Експериментальна оцінка ентропії на символ джерела відкритого тексту

1 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.

2 Хід роботи

Для виконання поставленого завдання, після короткого аналізу, я вирішив розбити його на три різних частини:

- робота з файлами та вхідним текстом
- робота з літерами
- робота з біграмами

2.1 Робота з файлами та вхідним текстом

Не буду вдаватись в деталі реалізації, так як це не є настільки важливим в даній роботі, лише хочу зазначити, що вхідний текст має назву `boloto.txt`, цей же текст, але вже опрацьований препроцесором має назву `boloto_processed.txt`, а вже оброблений файл та ще й без пробілів має назву `boloto_without_spaces.txt`.

2.2 Робота з літерами

В даній секції ми розв'язуємо декілька задач:

- підрахунок загальної к-ті літер у тексті

```
fn letters_count(letter_frequencies: &HashMap<char, i64>) -> i64
    {
    let mut count = 0;
    for (_key, _value) in letter_frequencies {
        count += _value;
    }
    count
}
```

- підрахунок к-ті кожної літери

```rust
fn get_letter_frequency(text: &str) -> HashMap<char, i64> {
    let mut frequencies: HashMap<char, i64> = HashMap::new();

    for c in text.chars() {
        *frequencies.entry(c).or_insert(0) += 1;
    }

    frequencies
}
```

- підрахунок ймовірності зустірти кожну окрему літеру

```rust
fn count_letters_probabilities(letter_frequencies: &HashMap<char, i64>) -> HashMap<char, f64> {
    let mut probabilities: HashMap<char, f64> = HashMap::new();
    let number_of_characters = letters_count(letter_frequencies) as f64;

    for (_key, _value) in letter_frequencies {
        probabilities.insert(*_key, (*_value as f64) / number_of_characters);
    }
    probabilities
}
```

- вивід знайдених значень

```rust
fn print_letters_probabilities(probabilities: &HashMap<char, f64>) {
    let mut sorted_probabilities: Vec<(&char, &f64)> = probabilities.iter().collect();
    sorted_probabilities.sort_by(|a, b| b.1.partial_cmp(a.1).unwrap());
    for (&letter, &probability) in sorted_probabilities {
        println!("{}: {}", letter, probability);
    }
    println!();
}

fn print_letter_frequencies(letter_frequencies: &HashMap<char, i64>) {
    let mut sorted_frequencies: Vec<(&char, &i64)> = letter_frequencies.iter().collect();
    sorted_frequencies.sort_by_key(|&(_, frequency)| *frequency);
    for (&letter, &frequency) in sorted_frequencies.iter().rev() {
        println!("{}: {}", letter, frequency);
    }
    println!();
}
```

2.3 Робота з біграмами

Аналогічні задачі доводиться розв'язувати і у випадку з біграмами, правда реалізація буде досить сильно відрізнятись.

- підрахунок загальної к-ті біграм у тексті

```
fn bigram_count(bigram_frequencies: &HashMap<String, i64>) -> i64 {
    let mut count = 0;
    for (_key, _value) in bigram_frequencies {
        count += _value;
    }
    count
}
```

- підрахунок к-ті кожної з біграм

```
fn get_bigram_frequency(text: &str) -> HashMap<String, i64> {
    let mut frequencies: HashMap<String, i64> = HashMap::new();

    let mut chars = text.chars().peekable();
    while let (Some(curr), Some(&next)) = (chars.next(), chars.peek()) {
        if curr.is_alphabetic() && next.is_alphabetic() {
            let bigram = format!("{}{}", curr.to_lowercase(), next.to_lowercase());
            *frequencies.entry(bigram).or_insert(0) += 1;
        } else if curr.is_alphabetic() && next.is_whitespace() {
            let bigram = format!("{} ", curr.to_lowercase());
            *frequencies.entry(bigram).or_insert(0) += 1;
        } else if curr.is_whitespace() && next.is_alphabetic() {
            let bigram = format!(" {}", next.to_lowercase());
            *frequencies.entry(bigram).or_insert(0) += 1;
        }
    }
    frequencies
}
```

- підрахунок ймовірності зустірти кожну окрему біграму

```
fn count_bigram_probabilities(bigram_frequencies: &HashMap<String, i64>) -> HashMap<String, f64> {
    let mut probabilities: HashMap<String, f64> = HashMap::new();
    let number_of_bigrams = bigram_count(bigram_frequencies) as f64;

    for (_key, _value) in bigram_frequencies {
        probabilities.insert(_key.clone(), (*_value as f64) / number_of_bigrams);
    }

    probabilities
}
```

- вивід знайдених значень

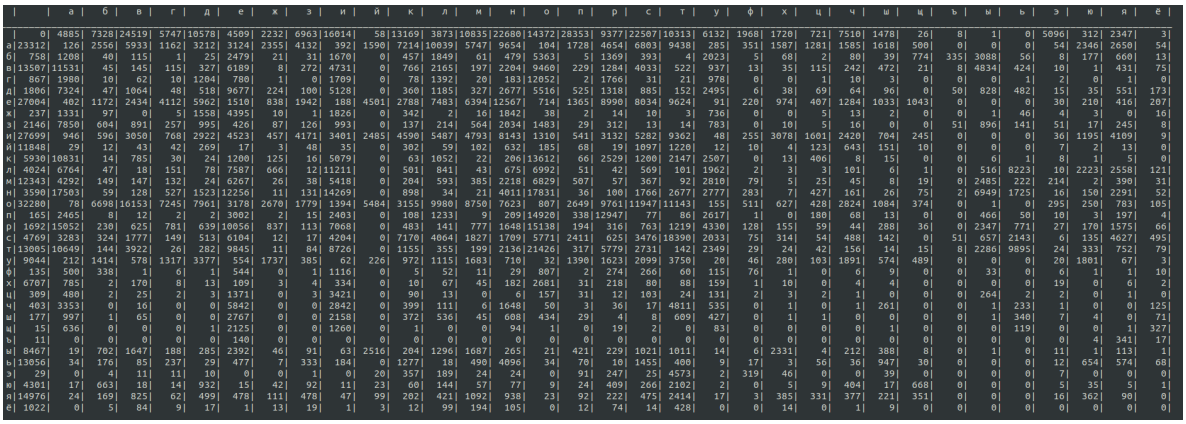
3 Результати роботи

3.1 Текст з пробілами

3.1.1 Робота з літерами

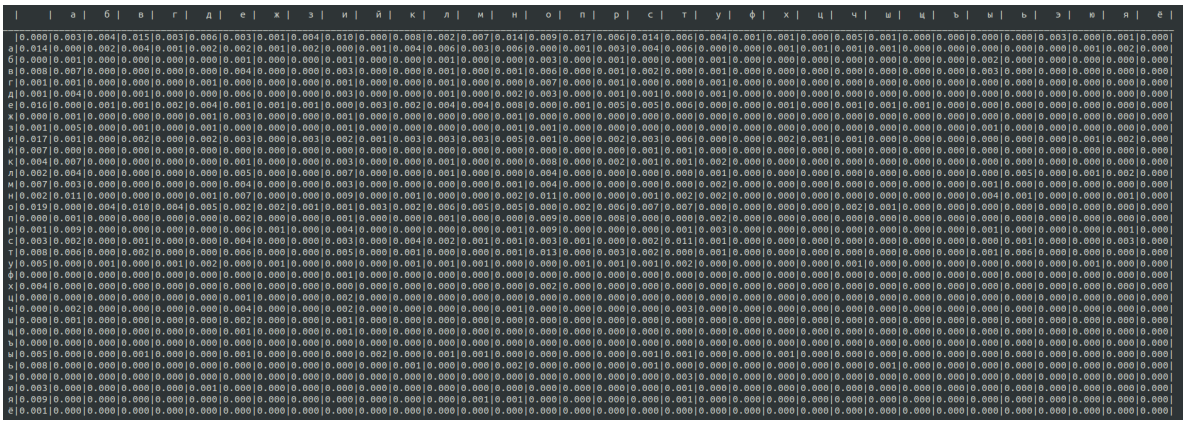
Літера	Кількість	Ймовірність
’ ’	248583	0.0149
а	115586	0.06929
б	23656	0.01418
в	65348	0.03918
г	23164	0.01389
д	43783	0.02625
е	113939	0.06831
ё	2136	0.00128
ж	12555	0.00753
з	21398	0.01283
и	107358	0.06436
й	17005	0.01019
к	47901	0.02872
л	56102	0.03363
м	46423	0.02783
н	92312	0.05534
о	158172	0.09482
п	41612	0.02495
р	67768	0.04063
с	77815	0.04665
т	96940	0.05811
у	37574	0.02253
ф	4406	0.00264
х	11742	0.00704
ц	6460	0.00387
ч	20695	0.01241
ш	9662	0.00579
щ	4685	0.00281
ы	25661	0.01538
ь	24874	0.01491
э	6029	0.00361
ю	10427	0.00625
я	25800	0.01547
ъ	513	0.00031

3.1.2 Робота з біграмами



	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	ъ	ы	ь	э	ю	я	ё
	0	4885	7328	24519	5747	10578	4509	2232	6963	16014	58	13169	3873	10835	22680	14372	28353	9377	22507	10313	6132	1968	1720	721	7510	1478	26	8	1	0	5096	312	2347	3
а	23312	126	2556	5933	1162	3212	3124	2355	4132	392	1590	7214	10039	5747	9654	104	1728	4654	6803	9438	285	351	1587	1281	1585	1618	500	0	0	0	54	2346	2650	54
б	758	1208	40	115	1	25	2479	21	31	1670	0	457	1849	61	479	5363	5	1369	393	4	2023	5	68	2	80	39	774	335	3088	56	8	177	660	13
в	13507	11531	45	145	115	327	6189	8	272	4731	0	766	2165	197	2204	9460	229	1284	4033	522	937	13	35	115	242	472	21	8	4834	424	10	1	431	75
г	867	1980	10	62	10	1204	780	1	0	1709	0	78	1392	20	183	1285	2	1766	31	21	978	0	0	1	10	3	0	0	0	1	2	0	1	0
д	1806	7324	47	1064	48	518	9677	224	100	5128	0	360	1185	327	2677	5516	525	1318	885	152	2495	6	38	69	64	96	0	50	828	482	15	35	551	173
е	27004	402	1172	2434	4112	5962	1510	838	1942	188	4501	2788	7483	6394	12567	714	1365	8990	8034	9624	91	220	974	407	1284	1033	1043	0	0	0	0	30	210	416	207
ж	237	1331	97	0	5	1558	4395	10	1	1826	0	342	2	16	1842	38	2	14	10	3	736	0	0	5	13	2	0	0	1	46	4	3	0	16
з	2146	7850	604	891	257	995	426	87	126	993	0	137	214	564	2034	1483	29	312	13	14	783	0	10	5	16	0	0	51	896	141	51	17	245	8
и	27699	946	596	3050	768	2922	4523	457	4171	3401	2485	4590	5487	4793	8143	1310	541	3132	5282	9362	48	255	3078	1601	2420	704	245	0	0	0	36	1195	4109	9
й	11848	29	12	43	42	269	17	3	48	35	0	302	59	102	632	185	68	19	1097	1220	12	10	4	123	643	151	10	0	0	0	7	2	13	0
к	5930	10831	14	785	30	24	1200	125	16	5079	0	63	1052	22	206	13612	66	2529	1200	2147	2507	0	13	406	8	15	0	0	6	1	8	1	5	0
л	4024	6764	47	18	151	78	7587	666	12	11211	0	501	841	43	675	6992	51	42	569	101	1962	2	3	3	101	6	1	0	516	8223	10	2223	2558	121
м	12343	4292	149	147	132	24	6267	26	38	5418	0	204	593	385	2218	6829	507	57	367	92	2810	79	5	25	45	8	19	0	2485	222	214	2	390	31
н	3590	17503	59	128	527	1523	12256	11	131	14269	0	898	34	21	4011	17831	36	100	1766	2677	2777	283	7	427	161	26	75	2	6949	1725	16	150	2291	52
о	32280	78	6698	10153	7245	7961	3178	2670	1779	1394	5484	3155	9900	8750	7623	807	2649	9761	11947	11143	155	511	627	428	2824	1084	374	0	1	0	295	250	783	105
п	165	2465	8	12	2	2	3002	2	15	2403	0	108	1233	9	209	14920	338	12947	77	86	2617	1	0	180	68	13	0	0	466	50	10	3	197	4
р	1692	15052	230	625	781	639	10056	837	113	7068	0	483	141	777	1648	15138	194	316	763	1219	4330	128	155	59	44	288	36	0	2347	771	27	170	1575	66
с	4769	3283	324	1777	149	513	6104	12	17	4204	0	7170	4064	1827	1709	5771	2411	625	3476	18390	2033	75	314	54	488	142	0	51	657	2143	6	135	4627	495
т	13085	10649	144	3922	26	282	9845	11	84	8726	0	1155	355	199	2136	21426	317	5779	2731	142	2349	29	24	42	156	14	15	8	2286	9895	24	333	752	79
у	9044	212	1414	578	1317	3377	554	1737	385	62	226	972	1115	1683	710	32	1390	1623	2099	3750	20	46	280	103	1891	574	489	0	0	0	20	1801	67	3
ф	135	500	338	1	6	1	544	0	1	1116	0	5	52	11	29	807	2	274	266	60	115	76	1	0	6	9	0	0	33	0	6	1	1	10
х	6707	785	2	170	8	13	109	3	4	334	0	10	67	45	182	2681	31	218	80	88	159	1	10	0	4	4	0	0	0	0	19	0	6	2
ц	309	480	2	25	2	3	1371	0	3	3421	0	90	13	0	6	157	31	12	103	24	131	2	3	2	1	0	0	0	264	2	2	0	1	0
ч	403	3353	0	16	0	0	5842	0	0	2842	0	399	111	6	1648	50	3	36	17	4811	535	0	1	0	1	261	0	0	1	233	1	0	0	125
ш	177	997	1	65	0	0	2767	0	0	2158	0	372	536	45	608	434	29	4	8	609	427	0	1	1	0	0	0	0	1	340	7	4	0	71
щ	15	636	0	0	0	1	2125	0	0	1260	0	1	0	0	94	1	0	19	2	0	83	0	0	0	0	1	0	0	0	119	0	0	1	327
ъ	11	0	0	0	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	341	17
ы	8467	19	702	1647	188	285	2392	46	91	63	2516	204	1296	1687	265	21	421	229	1021	1011	14	6	2331	4	212	388	8	0	1	0	11	1	113	1
ь	13056	34	176	85	237	29	477	7	333	184	0	1277	18	490	4096	34	70	10	1455	400	9	17	3	56	36	947	30	0	0	0	12	654	574	68
э	29	0	4	11	11	10	0	0	1	0	20	357	109	24	24	0	91	247	25	4573	2	319	46	0	0	39	0	0	0	0	7	0	0	0
ю	4301	17	663	18	14	932	15	42	92	11	23	60	144	57	77	9	24	409	266	2102	2	0	5	9	404	17	668	0	0	0	5	35	5	1
я	14976	24	169	825	62	499	478	111	478	47	99	202	421	1092	938	23	92	222	475	2414	17	3	385	331	377	221	351	0	0	0	16	362	90	0
ё	1022	0	5	84	9	17	1	13	19	1	3	12	99	194	105	0	12	74	14	428	0	0	14	0	1	9	0	0	0	0	0	0	0	0

Рис. 1: к-ть біграм



	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	ъ	ы	ь	э	ю	я	ё
	0.000	0.003	0.004	0.015	0.003	0.006	0.003	0.001	0.004	0.010	0.000	0.008	0.002	0.007	0.014	0.009	0.017	0.006	0.014	0.006	0.004	0.001	0.001	0.000	0.005	0.001	0.000	0.000	0.000	0.000	0.003	0.000	0.001	0.000
а	0.014	0.000	0.002	0.004	0.001	0.002	0.002	0.001	0.002	0.000	0.001	0.004	0.006	0.003	0.006	0.000	0.001	0.003	0.004	0.006	0.000	0.000	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.000
б	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.003	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000
в	0.008	0.007	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.003	0.000	0.000	0.001	0.000	0.001	0.006	0.000	0.001	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000
г	0.001	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.007	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
д	0.001	0.004	0.000	0.001	0.000	0.000	0.006	0.000	0.000	0.003	0.000	0.001	0.000	0.002	0.003	0.000	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
е	0.016	0.000	0.001	0.001	0.002	0.004	0.001	0.001	0.001	0.000	0.003	0.002	0.004	0.004	0.008	0.000	0.001	0.005	0.005	0.006	0.000	0.000	0.001	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ж	0.000	0.001	0.000	0.000	0.000	0.001	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
з	0.001	0.005	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
и	0.017	0.001	0.000	0.002	0.000	0.002	0.003	0.000	0.003	0.002	0.001	0.003	0.003	0.003	0.005	0.001	0.000	0.002	0.003	0.006	0.000	0.000	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.000
й	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
к	0.004	0.007	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.003	0.000	0.000	0.001	0.000	0.000	0.006	0.000	0.002	0.001	0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
л	0.002	0.004	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.007	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.001	0.002	0.000
м	0.007	0.003	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.001	0.004	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
н	0.002	0.011	0.000	0.000	0.000	0.001	0.007	0.000	0.000	0.009	0.000	0.001	0.000	0.000	0.002	0.011	0.000	0.000	0.001	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.001	0.000	0.000	0.001	0.000
о	0.019	0.000	0.004	0.010	0.004	0.005	0.002	0.002	0.001	0.001	0.003	0.002	0.006	0.005	0.005	0.000	0.002	0.006	0.007	0.007	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
п	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
р	0.001	0.009	0.000	0.000	0.000	0.000	0.006	0.001	0.000	0.004	0.000	0.000	0.000	0.000	0.001	0.009	0.000	0.000	0.000	0.001	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000
с	0.003	0.002	0.000	0.001	0.000	0.000	0.004	0.000	0.000	0.003	0.000	0.004	0.002	0.001	0.001	0.003	0.001	0.000	0.002	0.011	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.003	0.000
т	0.008	0.006	0.000	0.002	0.000	0.000	0.006	0.000	0.000	0.005	0.000	0.001	0.000	0.000	0.001	0.013	0.000	0.003	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.006	0.000	0.000	0.000	0.000
у	0.005	0.000	0.001	0.000	0.001	0.002	0.000	0.001	0.000	0.000	0.000	0.001	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000
ф	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
х	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ц	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ч	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ш	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
щ	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ъ	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ы	0.005	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.002	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ь	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.002	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
э	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ю	0.003	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
я	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ё	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

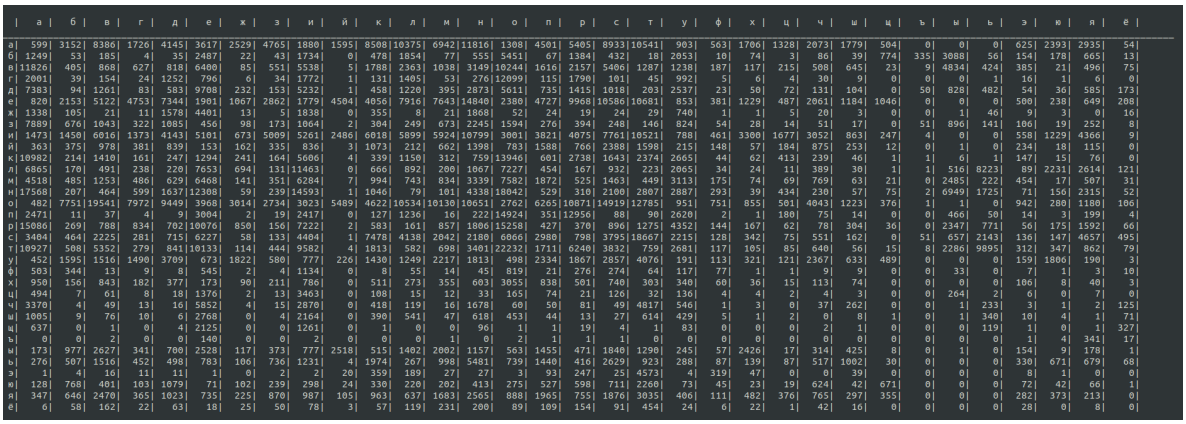
Рис. 2: ймовірність появи біграм

3.2 Текст без пробілів

3.2.1 Робота з літерами

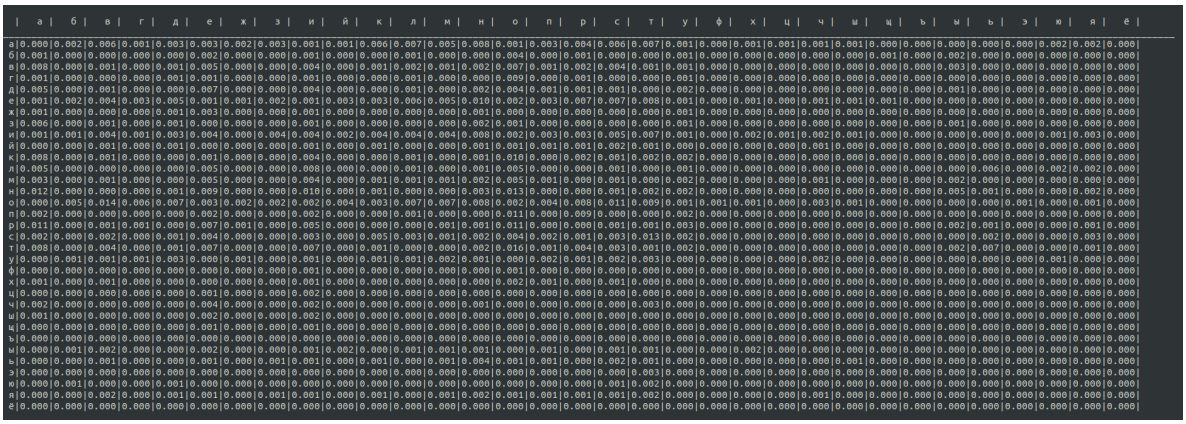
Літера	Кількість	Ймовірність
а	115586	0.08143
б	23656	0.01667
в	65348	0.04604
г	23164	0.01632
д	43783	0.03084
е	113939	0.08027
ё	2136	0.00150
ж	12555	0.00884
з	21398	0.01507
и	107358	0.07563
й	17005	0.012
к	47901	0.03374
л	56102	0.03952
м	46423	0.0327
н	92312	0.06503
о	158172	0.11143
п	41612	0.02931
р	67768	0.04774
с	77815	0.05482
т	96940	0.06829
у	37574	0.02647
ф	4406	0.0031
х	11742	0.00827
ц	6460	0.00455
ч	20695	0.01458
ш	9662	0.00681
щ	4685	0.0033
ы	25661	0.01808
ь	24874	0.01752
э	6029	0.00425
ю	10427	0.00735
я	25800	0.01818
ъ	513	0.00036

3.2.2 Робота з біграмами



	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	ъ	ы	ь	э	ю	я	ё
а	599	3152	8386	1726	4145	3617	2529	4765	1880	1595	8508	10375	6942	11816	1308	4501	5405	8933	10541	903	563	1706	1328	2073	1779	504	0	0	0	625	2393	2935	54
б	1249	53	185	4	35	2487	22	43	1734	0	478	1854	77	555	5451	67	1384	432	18	2053	10	74	3	86	39	774	335	3088	56	154	178	665	13
в	11826	405	868	627	818	6400	85	551	5538	5	1788	2363	1038	3149	10244	1616	2157	5406	1287	1238	187	117	215	508	645	23	9	4834	424	385	21	496	75
г	2001	39	154	24	1252	796	6	34	1772	1	131	1405	53	276	12099	115	1790	101	45	992	5	6	4	30	9	0	0	0	1	16	1	6	0
д	7383	94	1261	83	583	9708	232	153	5232	1	458	1220	395	2873	5611	735	1415	1018	203	2537	23	50	72	131	104	0	50	828	482	54	36	585	173
е	820	2153	5122	4753	7344	1901	1067	2862	1779	4504	4056	7916	7643	14840	2380	4727	9968	10586	10681	853	381	1229	487	2061	1184	1046	0	0	0	500	238	649	208
ж	1338	105	21	11	1578	4401	13	5	1838	0	355	8	21	1868	52	24	19	24	29	740	1	1	5	20	3	0	0	1	46	9	3	0	16
з	7889	676	1043	322	1085	456	98	173	1064	2	304	249	673	2245	1594	276	394	248	146	824	54	28	14	51	17	0	51	896	141	106	19	252	8
и	1473	1458	6016	1373	4143	5101	673	5009	5261	2486	6018	5899	5924	10799	3001	3821	4075	7761	10521	788	461	3300	1677	3052	863	247	4	0	0	558	1229	4366	9
й	363	375	970	381	839	153	162	335	836	3	1073	212	662	1398	783	1588	766	2388	1598	215	148	57	184	875	253	12	0	1	0	234	18	115	0
к	10982	214	1410	161	247	1294	241	164	5606	4	339	1150	312	759	13946	601	2738	1643	2374	2665	44	62	413	239	46	1	1	6	1	147	15	76	0
л	6865	170	491	238	220	7653	694	131	11463	0	666	892	200	1067	7227	454	167	932	223	2065	34	24	11	389	30	1	1	516	8223	89	2231	2614	121
м	4518	485	1253	486	629	6468	141	351	6284	7	994	743	834	3339	7582	1872	525	1463	449	3113	175	74	69	769	63	21	0	2485	222	454	17	507	31
н	17568	287	464	599	1637	12308	59	239	14593	1	1046	79	101	4338	18042	529	310	2100	2807	2887	293	39	434	230	57	75	2	6949	1725	71	156	2315	52
о	482	7751	19541	7972	9449	3968	3014	2734	3023	5489	4622	10534	10136	16651	2762	6265	10871	14919	12785	951	751	855	501	4043	1223	376	1	1	0	942	280	1180	106
п	2471	11	37	4	9	3004	2	19	2417	0	127	1236	16	222	14924	351	12956	88	90	2620	2	1	180	75	14	0	0	466	50	14	3	199	4
р	15086	269	788	834	702	10076	850	156	7222	2	583	161	857	1806	15258	427	370	896	1275	4352	144	167	62	78	304	36	0	2347	771	56	175	1592	66
с	3404	464	2225	281	715	6227	58	133	4404	1	7478	4138	2042	2180	6066	2980	798	3795	18667	2215	128	342	75	551	162	0	51	657	2143	136	147	4657	495
т	10927	508	5352	279	841	10133	114	444	9582	4	1813	582	698	3401	22232	1711	6240	3832	759	2681	117	105	85	640	56	15	8	2286	9895	312	347	862	79
у	452	1595	1516	1490	3709	673	1822	580	777	226	1430	1249	227	1813	498	2334	1867	2857	4076	191	113	321	121	2367	633	489	0	0	0	159	1806	190	3
ф	503	344	13	9	8	545	2	4	1134	0	8	55	14	45	819	21	276	274	64	117	77	1	1	9	9	0	0	33	0	7	1	3	10
х	950	156	843	182	377	173	90	211	786	0	511	273	355	603	3055	838	501	740	303	340	60	36	15	113	74	0	0	0	0	106	8	40	3
ц	494	7	61	8	18	1376	2	13	3463	0	108	15	12	33	165	74	21	126	32	136	4	4	2	4	3	0	0	264	2	6	0	7	0
ч	3370	4	49	13	16	5852	4	15	2870	0	418	119	16	1678	60	50	81	49	4817	546	1	3	0	37	262	0	0	1	233	3	1	2	125
ш	1005	9	76	10	6	2768	0	4	2164	0	390	541	47	618	453	44	13	27	614	429	5	1	2	0	8	1	0	1	340	10	4	1	71
щ	637	0	1	0	4	2125	0	0	1261	0	1	0	0	96	1	1	19	4	1	83	0	0	0	2	1	0	0	0	119	1	0	1	327
ъ	0	0	2	0	0	140	0	0	2	0	0	0	1	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	4	341	17
ы	173	977	2627	341	700	2528	117	373	777	2518	515	1402	2002	1157	563	1455	471	1840	1290	245	57	2426	17	314	425	8	0	1	0	154	9	178	1
ь	276	507	1516	452	498	783	106	736	1231	4	1974	267	998	5481	739	1440	416	2629	923	288	87	139	87	517	1002	30	0	0	0	330	671	679	68
э	1	4	16	11	11	1	0	2	2	20	359	189	27	27	1	93	247	25	4573	4	319	47	0	0	39	0	0	0	0	8	1	0	0
ю	128	768	401	103	1079	71	102	239	298	24	330	220	202	413	275	527	598	711	2260	73	45	23	19	624	42	671	0	0	0	72	42	66	1
я	347	646	2470	365	1023	735	225	870	987	105	963	637	1683	2565	888	1965	755	1876	3035	406	111	482	376	765	297	355	0	0	0	282	373	213	0
ё	6	58	162	22	63	18	25	50	78	3	57	119	231	200	89	109	154	91	454	24	6	22	1	42	16	0	0	0	0	28	0	8	0

Рис. 3: к-ть біграм



	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	ъ	ы	ь	э	ю	я	ё
а	0.000	0.002	0.006	0.001	0.003	0.003	0.002	0.003	0.001	0.001	0.006	0.007	0.005	0.008	0.001	0.003	0.004	0.006	0.007	0.001	0.000	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.002	0.000
б	0.001	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.000	0.000
в	0.008	0.000	0.001	0.000	0.001	0.005	0.000	0.000	0.004	0.000	0.001	0.002	0.001	0.002	0.007	0.001	0.002	0.004	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000
г	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.008	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
д	0.005	0.000	0.001	0.000	0.000	0.007	0.000	0.000	0.004	0.000	0.000	0.001	0.000	0.002	0.004	0.001	0.001	0.001	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
е	0.001	0.002	0.004	0.003	0.005	0.001	0.001	0.002	0.001	0.003	0.003	0.006	0.005	0.010	0.002	0.003	0.007	0.007	0.008	0.001	0.000	0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ж	0.001	0.000	0.000	0.000	0.001	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
з	0.006	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
и	0.001	0.001	0.004	0.001	0.003	0.004	0.001	0.004	0.004	0.002	0.004	0.004	0.004	0.008	0.002	0.003	0.005	0.007	0.001	0.001	0.000	0.002	0.001	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.000
й	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.002	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
к	0.008	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.000	0.001	0.000	0.001	0.010	0.000	0.002	0.001	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
л	0.005	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.008	0.000	0.000	0.001	0.000	0.001	0.005	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.002	0.002	0.000
м	0.003	0.000	0.001	0.000	0.000	0.005	0.000	0.000	0.004	0.000	0.001	0.001	0.001	0.002	0.005	0.001	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000
н	0.012	0.000	0.000	0.001	0.001	0.009	0.000	0.000	0.010	0.000	0.001	0.001	0.001	0.003	0.013	0.000	0.001	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.001	0.000	0.000	0.002	0.000
о	0.000	0.005	0.014	0.006	0.007	0.003	0.002	0.002	0.002	0.004	0.003	0.007	0.007	0.008	0.002	0.004	0.008	0.011	0.009	0.001	0.001	0.001	0.000	0.003	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.001	0.000
п	0.002	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0.011	0.000	0.009	0.000	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
р	0.011	0.000	0.001	0.001	0.000	0.007	0.001	0.000	0.005	0.000	0.000	0.000	0.001	0.001	0.011	0.000	0.000	0.001	0.001	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.001	0.000
с	0.002	0.000	0.002	0.000	0.001	0.004	0.000	0.000	0.003	0.000	0.005	0.003	0.001	0.002	0.004	0.002	0.001	0.003	0.013	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.003	0.000
т	0.008	0.000	0.004	0.000	0.001	0.007	0.000	0.000	0.007	0.000	0.001	0.000	0.000	0.002	0.016	0.001	0.004	0.003	0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.007	0.000	0.000	0.001	0.000
у	0.000	0.001	0.001	0.001	0.003	0.000	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.002	0.001	0.000	0.002	0.001	0.002	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ф	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
х	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ц	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ч	0.002	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ш	0.001	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
щ	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ъ	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ы	0.000	0.001	0.002	0.000	0.000	0.002	0.000	0.000	0.001	0.002	0.000	0.001	0.001	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ь	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.001	0.001	0.000	0.001	0.000	0.001	0.004	0.001	0.001	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
э	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ю	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
я	0.000	0.000	0.002	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.000	0.001	0.002	0.001	0.001	0.001	0.001	0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ё	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Рис. 4: ймовірність появи біграм

4 Обчислення ентропій H_1 та H_2

- Обчислення H_1:

```
fn compute_h1(letter_frequencies: &HashMap<char, i64>) -> f64 {
    let mut h1 = 0.0;
    let probabilities = count_letters_probabilities(&
   letter_frequencies);

    for (_key, _value) in probabilities {
        h1 += _value * f64::log2(_value);
    }

    h1 = -h1;
    h1
}
```

- Обчислення H_2:

```
fn compute_h2(bigram_frequencies: &HashMap<String, i64>) -> f64 {
    let mut h2 = 0.0;
    let probabilities = count_bigram_probabilities(&
   bigram_frequencies);

    for (_key, _value) in probabilities {
        h2 += _value * f64::log2(_value);
    }
    h2 = -h2/2.0;

    h2
}
```

Якщо у тексті наявні пробіли, то $H_1 = 4.404$, якщо ж їх немає, то $H_1 = 4.461$. Так само й для H_2, з пробілами $H_2 = 4.021$, без $H_2 = 4.152$.

5 Оціночні значення величин $H^{(10)}$, $H^{(20)}$ та $H^{(30)}$

Використовуючи програму CoolPinkProgram.exe визначаємо приблизні значення:

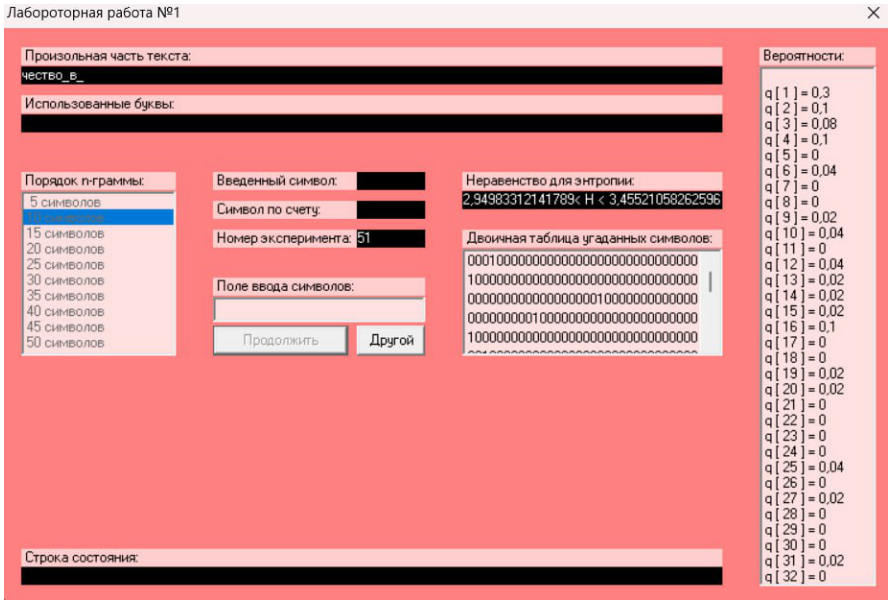


Рис. 5: значення $H^{(10)}$

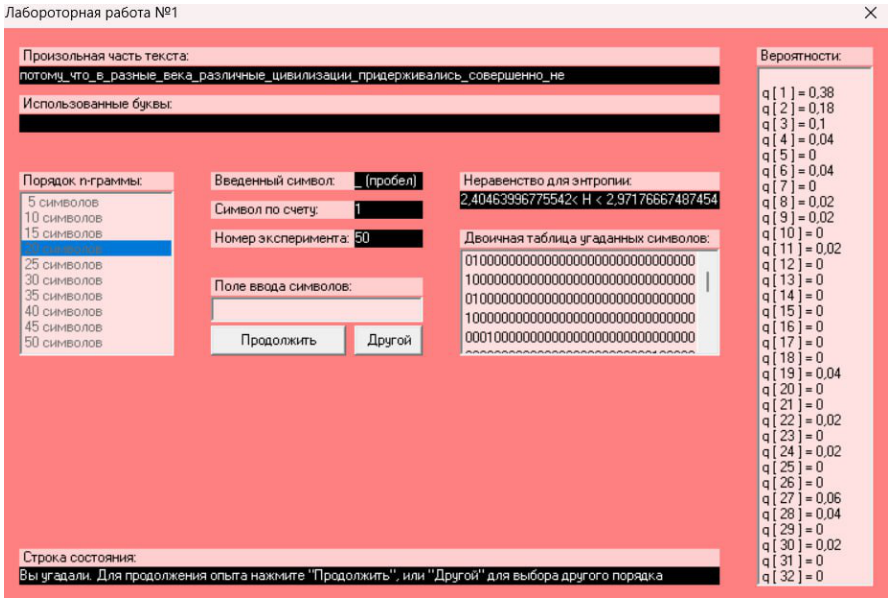


Рис. 6: значення $H^{(20)}$

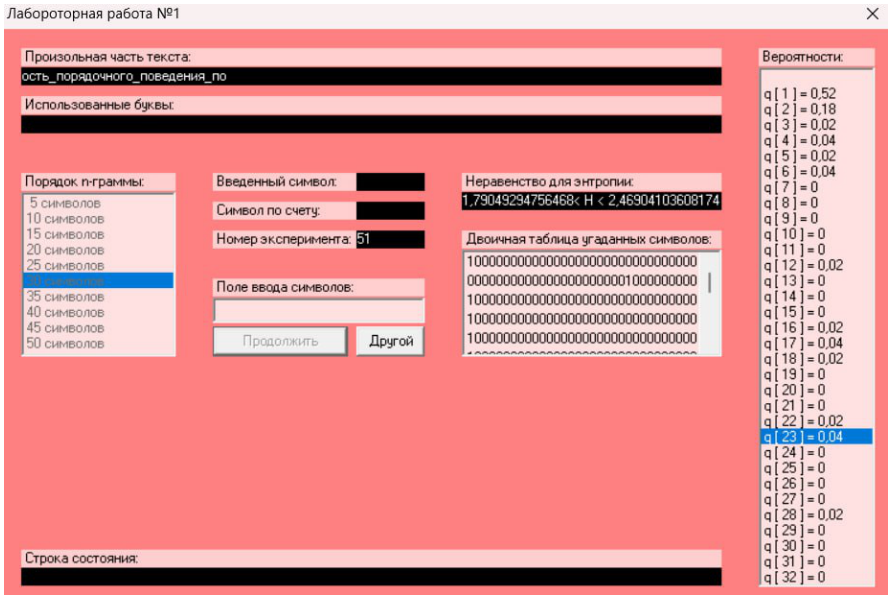


Рис. 7: значення $H^{(30)}$

Маємо, що:

$$a \leq H^{(10)} \leq c$$
$$a \leq H^{(20)} \leq c$$
$$a \leq H^{(30)} \leq c$$

Тоді можемо обчислити надлишковість $_p$осійської мови R використовуючи формулу:

$$R = 1 - \frac{H_{\text{inf}}}{H_0}, \quad \text{де } H_0 = \log 32 \approx 5$$

$$H^{(30)} : 0.642 \geq R \geq 0.506$$
$$R \sim 0.57$$

6 Висновок

З результатів обрахунків бачимо, що $_p$осійської мова є досить надлишковою, тому ми можемо ущільнювати тексти цієї мови майєе вдвічі без втрати змісту.