269e-05
51 цу ~ 0.0001386644
52 ка ~ 0.0094446794
53 му ~ 0.0020687689
```

*** Ентропія біграми з пробілами без перетину ***
3.8769590123579682

*** Надлишковість біграми з пробілами без перетину ***
0.22460819752840633

2. За допомогою програми CoolPinkProgram оцінити значення $H(10)$, $H(20)$, $H(30)$.
 $H(10)$:

[illegible]

H(20):

[illegible]

H(30):

Произвольная часть текста:
они_всегда_были_согласны_в_т

Использованные буквы:

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

Введенный символ:
Символ по счету:
Номер эксперимента: 51

Неравенство для энтропии:
 $1.62386930407941 < H < 2.44360515566949$

Двоичная таблица угаданных символов:

000000100000000000000000000000
001000000000000000000000000000
000000001000000000000000000000
100000000000000000000000000000
100000000000000000000000000000

Вероятности:

q[1] = 0.5
q[2] = 0.14
q[3] = 0.12
q[4] = 0.06
q[5] = 0.02
q[6] = 0
q[7] = 0.04
q[8] = 0
q[9] = 0.04
q[10] = 0.02
q[11] = 0.02
q[12] = 0.02
q[13] = 0
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
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

Строка состояния:

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

R(10):

$$1 - \frac{2.22156706429134}{5} = 0.555686587141732$$

$$1 - \frac{2.98262668981542}{5} = 0.403474662036916$$

$$0.403474662036916 < R(10) < 0.555686587141732$$

R(20):

$$1 - \frac{1.43069731699498}{5} = 0.713860536601004$$

$$1 - \frac{2.13366068968819}{5} = 0.573267862062362$$

$$0.555686587141732 < R(20) < 0.713860536601004$$

R(30):

$$1 - \frac{1.62386930407941}{5} = : 0.675226139184118$$

$$1 - \frac{2.44360515566949}{5} = 0.511278968866102$$

$$0.511278968866102 < R(30) < : 0.675226139184118$$

Висновок:

В ході виконання першого комп'ютерного практикуму ми визначили означення ентропії та надлишковості тексту, підвищили навички з процедурного програмування.

Найскладнішим елементом комп'ютерного практикуму була реалізація обробки тексту мовою Python, якщо конкретніше, то з визначенням ентропії для біграм: аналіз проблеми було розтягнуто на довгі години, хоча помилка закралась лише в формулі для обчислення ентропії. Крім цього варто відмітити специфіку CoolPinkProgram, ентропія для 10-грамми постійно була доволі високою, адже не завжди виходило відгадати символ з короткого набору слів, але в результаті значення можна назвати "нормальним".