

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ НАЦІОНАЛЬНИЙ
ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ "КИЇВСЬКИЙ
ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО"
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Виконали:
Студентки групи ФБ-33
Яремко А.В та Журавльова М.В

Варіант №3

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.

Порядок виконання роботи

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: “Злочин і кара” Ф. М. Достоевський (~ 2.1 МБ)

Перед початком обчислень було проведено попереднє очищення тексту за методичними вказівками, а результат збережено у файли clean_with_spaces.txt та clean_no_spaces.txt .

Ймовірність появи кожного символу обчислювалась як відношення кількості входжень символу до загальної кількості символів у тексті: $p_i = \frac{n_i}{N}$, де n_i - кількість появ символу, N - довжина тексту.

Для біграм застосовувалися два підходи: перекривні та неперекривні біграми.

Для обчислення ентропії літер використовувалась формула: $H_1 = - \sum_{i=1}^n p_i \log(p_i)$, а для

ентропії біграм: $H_2 = - \frac{1}{2} \sum_{i,j} p_{ij} \log(p_{ij})$

Для оцінки надлишковості тексту була застосована формула: $R = 1 - \frac{H}{\log |A|}$, де $|A|$ - розмір алфавіту тексту.

Результат роботи коду:

```
АНАЛІЗ ТЕКСТУ: «Злочин і кара» Ф.М. Достоевського
=====

--- ТЕКСТ З ПРОБІЛАМИ ---

Аналіз файлу: prestuplenie-i-nakazanie-fedor-dostoevskij.txt
Довжина сирого тексту: 1,248,151 символів
Після фільтрації: 1,165,388 символів
Збережено очищений текст -> results\clean_with_spaces.txt

- ЕНТРОПІЇ:
H1 = 4.355119
H2 (перекривні) = 3.949231
H2 (неперекривні)= 3.948646

- НАДЛИШКОВОСТІ:
|A| = 32, log2(|A|) = 5.000000
R1 (H1) = 0.128976
R2 (H2, перекривні) = 0.210154
R2 (H2, неперекривні) = 0.210271
```

```
--- ТЕКСТ БЕЗ ПРОБІЛІВ ---

Аналіз файлу: prestuplenie-i-nakazanie-fedor-dostoevskij.txt
Довжина сирого тексту: 1,248,151 символів
Після фільтрації: 969,927 символів
Збережено очищений текст -> results\clean_no_spaces.txt

- ЕНТРОПІЇ:
H1 = 4.448816
H2 (перекривні) = 4.131265
H2 (неперекривні) = 4.130387

- НАДЛИШКОВОСТІ:
|A| = 31, log2(|A|) = 4.954196
R1 (H1) = 0.102011
R2 (H2, перекривні) = 0.166108
R2 (H2, неперекривні) = 0.166285

===== ПОРІВНЯННЯ РЕЗУЛЬТАТІВ =====
Параметр | 3 пробілами | Без пробілів |
-----|-----|-----|
H1 | 4.355119 | 4.448816 |
H2 (перекривні) | 3.949231 | 4.131265 |
H2 (неперекривні) | 3.948646 | 4.130387 |
R1 (H1) | 0.128976 | 0.102011 |
R2 (H2, перекривні) | 0.210154 | 0.166108 |
R2 (H2, неперекривні) | 0.210271 | 0.166285 |
=====
```

Монограми з пробілами:

Символ	Ймовірність
	0,167721823
о	0,09501814
е	0,072889887
а	0,065776377
и	0,055329212
н	0,054290931
т	0,053166842
с	0,045092278
в	0,038981009
л	0,037626095
р	0,035563263
к	0,027648303
д	0,026359461
м	0,02618527
у	0,024091547
п	0,022923696
ь	0,018311498
я	0,01744655
ч	0,014608868
г	0,014414942
б	0,014215866
ы	0,013722468
з	0,012867817
ж	0,009202085
й	0,008731856
х	0,007284269
ш	0,006584073
ю	0,004752923
э	0,002846262
щ	0,002662633
ц	0,002503029
ф	0,001180723

Монограми без пробілів:

Символ	Ймовірність
о	0,114166324
е	0,087578756
а	0,079031721
и	0,06647923
н	0,065231713
т	0,063881096
с	0,054179335
в	0,046836515
л	0,045208557
р	0,04273002
к	0,033220026
д	0,031671456
м	0,031462162
у	0,028946508
п	0,02754331
ь	0,022001656
я	0,020962402
ч	0,017552867
г	0,01731986
б	0,017080667
ы	0,016487839
з	0,015460957
ж	0,011056502
й	0,010491511
х	0,008752205
ш	0,007910905
ю	0,005710739
э	0,003419845
щ	0,00319921
ц	0,003007443
ф	0,001418663

Біграми з пробілами перекривні:

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ь	э	ю	я			
а	2988	7651	19237	3798	9315	4591	2567	4990	13545	13	9307	2853	6720	18183	12970	18775	5981	18874	11407	5091	619	1852	362	6880	651	71	0	0	3274	76	2820		
б	19623	3	597	3123	740	2086	1417	1809	3944	126	993	6163	7965	3753	4231	12	926	2158	5212	4910	87	71	1098	71	1044	974	294	0	0	5	988	2332	
в	514	695	1	63	24	2372	4	2	954	0	136	850	46	319	2208	0	1379	232	4	1481	0	48	1	20	12	281	4042	207	0	8	664		
г	7135	6594	4	29	26	887	5856	0	563	3597	0	141	710	115	1725	7229	418	671	4275	266	726	0	71	21	196	737	12	3016	170	0	0	238	
д	1311	1183	0	8	12	1218	339	0	1	697	0	59	1459	39	290	8953	0	644	51	6	500	0	0	0	35	2	0	0	0	0	12	0	
е	1111	5738	31	998	20	47	5071	23	16	2640	0	237	682	87	2067	4359	99	2023	368	286	2208	1	47	282	71	91	2	504	1131	1	11	466	
ж	21964	41	1455	1647	3702	3283	1930	925	1338	126	2502	1341	6411	4914	8110	334	1281	6866	5353	6266	195	28	515	372	1420	1088	1022	0	0	0	263	253	
з	739	1361	25	0	5	873	4882	11	0	1447	0	107	18	4	828	26	0	5	12	0	316	0	0	1	6	0	0	0	56	0	2	0	
и	1477	5114	154	1027	366	815	261	52	12	443	0	141	274	271	2023	596	35	274	35	8	665	0	0	3	22	7	0	345	151	0	2	458	
й	20205	137	722	2805	628	1940	2250	253	2129	749	1369	2812	5344	2963	3556	273	275	1005	2458	4724	10	26	2049	948	1859	674	167	0	0	1	399	1610	
к	7682	0	0	0	0	280	4	0	4	0	0	137	239	50	396	8	1	2	556	376	0	0	0	46	181	212	2	0	0	0	0	0	
л	5594	7768	0	278	0	2	466	42	19	2643	0	11	581	3	599	10240	0	1688	247	623	1348	0	35	18	0	16	0	0	0	0	0	0	
м	7593	6003	27	14	139	21	4652	330	5	6239	0	321	75	16	286	6110	62	0	1507	105	1657	2	0	0	219	2	1	639	5427	0	889	1508	
н	8972	3516	17	13	79	0	4485	0	0	3097	0	84	105	37	1686	4114	62	68	116	30	2443	11	0	10	52	6	1	994	91	1	3	423	
о	4782	11587	11	27	75	345	11157	3	25	9352	0	302	0	0	3059	11275	0	95	399	525	3008	38	5	263	222	8	97	3219	1288	0	230	1913	
п	27240	14	4039	9230	4976	5634	2260	2321	1089	732	3534	1852	6674	3189	6997	261	1572	5732	6826	7368	101	224	526	180	2482	1065	266	0	0	22	646	731	
р	79	917	13	0	1	0	3326	0	0	1110	0	77	851	0	129	10341	30	7783	80	74	854	0	0	16	22	7	0	239	171	0	0	595	
с	895	8782	193	436	256	346	6049	287	54	5873	0	238	74	362	819	8396	120	13	186	751	2907	299	72	22	95	265	17	987	1032	0	208	1111	
т	4366	1010	99	1579	31	219	4875	25	10	1603	0	4763	8114	1226	993	3271	1921	238	822	11894	837	6	229	63	407	94	1	253	3341	0	271	3988	
у	6697	6521	14	2699	10	114	6717	0	25	4013	0	407	236	47	1279	16566	76	3460	1229	74	1997	2	16	66	353	10	32	1537	7218	10	91	444	
ф	7165	67	887	718	1578	2184	226	1917	266	8	227	707	1579	1697	571	4	888	638	1137	1626	1	16	450	11	916	656	364	0	0	1	901	70	
х	49	268	0	0	0	88	0	0	379	0	3	15	6	1	145	0	128	22	15	115	10	0	0	1	1	0	23	107	0	0	0	0	
ц	3623	498	0	198	0	0	326	0	8	587	0	1	145	68	156	2539	0	96	30	7	183	0	0	0	1	21	0	0	0	2	0	0	
ч	440	586	0	24	0	0	911	0	0	312	0	25	0	3	0	198	0	0	0	5	218	0	0	0	0	0	0	0	0	1	194	0	0
ш	534	2844	0	2	0	0	4524	0	0	1810	0	533	21	4	986	85	0	92	0	4451	731	0	0	0	0	0	0	147	0	0	261	0	0
щ	71	988	0	9	0	0	2383	0	0	1549	0	494	571	12	456	213	0	0	0	53	225	0	0	0	0	0	0	0	0	648	0	1	0
ь	6	452	0	1	0	0	1685	0	0	688	0	0	0	0	0	55	2	0	3	0	0	0	0	0	0	0	0	0	0	41	0	0	
э	4990	0	156	910	121	171	909	24	58	8	1426	172	2095	1091	211	0	115	231	733	777	2	2	0	1070	1	191	517	12	0	0	0	0	3
ю	13100	0	101	32	146	21	521	0	176	151	0	1376	0	987	2491	5	1	0	939	189	0	13	128	69	63	333	29	0	0	0	0	381	708
я	28	0	0	5	9	6	0	0	1	0	16	63	9	7	24	0	38	3	10	3046	0	9	43	0	0	0	0	0	0	0	0	0	0
а	3507	0	349	4	13	249	0	18	8	5	1	17	16	47	35	0	0	71	198	392	0	0	6	13	111	65	276	0	0	0	0	42	0
б	13269	0	23	312	68	539	112	113	253	19	95	193	883	402	669	0	53	97	683	1705	0	0	229	78	156	11	156	0	0	0	0	115	97

Біграми з пробілами неперекривні:

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	м	ь	э	ю	я		
а	1496	3857	9643	1881	4685	2236	1297	2504	6733	6	4702	1425	3347	9101	6524	9341	2595	9309	5669	2602	314	945	170	3435	330	32	0	0	1640	35	1373		
б	9810	2	285	1551	360	1052	681	860	1918	62	484	3065	3920	1905	2165	5	473	1052	2603	2454	48	41	547	35	528	488	151	0	0	3	505	1105	
в	241	938	0	31	0	10	1190	2	2	511	0	73	410	24	172	1104	0	667	112	2	749	0	28	1	8	4	133	2006	109	0	3	329	
г	3523	3294	2	14	12	447	2890	0	283	1817	0	74	339	56	833	3636	207	334	2171	119	374	0	33	10	90	389	3	1517	89	0	0	139	
д	655	568	0	5	5	597	171	0	1	337	0	35	730	21	149	4562	0	331	22	2	244	0	0	0	13	2	0	0	0	0	7	0	
е	546	2866	18	517	10	24	2562	8	7	1337	0	121	320	49	1045	2175	42	1004	183	145	1107	0	25	148	32	51	1	262	542	0	5	234	
ж	11016	25	713	836	1859	1652	963	456	662	64	1260	691	3132	2511	4003	185	654	3440	2685	3198	97	14	260	192	713	550	521	0	0	0	137	136	
з	373	680	12	0	3	447	2458	7	0	718	0	51	7	3	426	10	0	1	6	0	0	0	0	1	4	0	0	0	0	27	0	1	
и	754	2577	75	496	175	405	130	29	6	215	0	74	140	127	1021	296	0	123	14	5	343	0	0	0	17	5	0	190	82	0	1	245	
й	10179	62	375	1387	316	951	1118	128	1076	366	669	1399	2640	1476	1805	146	136	473	1237	2366	6	12	995	455	901	333	84	0	0	0	197	837	
к	3868	0	0	0	0	133	1	0	2	0	0	74	115	28	194	5	0	2	308	188	0	0	0	21	96	102	2	0	0	0	0	0	
л	2779	3842	0	144	0	2	241	20	8	1330	0	6	295	1	309	5092	0	809	129	323	671	0	15	9	0	11	0	0	0	0	0	0	
м	3828	3085	16	5	77	16	2373	157	1	3143	0	169	36	7	145	3058	29	809	129	323	671	0	15	9	0	11	0	0	0	0	0	0	
н	4500	1757	8	6	28	0	2277	0	0	1500	0	145	40	19	844	2033	35	30	59	16	1195	3	0	5	23	4	1	512	49	1	4	484	0
о	2456	5796	7	12	38	160	5509	0	14	4706	0	145	0	0	1547	5663	0	51	190	249	1467	17	1	139	118	3	44	1853	641	0	117	0	
п	13546	6	2047	4616	2461	2801	1133	1156	558	373	1755	912	3298	3073	3453	117	795	2889	3937	3748	57	105	245	95	1244	53	128	0	0	0	8	312	
р	38	450	7	7	5	580	467	44	467	0	550	47	44	467	0	580	467	44	467	0	580	467	44	467	0	580	467	44	467	0	580	467	44
с	47	450	7	7	5	580	467	44	467	0	550	47	44	467	0	580	467	44	467	0	580	467	44	467	0	580	467	44	467	0	580	467	44
т	2230	992	43	774	14	112	2398	12	6	814	0	2369	1595	6251	6262	1669	963	123	437	5886	409	3	130	32	210	38	1	130	1693	0	129	50	
у	55	3295	5	1355	5	10	3295	5	10	2065	15	5	10	3295	5	10	2065	15	5	10	3295	5	10	2065	15	5	10	3295	5	10	2065	15	5
ф	3950	32	458	595	779	1068	102	975	127	6	115	339	836	830	270	0	451	294	584	807	1	8	206	6	464	319	189	0	0	0	0	478	0
х	27	134	0	0	0	0	152	0	0	214	0	2	9	3	1	70	0	66	12	5	44	0	0	0	0	0	0	0	0	0	0	0	
ц	1857	269	0	94	0	0	55	0	0	300	0	1	75	28	1	75	28	1	75	28	1	75	28	1	75	28	1	75	28	1	75	28	1
ч	224	307	0	12	477	0	477	0	12	152	0	152	0	12	477	0	12	477	0	12	477	0	12	477	0	12	477	0	12	477	0	12	477
ш	268	1463	0	0	0	0	2779	0	0	899	0	291	10	3	479	32	0	35	0	2219	357	0	0	0	0	0	0	0	0	0	0	138	0
щ	34	485	0	5	0	0	1194	0	0	740	0	255	278	11	240	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	325	1	
м	3	222	0	0	0	0	828	0	0	368	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	
н	2490	0	73	451	64	44	246	10	31	4	698	80	1059	568	104	0	63	105	566	388	2	0	0	514	1	98	274	0	0	0	0	0	
о	5635	0	51	21	78	12	857	0	88	69	0	682	0	188	1245	3	1	0	435	91	0	12	75	32	29	170	152	0	0	0	0	181	
п	5	15	0	0	2	2	2	0	0	1	0	8	13	7	164	0	20	1	5	1525	0	4	20	0	0	0	0	0	0	0	0	0	
р	1752	0	1776	3	7	176	7	0	0	3	16	7	164	0	20	1	5	1525	0	4	20	0	0	0	0	0	0	0	0	0	0	0	
с	6508	0	9	166	32	285	50	60	115	10	41	89	455	205	339	14	40	44	365	813	0	0	115	33	68	7	79	0	0	0	0	64	

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	м	ь	э	ю	я	
а	307	1247	5052	1110	3011	1957	1999	4435	1617	993	7145	8274	4481	6076	1299	2761	2700	7090	6073	598	126	1276	103	1722	1040	306	0	0	328	991	2538	
б	701	5	87	1	26	2389	4	17	991	0	149	850	52	346	2302	11	1382	265	26	1527	0	50	1	42	12	281	4042	207	94	8	699	
в	6682	199	473	344	1213	6036	64	711	3891	6	671	820	275	2357	7654	1149	981	5070	812	924	19	128	41	364	779	16	3016	170	901	7	275	
г	1182	28	170	40	1286	361	9	60	767	0	115	1469	72	406	9083	184	707	175	45	536	5	7	0	54	4	0	0	0	0	8	13	13
д	5759	63	1100	32	96	5100	30	58	2718	0	321	709	121	2220	4417	192	2062	482	327	2228	5	51	284	98	95	2	504	1131	25	11	478	
е	295	2400	3963	4142	4366	2329	1151	2081	1575	2502	2175	6796	5858	9861	1656	3635	7507	7995	7532	776	81	865	403	2014	1163	1028	0	0	429	273	500	
ж	1368	56	52	10	897	4894	12	14	1515	0	187	28	19	919	100	53	15	62	46	327	1	1	1	24	0	0	0	56	38	2	27	
з	5131	182	1163	404	903	292	65	66	485	0	252	296	361	2158	671	174	306	173	111	708	5	9	17	61	11	3	345	151	25	2	466	
и	393	1544	4937	993	3004	2675	449	2699	1880	1369	3826	5604	3580	5869	1627	2339	1591	4643	5823	584	93	2267	991	2475	735	171	0	0	237	411	1891	
й	122	247	607	249	690	96	122	159	646	0	534	353	396	984	413	795	348	1345	748	173	54	70	74	515	260	4	0	95	10	69		
к	7823	543	747	84	235	551	168	104	3035	2	252	673	194	1146	10576	393	1854	753	1058	1475	12	75	22	205	31	4	0	0	100	1	105	
л	6159	263	712	270	309	4996	358	181	6794	0	662	165	159	956	7094	665	394	2158	410	1783	24	35	5	579	24	3	639	5427	102	889	1634	
м	3639	292	850	271	474	4637	156	224	3775	1	446	217	280	2478	4780	987	354	1067	501	2717	52	74	19	394	44	5	994	91	121	9	567	
н	11644	255	394	158	524	11247	28	153	9650	0	520	64	131	3455	11568	526	320	894	762	3169	50	54	270	351	30	99	3219	1288	53	230	1954	
о	257	5405	12271	5450	6863	2963	2841	1773	2335	3536	2977	7081	7142	9452	2129	4089	6424	9328	9098	983	305	751	225	3509	1136	275	0	0	419	653	1247	
п	919	19	4	8	4	3328	0	3	1121	0	79	855	1	131	10344	41	7784	87	77	855	0	2	16	25	7	0	239	171	0	0	595	
р	8803	225	503	268	376	6361	291	71	5961	1	277	79	380	898	8457	299	42	276	790	2922	301	79	23	119	267	17	987	1032	9	209	1122	
с	2090	250	1354	113	448	4943	234	97	1795	1	4999	3202	1407	1470	3524	2348	352	1186	12215	2015	22	267	70	501	112	4	253	3341	80	275	4021	
т	6718	395	3321	111	449	6908	120	171	4515	0	729	314	270	1874	16987	620	3839	1834	436	2176	23	55	80	633	28	33	1537	7218	134	92	540	
у	253	1071	1461	1696	2562	345	2040	447	619	227	1059	1674	2039	1430	433	1558	833	1852	2113	144	44	525	21	1358	677	366	0	0	108	903	218	
ф	269	1	2	1	3	88	0	0	384	0	10	16	19	1	145	7	128	25	17	115	10	0	0	2	2	0	23	107	0	0	1	
х	545	132	505	131	226	399	36	85	895	0	182	252	181	418	2749	444	204	379	196	275	13	18	9	95	45	2	0	0	57	1	35	
ц	596	11	63	9	17	935	2	10	344	4	44	4	16	34	332	64	11	40	33	228	4	1	0	17	4	0	194	10	6	1	9	
ч	2858	21	65	15	28	4531	1	11	1848	0	571	30	22	1037	115	58	117	38	4477	741	0	7	0	11	147	0	0	261	5	0	10	
ш	989	5	16	2	4	2384	1	2	1553	0	499	573	14	462	217	6	4	5	54	226	3	0	1	1	1	0	0	648	0	1	2	
щ	453	0	1	0	0	1688	0	0	689	0	0	0	0	0	55	2	0	4	0	170	0	0	0	0	0	0	0	0	41	0	0	
м	54	360	1399	199	425	1044	73	210	465	1426	382	2140	1327	712	390	517	364	1231	1127	134	6	1110	7	323	526	14	0	0	91	1	54	
я	260	489	1300	305	584	959	97	484	1304	0	2094	192	771	7940	864	1145	291	2035	911	271	47	269	99	678	357	33	0	0	214	384	963	
э	1	1	7	10	7	0	0	1	0	16	64	9	7	24	1	41	3	11	3047	0	9	43	0	2	0	0	0	0	13	0	0	
ю	80	420	291	75	517	58	87	86	284	1	243	71	179	307	306	352	200	538	606	77	20	30	25	270	80	282	0	0	51	43	60	
я	305	438	1780	238	1172	441	286	583	1133	95	759	1059	762	2024	798	1262	544	1932	2489	319	41	372	110	583	56	137	0	0	176	119	239	

Біграми без пробілів неперекривні:

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	м	ь	э	ю	я	
а	170	636	2610	543	1526	958	984	2216	824	461	3605	4177	2289	3010	638	1357	1369	3567	3003	322	61	626	51	888	524	155	0	0	157	518	1249	
б	340	1	42	0	13	1169	2	7	511	0	72	452	26	183	1166	6	702	135	11	740	0	24	0	19	4	136	2037	107	49	2	350	
в	3305	96	237	172	604	3047	33	354	1854	4	337	386	140	1171	3850	553	495	2548	424	455	9	62	21	181	363	10	1491	78	138	2	131	
г	595	12	81	16	642	175	7	28	384	7	55	377	39	196	4607	36	358	83	19	257	3	1	0	28	1	0	0	0	7	8	5	
д	2822	20	531	14	57	2526	10	30	1343	0	161	344	59	1096	2212	86	1050	253	163	1120	3	29	150	46	51	1	238	553	12	5	213	
е	156	1213	2049	2402	2206	1179	591	1032	751	1241	1076	3462	3033	4964	816	1817	3669	3814	3787	404	45	418	197	1049	593	532	0	0	216	135	251	
ж	700	19	29	5	469	2368	5	7	735	0	115	13	10	469	42	18	5	32	24	161	1	0	0	14	0	0	0	0	32	21	0	13
з	2539	88	597	212	464	187	24	31	245	0	125	152	191	1127	321	90	144	85	66	364	3	3	10	25	4	2	174	66	12	1	235	
и	193	798	2504	486	1513	1346	222	1334	952	669	1872	2776	1774	2826	803	1183	811	2366	2956	316	37	1104	509	1216	368	85	0	0	111	196	965	
й	61	115	304	106	342	47	70	98	326	0	281	178	217	470	209	393	176	685	388	88	23	38	34	253	122	2	0	0	37	4	35	
к	3926	274	383	41	117	258	97	47	1517	0	121	345	91	542	5341	197	959	381	541	692	5	40	15	106	16	2	0	0	48	1	46	
л	2078	131	350	147	164	2514	185	92	3363	0	319	85	76	467	2489	339	197	1049	198	892	12	20	3	284	13	2	317	2687	47	463	791	
м	1868	138	439	156	239	2302	77	110	1862	1	214	116	146	1196	2368	472	183	515	261	1327	23	37	8	194	27	4	483	41	59	7	275	
н	5788	121	286	79	251	5639	16	68	4843	0	260	35	76	1739	5826	267	153	453	362	1561	21	25	133	163	15	55	1624	639	22	111	937	
о	123	2726	6079	2737	3459	1458	1468	875	1107	1809	1490	3498	3538	4797	1037	2063	5193	4602	4540	450	153	375	119	1776	597	141	0	0	211	330	607	
п	453	11	2	5	2	1674	0	2	555	0	43	440	0	66	5223	19	3890	46	36	436	0	0	7	15	5	0	122	85	0	0	301	
р	4403	108	253	126	177	3159	154	31	3068	1	143	41	191	466	4223	144	20	150	384	1429	155	42	12	55	134	10	485	538	4	108	549	
с	1025	122	976	335	1127	2753	896	37	2453	19	1599	4779	1774	703	1727	513	113	1274	1772	1510	669	37	134	207	14	37	146	27	35	140	225	
т	3332	199	1697	59	221	3441	62	89	2273	0	361	169	121	945	8405	321	1789	901	232	1095	11	282	11	321	11	792	3615	69	41	262	391	
у	141	519	710	872	1304	274	190	212	317	115	511	843	1001	705	237	777	433	963	1103	63	25	281	11	710	333	187	0	0	57	440	111	
ф	184	1	0	0	44	0	0	191	0	6	9	10	1	71	5	72	12	10	67	3	0	0	1	2	0	0	0	0	79	0	0	10
х	270	37	57	290	113	290	113	45	259	0	216	1381	98	216	1381	98	216	1381	98	216	1381	98	216	1381	98	216	1381	98	216	1381	98	216
ц	287	5	32	5	9	493	0	0	177	0	21	2	8	19	118	30	5	23	14	112	2	0	0	8	2	0	0	98	0	3	1	4
ч	1392	14	26	9	16	2245	0	4	937	0	274	274	17	11	500	58	30	56	19	2220	363	0	5	0	0	3	70	0	0	131	4	0
ш	516	2	9	1	2	1189	1	2	777	0	236	300	5	246	98	2	3	2	24	111	1	0	1	0	0	0	0	0	314	0	1	6
щ	213	5	0	0	815	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
м	27	167	722	103	208	551	37	98	228	718	188	1094	652	357	143	245	167	576	568	69	1	584	3	158	258	17	0	0	58	1	25	1
ь	125	237	630	146	301	474	51	251	646	0	1071	90	381	1994	417	588	141	1004	465	131	21	144	63	343	171	17	0	0	114	203	500	
э	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ю	39	206	146	37	269	32	44	39	146	0	37	82	159	100	174	305	115	254	303	39	13	13	13	124	38	194	0	0	18	154	229	
я	142	224	895	141	594	226	151	312	563	46	377	561	417	1015	400	602	262	984	1218	150	29	183	45	312	30	30	0	0	85	59	30	0

Лабораторная работа №1

Произвольная часть текста:
апельсина_совсем_пр

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:
 $1,98287088738314 < H < 2,7777951914117$

Двоичная таблица угаданных символов:

01000000000000000000000000000000
00000001000000000000000000000000
00100000000000000000000000000000
00000000000000000100000000000000
10000000000000000000000000000000

Вероятности:

q[1] = 0,46
q[2] = 0,14
q[3] = 0,1
q[4] = 0,04
q[5] = 0,04
q[6] = 0
q[7] = 0,02
q[8] = 0,02
q[9] = 0,02
q[10] = 0,02
q[11] = 0
q[12] = 0,02
q[13] = 0,02
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0,04
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] = 0,04
q[30] = 0
q[31] = 0,02
q[32] = 0

Строка состояния:

Для 20 символов энтропия уменьшается через повтор символов.

Лабораторная работа №1

Произвольная часть текста:
а_разделяют_любопытную_идею_о_том_что_они_должны_вести_себя_определенным_об

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: _ (пробел)

Символ по счету: 1

Номер эксперимента: 51

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
 $1,27061792149046 < H < 1,99438444742261$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
10000000000000000000000000000000
00000000000000000000000000000001
01000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

q[1] = 0,6078431
q[2] = 0,0784313
q[3] = 0,1568627
q[4] = 0,0196078
q[5] = 0,0196078
q[6] = 0,0392156
q[7] = 0,0196078
q[8] = 0,0196078
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0,019607
q[15] = 0
q[16] = 0
q[17] = 0
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] = 0,019607

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

Для 30 символов энтропия еще уменьшается, поскольку больший блок охватывает больше символов, которые легче предсказать.

Розрахунок ентропії для блоків довжини 10, 20 і 30 символів дозволяє оцінити ступінь випадковості тексту на різних масштабах. Зі збільшенням довжини блоку спостерігається **зменшення ентропії**, що вказує на прояв закономірностей у тексті. Виконання не менше 50 експериментів гарантує статистично надійні результати. Порівняння з CoolPinkProgram дозволяє перевірити коректність власних обчислень. Таку саму закономірність бачимо і в H1 та H2. $H_1 > H_2$ блоки із двох символів менш випадкові, бо з'являються повторювані комбінації.

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

Аналогічно до формули з першого пункту розрахуємо надлишковість:

$$|A| = 32, \log 32 = 5$$

$$2.7919919623603 < H < 3.28153501408779 \\ 0.500802720470174 > R(10) > 0.372661087473526$$

$$1.98287088738314 < H < 2.7777951914117 \\ 0.583395773390644 > R(20) > 0.474633931031182$$

$$1.27061792149046 < H < 1.9943844742261 \\ 0.378946628238864 > R(30) > 0.3020768941536886$$

Висновки.

Аналізуючи результати з першого пункту можна помітити, що H1 та H2 для тексту без пробілів є вищими, що можна пояснити тим, що при вилученні пробіла, який є найчастішим символом, довжина тексту зменшується, а ймовірності стають більш рівномірно розподіленими, що в свою чергу збільшує ентропію. Якщо подивитись на надлишковість, то R1 та R2 навпаки для тексту з пробілами вища, що свідчить про те, що наявність пробілу підвищує загальну надлишковість.

Аналізуючи результати роботи з CoolPinkProgram можна побачити, що ентропії вийшли меншого розміру, ніж ентропії з пробілами з першого пункту, а надлишковості більше. Це може свідчити про те, що надлишковість зростає зі збільшенням порядку моделі.

Загалом, виконання комп'ютерного практикуму допомогло нам узагальнити поняття ентропії та надлишковості. Ми навчились на практиці визначати частоти n-грам, розраховувати значення ентропії, надлишковості.