

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ**  
**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**  
**«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМ.ІГОРЯ СІКОРСЬКОГО»**  
**ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ**  
**Кафедра інформаційної безпеки**

**Комп'ютерний практикум №1**

Виконали:

Студенти 3 курсу

ФБ-01 Літвінчук Софія  
та  
ФБ-02 Косарик Дарія

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.

Постановка задачі:

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  $(10) H$ ,  $(20) H$ ,  $(30) H$ .
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Файл з текстом: text.txt

Файл з кодом: main.py

Під час роботи було створено такі таблиці, які містять відповідну інформацію:

- fr.bi.cross.off.xlsx - частоти біграм, букви яких перетинаються у тексті без пробілів
- fr.bi.cross.xlsx - частоти біграм, букви яких перетинаються у тексті з пробілами
- fr.bi.notcross.off.xlsx - частоти біграм, букви яких не перетинаються у тексті без пробілів
- fr.bi.notcross.xlsx - частоти біграм, букви яких не перетинаються у тексті з пробілами
- fr.letters.off.xlsx - частоти букв у тексті без пробілів
- fr.letters.xlsx - частоти букв у тексті з пробілами

Хід роботи:

0. Підготовка тексту.

- залишили лише слова російською мовою, змінили регістр, видалили зайві пробіли.

```

6 f = open("text.txt", encoding='utf-8')
7 text = f.read()
8 text = text.lower()
9 text = re.sub(r"[\W\d]", " ", text)
10 text = re.sub(r"[A-Za-z]", "", text)
11 text = " ".join(text.split())

```

## 1. Обхоуємо частоту

### Для тексту без пробілів

➤ для букв

```

54 frequency1_l_off = {} #частота букв без пробілів
55 for i in letters_off:
56     frequency1_l_off[i] = text_off.count(i)/len(text_off)

```

	о	е	а	н	и	т	с	л	в	р	к	д	м	у	п
frequency	0,114716	0,086823	0,083203	0,069085	0,066399	0,060293	0,053606	0,05023	0,046488	0,039501	0,034875	0,02965	0,028789	0,027076	0,024093

я	ь	ы	г	б	ч	з	ж	й	ш	х	ю	э	щ	ц	ф	ъ
0,021561	0,020054	0,018577	0,018072	0,017715	0,016917	0,016648	0,011329	0,010739	0,008585	0,007913	0,006383	0,003613	0,002832	0,002716	0,001124	0,000307

➤ для біграм, букви яких перетинаються

```

frequency2_bi_off_cross = {} #частота біграм в тексті без пробілів + перетин
for i in letters_off_bi:
    frequency2_bi_off_cross[i] = all_off_bi_cross.count(i)/(2*len(all_off_bi_cross))

```

	а	б	в	г	д	е	ж	з	и
а	0,000297	0,000864	0,002884	0,00055	0,001766	0,00106	0,000771	0,002766	0,000827
б	0,000527	2,86E-06	5,14E-05	2,86E-06	2E-05	0,001117	4,29E-06	0	0,000513
в	0,003229	7,29E-05	0,000169	0,000159	0,000333	0,002663	3,14E-05	0,000409	0,00278
г	0,000474	2,29E-05	4,14E-05	7,14E-06	0,000704	0,000173	1E-05	1,14E-05	0,00048
д	0,002542	3,43E-05	0,000491	1,14E-05	3,86E-05	0,002766	1,43E-05	1,57E-05	0,001179
е	0,000167	0,001344	0,002397	0,002407	0,001979	0,001587	0,000533	0,001166	0,000767
ж	0,000819	3,29E-05	1,43E-05	5,71E-06	0,000441	0,002416	1,43E-06	4,29E-06	0,000904
з	0,00312	0,000111	0,000529	0,000314	0,000513	0,000159	7,71E-05	3,29E-05	0,000217
и	0,000169	0,000676	0,002666	0,000486	0,00132	0,001694	0,000303	0,001392	0,000956
й	0,000243	0,000161	0,000349	0,000119	0,000369	5,71E-05	9,29E-05	8,43E-05	0,000364
к	0,004713	0,000194	0,000297	4,86E-05	0,000107	0,000354	0,000109	4,86E-05	0,001963
л	0,00469	0,000123	0,000433	0,000156	0,000147	0,003436	0,000253	9,29E-05	0,003309
м	0,001534	0,000154	0,000383	0,000139	0,000193	0,00185	6,86E-05	9,43E-05	0,001917
н	0,006626	0,000186	0,00031	0,000103	0,000507	0,00584	3,43E-05	0,000104	0,004607
о	0,000131	0,002782	0,005482	0,003022	0,003124	0,001876	0,001361	0,000956	0,001247
п	0,000697	1,43E-06	1,43E-06	0	0	0,001204	1,43E-06	2,86E-06	0,000421
р	0,003949	0,000166	0,000211	0,000201	0,000236	0,003189	0,00019	2,14E-05	0,00273
с	0,001076	8,71E-05	0,001257	6,43E-05	0,000306	0,002616	0,000124	4,43E-05	0,000817
т	0,002982	0,00019	0,001857	3,29E-05	0,00018	0,003073	4,57E-05	3,86E-05	0,002446
у	0,00012	0,000393	0,000993	0,000639	0,001089	0,000239	0,001147	0,000253	0,000374
ф	4,71E-05	1,43E-06	0	0	0	7,86E-05	0	0	0,000176
х	0,000551	7,29E-05	0,000163	3,86E-05	0,000109	5,71E-05	2,71E-05	3,86E-05	0,000197
ц	0,000271	5,71E-06	3,71E-05	2,86E-06	0	0,000349	0	4,29E-06	0,000151
ч	0,001459	8,57E-06	6,14E-05	5,71E-06	1,71E-05	0,001917	1,43E-06	2,43E-05	0,00079
ш	0,00048	7,14E-06	3,57E-05	2,86E-06	0	0,001336	0	0	0,00092
щ	0,000176	0	1,43E-06	0	0	0,000766	0	0	0,000429
ъ	0	0	0	0	0	7,57E-05	0	0	0
ы	3,29E-05	0,000427	0,000783	0,000119	0,000219	0,000599	4,29E-05	0,000123	0,000226

➤ для біграм, букви яких не перетинаються

```
frequency2_bi_off_not_cross = {} #частота біграм із тексту без проблів + не перетин
for i in letters_off_bi:
    frequency2_bi_off_not_cross[i] = all_off_bi_not_cross.count(i) / (2*len(all_off_bi_not_cross))
```

	а	б	в	г	д	е	ж	з	и	й
а	0,000297	0,000864	0,002884	0,00055	0,001766	0,00106	0,000771	0,002766	0,000827	0,000313
б	0,000527	2,86E-06	5,14E-05	2,86E-06	2E-05	0,001117	4,29E-06	0	0,000513	0
в	0,003229	7,29E-05	0,000169	0,000159	0,000333	0,002663	3,14E-05	0,000409	0,00278	0
г	0,000474	2,29E-05	4,14E-05	7,14E-06	0,000704	0,000173	1E-05	1,14E-05	0,00048	0
д	0,002542	3,43E-05	0,000491	1,14E-05	3,86E-05	0,002766	1,43E-05	1,57E-05	0,001179	0
е	0,000167	0,001344	0,002397	0,002407	0,001979	0,001587	0,000533	0,001166	0,000767	0,001516
ж	0,000819	3,29E-05	1,43E-05	5,71E-06	0,000441	0,002416	1,43E-06	4,29E-06	0,000904	0
з	0,00312	0,000111	0,000529	0,000314	0,000513	0,000159	7,71E-05	3,29E-05	0,000217	0
и	0,000169	0,000676	0,002666	0,000486	0,00132	0,001694	0,000303	0,001392	0,000956	0,000806
й	0,000243	0,000161	0,000349	0,000119	0,000369	5,71E-05	9,29E-05	8,43E-05	0,000364	0
к	0,004713	0,000194	0,000297	4,86E-05	0,000107	0,000354	0,000109	4,86E-05	0,001963	0
л	0,00469	0,000123	0,000433	0,000156	0,000147	0,003436	0,000253	9,29E-05	0,003309	0
м	0,001534	0,000154	0,000383	0,000139	0,000193	0,00185	6,86E-05	9,43E-05	0,001917	0
н	0,006626	0,000186	0,00031	0,000103	0,000507	0,00584	3,43E-05	0,000104	0,004607	0
о	0,000131	0,002782	0,005482	0,003022	0,003124	0,001876	0,001361	0,000956	0,001247	0,001846
п	0,000697	1,43E-06	1,43E-06	0	0	0,001204	1,43E-06	2,86E-06	0,000421	0
р	0,003949	0,000166	0,000211	0,000201	0,000236	0,003189	0,00019	2,14E-05	0,00273	0
с	0,001076	8,71E-05	0,001257	6,43E-05	0,000306	0,002616	0,000124	4,43E-05	0,000817	0
т	0,002982	0,00019	0,001857	3,29E-05	0,00018	0,003073	4,57E-05	3,86E-05	0,002446	0
у	0,00012	0,000393	0,000993	0,000639	0,001089	0,000239	0,001147	0,000253	0,000374	8E-05
ф	4,71E-05	1,43E-06	0	0	0	7,86E-05	0	0	0,000176	0
х	0,000551	7,29E-05	0,000163	3,86E-05	0,000109	5,71E-05	2,71E-05	3,86E-05	0,000197	0
ц	0,000271	5,71E-06	3,71E-05	2,86E-06	0	0,000349	0	4,29E-06	0,000151	0
ч	0,001459	8,57E-06	6,14E-05	5,71E-06	1,71E-05	0,001917	1,43E-06	2,43E-05	0,00079	0
ш	0,00048	7,14E-06	3,57E-05	2,86E-06	0	0,001336	0	0	0,00092	0
щ	0,000176	0	1,43E-06	0	0	0,000766	0	0	0,000429	0
ъ	0	0	0	0	0	7,57E-05	0	0	0	0
ы	3,29E-05	0,000427	0,000783	0,000119	0,000219	0,000599	4,29E-05	0,000123	0,000226	0,00071

Для тексту з пробілами

➤ для букв

```
70 frequency1_l = {} #частота букв із пробілами
71 for i in letters:
72     frequency1_l[i] = text.count(i) / len(text)
```

		о	е	а	н	и	т	с	л	в	р	к	д	м	у	п
frequency	0,167352	0,095518	0,072293	0,069279	0,057523	0,055287	0,050203	0,044635	0,041824	0,038708	0,03289	0,029038	0,024688	0,023971	0,022544	0,020061

я	ь	ы	г	б	ч	з	ж	й	ш	х	ю	э	щ	ц	ф	ъ
0,017953	0,016698	0,015468	0,015048	0,014751	0,014086	0,013862	0,009433	0,008942	0,007148	0,006589	0,005315	0,003008	0,002358	0,002261	0,000936	0,000256

➤ для біграм, букви яких перетинаються

```
frequency2_bi_cross = {} #частота біграм із тексту з пробілів + перетин
for i in letters_bi:
    frequency2_bi_cross[i] = all_bi_cross.count(i) / (2 * len(all_bi_cross))
```

	а	б	в	г	д	е	ж	з	и	й	к
а	5,35E-06	0,000321	0,001496	0,000268	0,001087	0,000502	0,000559	0,002067	6,42E-05	0,000269	0,002059
б	0,000448	5,95E-07	3,39E-05	3,57E-06	1,67E-05	0,000969	2,97E-06	0	0,000407	0	0,000125
в	0,002677	6,54E-06	1,13E-05	1,9E-05	0,000136	0,002107	5,95E-07	0,000269	0,002254	0	7,79E-05
г	0,000404	5,95E-07	5,95E-07	0	0,000581	0,000139	0	0	0,000396	0	6,78E-05
д	0,002125	1,07E-05	0,000377	2,38E-06	1,67E-05	0,00223	5,35E-06	1,19E-06	0,000973	0	0,000124
е	1,96E-05	0,000649	0,001162	0,00181	0,001212	0,001055	0,000359	0,000632	4,76E-05	0,001259	0,000874
ж	0,000671	2,38E-05	0	4,16E-06	0,000379	0,001991	2,97E-06	0	0,000738	0	4,22E-05
з	0,002613	7,55E-05	0,000406	0,000247	0,000391	0,000123	5,59E-05	5,35E-06	0,000152	0	4,76E-05
и	5,59E-05	0,000189	0,001257	0,000197	0,00072	0,00115	0,000152	0,000887	0,000278	0,000676	0,000846
й	0	1,19E-06	0	0	8,86E-05	5,95E-07	0	5,95E-07	0	0	4,1E-05
к	0,003901	0	7,02E-05	0	5,95E-07	0,00024	7,14E-06	1,78E-06	0,00147	0	6,54E-06
л	0,003868	1,19E-05	1,19E-06	5,06E-05	6,54E-06	0,002602	0,000194	5,95E-07	0,002515	0	0,000149
м	0,001225	2,97E-06	1,78E-06	1,43E-05	0	0,001467	0	1,78E-06	0,001288	0	2,85E-05
н	0,005417	5,35E-06	2,97E-06	4,7E-05	0,000328	0,004853	5,95E-07	7,14E-06	0,003693	0	0,000117
о	0	0,001722	0,003469	0,002281	0,002111	0,00119	0,00094	0,000531	0,000371	0,001512	0,000765
п	0,000581	0	0	0	0	0,000991	0	0	0,000363	0	3,03E-05
р	0,003281	0,000118	0,000152	0,000156	0,000183	0,002653	0,000142	8,92E-06	0,00224	0	0,000269
с	0,000904	2,8E-05	0,000924	7,14E-06	0,000172	0,002201	1,19E-05	1,13E-05	0,000635	0	0,002482
т	0,002479	8,33E-06	0,001358	2,38E-06	5,17E-05	0,002467	1,19E-06	5,95E-07	0,001922	0	0,000237
у	1,72E-05	0,000225	0,000548	0,000481	0,000758	0,000118	0,00089	0,000149	1,19E-06	6,36E-05	0,000359
ф	3,39E-05	0	0	0	0	6,6E-05	0	0	0,000144	0	0
х	0,000404	0	4,94E-05	1,19E-06	0	2,2E-05	0	0	4,94E-05	0	0
ц	0,000213	5,95E-07	1,78E-05	0	0	0,000309	0	0	0,000105	0	7,73E-05
ч	0,001131	0	2,97E-06	0	0	0,001611	0	0	0,000642	0	0,000148
ш	0,00042	0	2,68E-05	0	0	0,001119	0	0	0,000738	0	0,000189
щ	0,000149	0	0	0	0	0,000595	0	0	0,000352	0	0
ъ	0	0	0	0	0	6,48E-05	0	0	0	0	0
ы	0	0,000236	0,000479	4,76E-05	6,9E-05	0,000469	1,13E-05	3,09E-05	1,19E-06	0,000607	8,51E-05

➤ для біграм, букви яких не перетинаються

```
78 frequency2_bi_not_cross = {} #частота біграм із тексту з пробілами + не перетин
79 for i in letters_off_bi:
80     frequency2_bi_not_cross[i] = all_bi_not_cross.count(i) / (2 * len(all_bi_not_cross))
```

	а	б	в	г	д	е	ж	з	и
а	4,76E-06	0,000291	0,001518	0,000263	0,001104	0,000473	0,000569	0,002053	5,83E-05
б	0,000478	0	3,09E-05	4,76E-06	2,02E-05	0,000924	2,38E-06	0	0,000414
в	0,002657	4,76E-06	1,55E-05	1,9E-05	0,000134	0,002114	0	0,000282	0,002274
г	0,0004	0	1,19E-06	0	0,000596	0,00014	0	0	0,000389
д	0,002089	8,33E-06	0,00034	2,38E-06	1,67E-05	0,002213	4,76E-06	2,38E-06	0,000966
е	2,02E-05	0,00069	0,001112	0,001856	0,001248	0,001041	0,000351	0,000651	5,23E-05
ж	0,000689	2,38E-05	0	5,95E-06	0,00039	0,001995	1,19E-06	0	0,000732
з	0,00269	7,61E-05	0,00042	0,000233	0,000358	0,000118	4,88E-05	5,95E-06	0,000153
и	6,07E-05	0,000177	0,001282	0,000193	0,000763	0,001159	0,000156	0,00091	0,000285
й	0	2,38E-06	0	0	9,52E-05	0	0	1,19E-06	0
к	0,00404	0	6,42E-05	0	1,19E-06	0,000232	9,52E-06	3,57E-06	0,00148
л	0,003898	9,52E-06	1,19E-06	4,28E-05	5,95E-06	0,002608	0,000201	0	0,002468
м	0,001262	3,57E-06	1,19E-06	1,43E-05	0	0,001463	0	1,19E-06	0,00124
н	0,005379	5,95E-06	0	4,76E-05	0,000308	0,004853	0	7,14E-06	0,003672
о	0	0,001755	0,00347	0,002255	0,002089	0,001217	0,000959	0,000528	0,000349
п	0,000594	0	0	0	0	0,001017	0	0	0,000364
р	0,003244	0,000109	0,000174	0,000164	0,000172	0,002631	0,000137	8,33E-06	0,00222
с	0,000903	2,14E-05	0,000927	5,95E-06	0,000163	0,002234	7,14E-06	1,67E-05	0,000639
т	0,002481	5,95E-06	0,001386	2,38E-06	5,71E-05	0,002491	2,38E-06	0	0,001985
у	1,43E-05	0,000232	0,000547	0,000502	0,000795	0,000115	0,000884	0,000139	0
ф	3,81E-05	0	0	0	0	8,09E-05	0	0	0,00014
х	0,0004	0	5,35E-05	2,38E-06	0	2,26E-05	0	0	5E-05
ц	0,000213	1,19E-06	1,9E-05	0	0	0,000305	0	0	9,75E-05
ч	0,001153	0	4,76E-06	0	0	0,001598	0	0	0,000678
ш	0,000394	0	2,14E-05	0	0	0,001135	0	0	0,000761
щ	0,000163	0	0	0	0	0,000579	0	0	0,000359
ъ	0	0	0	0	0	5,95E-05	0	0	0
ы	0	0,000236	0,000487	5,12E-05	8,09E-05	0,000446	9,52E-06	3,57E-05	2,38E-06

## 2. Обраховуємо ентропію

Для тексту без пробілів

➤ для букв

```

85 sum1 = 0
86 for i in frequency1_l_off.values():
87     sum1 = sum1 + (i*math.log(i, 2))
88 h1_off = -sum1 #для букв + текст без пробілів
89 print("H1 (без проб.): ", h1_off)

```

➤ для біграм, букви яких перетинаються

```

121 sum6 = 0
122 for i in frequency2_bi_off_cross.values():
123     if i != 0:
124         sum6 = sum6 + (i * math.log(i, 2))
125 h2_off_cross = -sum6 #біграми з тексту без пробілів + перетин.
126 print("H2 (без проб., перетин.) : ", h2_off_cross)

```

➤ для біграм, букви яких не перетинаються

```

100     sum3 = 0
101     for i in frequency2_bi_off_not_cross.values():
102         if i != 0:
103             sum3 = sum3 + (i * math.log(i, 2))
104     h2_off_not_cross = -sum3    #біграми з тексту без пробілів + не перетин.
105     print("H2 (без проб., не перетин.): ", h2_off_not_cross)

```

Для тексту з пробілами

➤ для букв

```

92     sum2 = 0
93     for i in frequency1_l.values():
94         sum2 = sum2 + (i * math.log(i, 2))
95     h1 = -sum2    #для букв + текст з пробілами
96     print("H1 (з проб.) : ", h1)

```

➤ для біграм, букви яких перетинаються

```

107     sum4 = 0
108     for i in frequency2_bi_not_cross.values():
109         if i != 0:
110             sum4 = sum4 + (i * math.log(i, 2))
111     h2_not_cross = -sum4    #біграми з тексту з пробілами + не перетин.
112     print("H2 ( з проб., не перетин.) : ", h2_not_cross)

```

➤ для біграм, букви яких не перетинаються

```

114     sum5 = 0
115     for i in frequency2_bi_cross.values():
116         if i != 0:
117             sum5 = sum5 + (i * math.log(i, 2))
118     h2_cross = -sum5    #біграми з тексту з пробілами + перетин.
119     print("H2 (з проб., перетин.) : ", h2_cross)

```

```

H1 (без проб.):  4.4437247660503365
H1 (з проб.) :  4.351650441166879
H2 (без проб., не перетин.):  4.622885811929124
H2 ( з проб., не перетин.) :  3.147158266504846
H2 (з проб., перетин.) :  4.441990097790937
H2 (без проб., перетин.) :  4.623223587995598

```

2. Робота з програмою CoolPinkProgram .

Результати:

$$3,32813 < H(10) < 3,96108$$

Лабораторная работа №1



Произвольная часть текста:

ми\_но\_тот

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:

3,32812539321919 < H < 3,961076880026

Двоичная таблица угаданных символов:

01000000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000  
00000000100000000000000000000000  
01000000000000000000000000000000  
00000000000000000000000000000000

Вероятности:

q[1] = 0,24  
q[2] = 0,1  
q[3] = 0,02  
q[4] = 0,08  
q[5] = 0,04  
q[6] = 0,04  
q[7] = 0,02  
q[8] = 0,02  
q[9] = 0,04  
q[10] = 0,06  
q[11] = 0,02  
q[12] = 0,04  
q[13] = 0,04  
q[14] = 0,02  
q[15] = 0,06  
q[16] = 0  
q[17] = 0,02  
q[18] = 0  
q[19] = 0,02  
q[20] = 0  
q[21] = 0,02  
q[22] = 0  
q[23] = 0  
q[24] = 0  
q[25] = 0  
q[26] = 0,02  
q[27] = 0,04  
q[28] = 0  
q[29] = 0,02  
q[30] = 0,02  
q[31] = 0  
q[32] = 0

$$2,1815722 < H(20) < 2,80936$$

Лабораторная работа №1



Произвольная часть текста:

дуг\_попасться\_инди

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

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:

2,18157222774133 < H < 2,80936468627839

Двоичная таблица угаданных символов:

00000000010000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000  
01000000000000000000000000000000  
.....

Вероятности:

q[1] = 0,4509803  
q[2] = 0,1960784  
q[3] = 0  
q[4] = 0,0196078  
q[5] = 0,0196078  
q[6] = 0,0196078  
q[7] = 0  
q[8] = 0  
q[9] = 0,0196078  
q[10] = 0,019607  
q[11] = 0  
q[12] = 0,039215  
q[13] = 0,019607  
q[14] = 0,058823  
q[15] = 0  
q[16] = 0,019607  
q[17] = 0  
q[18] = 0,039215  
q[19] = 0  
q[20] = 0  
q[21] = 0,019607  
q[22] = 0  
q[23] = 0  
q[24] = 0,019607  
q[25] = 0  
q[26] = 0,019607  
q[27] = 0  
q[28] = 0,019607  
q[29] = 0  
q[30] = 0  
q[31] = 0  
q[32] = 0

$$1,77114 < H(30) < 2,4520$$



Лабораторная работа №1

Произвольная часть текста:  
вам\_то\_мое\_место\_я\_его\_первый

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:

Символ по счету:

Номер эксперимента: 50

Неравенство для энтропии:  
 $1,77113965312456 < H < 2,45204727688637$

Двоичная таблица угаданных символов:

00000010000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Поле ввода символов:

Продолжить Другой

Вероятности:

q[1] = 0,5714285
q[2] = 0,1020408
q[3] = 0,0612244
q[4] = 0
q[5] = 0,0408163
q[6] = 0
q[7] = 0,0204081
q[8] = 0,0204081
q[9] = 0,0204081
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0,020408
q[18] = 0
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0,040816
q[24] = 0,020408
q[25] = 0
q[26] = 0,020408
q[27] = 0,020408
q[28] = 0,020408
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0,020408

Строка состояния:

### 3. Надлишковість:

```

R_h1 = 1 - (h1/(math.log(33, 2))) #для букв з тексту з пробілами
R_h1_off = 1 - (h1_off/(math.log(33, 2))) #для букв з тексту без пробілів
R_h2_cross = 1 - (h2_cross/(math.log(33, 2))) # для біграм, букви яких перетинаються у тексті з пробілами
R_h2_cross_off = 1 - (h2_off_cross/(math.log(33, 2))) # для біграм, букви яких перетинаються у тексті без пробілами
R_h2_notcross = 1 - (h2_not_cross/(math.log(33, 2))) #для біграм, букви яких не перетинаються у тексті з пробілами
R_h2_notcross_off = 1 - (h2_off_not_cross/(math.log(33, 2))) #для біграм, букви яких не перетинаються у тексті безпробілами

```

```

R1 - 0.13732941197696846
R2 - 0.11907661041055018
R3 - 0.1194204908089399
R4 - 0.08349278850884467
R5 - 0.37610777587198085
R6 - 0.08355974919004483

```

### 4. Використані формули у роботі:

- Для обчислення надлишковості:

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

$$H_0 = \log_2 m,$$

m - к-ть букв у алфавіті (33 букви у нашому випадку)

- Для обчислення ентропій H1 і H2

$$H_n = \frac{1}{n} H(x_1, x_2, \dots, x_n)$$

, де

$$H(x_1, x_2, \dots, x_n) = - \sum_{z_1, z_2, \dots, z_n} P(x_1 = z_1, \dots, x_n = z_n) \cdot \log_2 P(x_1 = z_1, \dots, x_n = z_n).$$

Отримані результати ентропій та надлишковостей занесли у таблицю:

	Ентропія	Надлишковість
Букви, текст з пробілами	4,351	0,137
Букви, текст без пробілів	4,444	0,119
Біграми, що не перетин., текст з пробілами	3,147	0,376
Біграми, що не перетин., текст без пробів	4,623	0,083
Біграми, що перетин., текст з пробілами	4,441	0,119
Біграми, що перетин., текст без пробілів	4,623	0,083

$3,32813 < H(10) < 3,96108$	$0,2147 < R(10) < 0,340258$
$2,1815722 < H(20) < 2,80936$	$0,443144 < R(20) < 0,567639$
$1,77114 < H(30) < 2,4520$	$0,513916 < R(30) < 0,648917$

## 5. Аналіз результатів

Ми ознайомилися з такими поняттями і ось як ми їх зрозуміли:

- ентропія - це те, скільки інформації про джерело нам невідомо.
- надлишковість - величина, що показує, на скільки коротшим може бути повідомлення, при чому зміст залишається незмінним.

Проаналізувавши отримані значення частот, ми переконалися в тому, що найчастішими літерами в російському алфавіті є:

```
('о', 0.11471573578678934)
('е', 0.08682291257420013)
('а', 0.08320273156514969)
('н', 0.06908488281556935)
('и', 0.06639903423742616)
('т', 0.060293014650732536)
('с', 0.05360553741972813)
('л', 0.05022965433985985)
('в', 0.04648803868764867)
('р', 0.03950054645589422)
```

Отримані значення ентропій, показують те, що вона більша у тексті без пробілів.

### Висновки

У результаті виконання практикума ми ознайомилися з поняттями ентропії на символ джерела та його надлишковості, набули практичних навичок щодо оцінки ентропії та надлишковості.