

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

```
1 fn letters_count(letter_frequencies: &HashMap<char, i64>) -> i64
2 {
3     let mut count = 0;
4     for (_key, _value) in letter_frequencies {
5         count += _value;
6     }
7     count
8 }
```

- підрахунок к-ті кожної літери

```

1 fn get_letter_frequency(text: &str) -> HashMap<char, i64> {
2     let mut frequencies: HashMap<char, i64> = HashMap::new();
3
4     for c in text.chars() {
5         *frequencies.entry(c).or_insert(0) += 1;
6     }
7
8     frequencies
9 }

```

- підрахунок ймовірності зустріти кожну окрему літеру

```

1 fn count_letters_probabilities(letter_frequencies: &HashMap<char,
2     i64>) -> HashMap<char, f64> {
3     let mut probabilities: HashMap<char, f64> = HashMap::new();
4     let number_of_characters = letters_count(letter_frequencies)
5     as f64;
6
7     for (_key, _value) in letter_frequencies {
8         probabilities.insert(*_key, (*_value as f64) /
9         number_of_characters);
10    }
11    probabilities
12 }

```

- вивід знайдених значень

```

1 fn print_letters_probabilities(probabilities: &HashMap<char, f64
2     >) {
3     let mut sorted_probabilities: Vec<(&char, &f64)> =
4     probabilities.iter().collect();
5     sorted_probabilities.sort_by(|a, b| b.1.partial_cmp(a.1).
6     unwrap());
7     for (&letter, &probability) in sorted_probabilities {
8         println!("{}", letter, probability);
9     }
10    println!();
11 }
12
13 fn print_letter_frequencies(letter_frequencies: &HashMap<char,
14     i64>) {
15     let mut sorted_frequencies: Vec<(&char, &i64)> =
16     letter_frequencies.iter().collect();
17     sorted_frequencies.sort_by_key(|&(_, frequency)| *frequency);
18     for (&letter, &frequency) in sorted_frequencies.iter().rev()
19     {
20         println!("{}", letter, frequency);
21     }
22     println!();
23 }

```

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

```

1 fn bigram_count(bigram_frequencies: &HashMap<String, i64>) -> i64
2 {
3     let mut count = 0;
4     for (_key, _value) in bigram_frequencies {
5         count += _value;
6     }
7     count
8 }

```

- підрахунок к-ті кожної з біграм

```

1 fn get_bigram_frequency(text: &str) -> HashMap<String, i64> {
2     let mut frequencies: HashMap<String, i64> = HashMap::new();
3
4     let mut chars = text.chars().peekable();
5     while let (Some(curr), Some(&next)) = (chars.next(), chars.
6     peek()) {
7         if curr.is_alphabetic() && next.is_alphabetic() {
8             let bigram = format!("{}", curr.to_lowercase(),
9             next.to_lowercase());
10            *frequencies.entry(bigram).or_insert(0) += 1;
11        } else if curr.is_alphabetic() && next.is_whitespace() {
12            let bigram = format!("{}", curr.to_lowercase());
13            *frequencies.entry(bigram).or_insert(0) += 1;
14        } else if curr.is_whitespace() && next.is_alphabetic() {
15            let bigram = format!("{}", next.to_lowercase());
16            *frequencies.entry(bigram).or_insert(0) += 1;
17        }
18    }
19    frequencies
20 }

```

- підрахунок ймовірності зустріти кожну окрему біграму

```

1 fn count_bigram_probabilities(bigram_frequencies: &HashMap<String
2 , i64>) -> HashMap<String, f64> {
3     let mut probabilities: HashMap<String, f64> = HashMap::new();
4     let number_of_bigrams = bigram_count(bigram_frequencies) as
5     f64;
6
7     for (_key, _value) in bigram_frequencies {
8         probabilities.insert(_key.clone(), (*_value as f64) /
9         number_of_bigrams);
10    }
11
12    probabilities
13 }

```

- 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	я	и	е		
а	0	4885	7328	24519	5747	10578	4599	2232	6963	16014	58	13169	3873	10835	22680	14372	28353	9377	22507	16313	6132	1968	1720	721	7510	1478	26	0	1	0	5096	312	2347	3	
б	22332	126	2556	5933	1162	3212	3124	2355	4321	992	1599	7214	10039	5747	9654	104	1728	4654	6803	9438	285	351	1587	1201	1585	1610	590	0	0	0	54	2346	2650	54	
в	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	50	8	177	660	13	
г	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	0	4834	424	10	1	431	75	
д	897	1980	19	62	10	1264	780	1	0	1769	0	78	1392	20	183	12052	2	1766	31	21	978	0	0	1	10	3	0	0	0	0	0	1	2	0	
е	1800	7324	47	1064	48	518	9677	224	100	5128	0	368	1185	327	2677	5516	525	1318	885	152	2495	0	38	69	64	96	0	50	828	482	15	35	551	173	
ж	2704	402	1172	2434	4122	5962	1510	838	1942	1188	4501	2788	7483	6394	12567	714	1365	8990	8034	9624	91	220	974	407	1284	1033	1043	0	0	0	30	210	416	207	
з	237	1331	97	0	5	1558	4395	10	1	1826	0	342	2	16	1842	38	2	14	18	3	736	0	0	5	13	2	0	0	0	1	46	4	3	0	
и	2146	7850	604	891	257	985	426	87	126	993	0	137	214	544	2034	1483	29	322	13	14	783	0	10	5	16	0	0	51	890	141	51	17	245	8	
й	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	
к	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	
л	5930	10831	14	785	30	24	1200	125	16	5079	0	63	1052	22	206	13612	66	2529	1280	2147	2507	0	13	406	8	15	0	0	0	0	0	1	5	0	
м	4624	6764	47	10	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	
н	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	
о	3590	17583	59	128	527	1523	12256	11	131	14269	0	898	34	21	4811	17831	36	180	1766	2677	2777	283	7	427	161	26	75	2	6949	1725	10	156	2291	52	
п	32280	78	6098	16153	7245	7961	3178	2670	1779	1104	5484	3158	9980	8780	7623	807	2649	9761	11947	11143	155	511	627	420	2824	1084	374	0	1	0	295	250	783	105	
р	165	2465	8	12	2	1	3002	2	15	2402	0	108	1233	9	209	14920	338	12947	77	86	2617	1	0	180	68	13	0	0	466	50	10	3	197	4	
с	1692	15952	236	625	781	639	10056	837	113	7668	0	483	141	77	1648	15138	194	316	763	1219	4338	120	155	59	44	288	36	0	2347	771	27	170	1575	66	
т	4769	3283	324	177	149	513	6104	12	37	4204	0	7170	4064	1827	1789	5771	2411	625	3476	18398	2032	75	314	54	488	142	0	51	657	2142	6	135	4627	495	
у	13865	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	0	2286	9895	24	333	752	79	
ф	9844	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	1881	67	3	
х	325	588	338	1	6	1	544	0	1	1116	0	5	52	11	29	807	2	274	266	88	115	76	1	0	0	9	0	0	33	0	6	1	1	0	
ц	6787	785	2	170	8	13	109	3	4	334	0	10	67	45	182	2681	31	218	88	88	159	1	10	4	4	0	0	0	0	19	0	6	2	0	
ч	309	480	2	25	2	3	1371	0	90	13	0	6	157	31	12	103	24	131	2	3	2	1	0	0	0	0	0	0	0	264	2	2	0	1	
ш	403	3353	0	16	0	0	5842	0	0	2842	0	399	111	0	1648	50	3	36	17	4811	535	0	1	0	1	261	0	0	1	233	1	0	0	125	
щ	177	977	1	65	0	0	2767	0	0	2158	0	372	536	45	688	434	29	4	8	0	427	0	1	1	1	0	0	0	1	340	7	4	0	71	
ъ	15	636	0	0	0	1	12125	0	0	1260	0	1	0	0	94	1	0	19	2	0	83	0	0	0	0	0	0	0	0	0	119	0	0	1	327
ы	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	1	4	341	17
ь	8467	191	702	1868	188	204	2392	46	0	0	0	2514	204	1298	168	26	121	220	102	1010	14	22	2331	3884	1	121	0	0	0	0	0	0	0	0	
э	13850	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	12	654	574	68	0	
я	29	0	4	11	11	10	0	0	1	0	0	20	357	189	24	24	0	91	247	25	4573	2	319	46	0	0	39	0	0	0	7	0	0	0	
и	4381	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	1	1	
е	14976	24	169	824	62	499	478	111	478	47	99	202	421	1892	938	23	92	222	475	2454	17	3	385	311	377	221	351	0	0	16	362	98	0	0	
и	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	я	и	е	
а	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
б	0.010	0.000	0.002	0.004	0.001	0.002	0.003	0.000	0.001	0.004	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.004	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.000	0.002	0.000	0.000	0.000	0.001	0.002	0.000	
в	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	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
г	0.008	0.007	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.003	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.002	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	
д	0.001	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	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	
е	0.001	0.004	0.000	0.001	0.000	0.000	0.006	0.000	0.000	0.003	0.000	0.000	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	
ж	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	0.001	0.000	0.000	0.000	0.000	0.000	
з	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	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
и	0.001	0.005	0.000	0.001	0.000	0.001	0.00	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	
й	0.017	0.000	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	0.001	0.002	0.000
к	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	0.001	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	
л	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	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
м	0.002	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	
н	0.002	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	
о	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	0.001	0.000	0.000	
п	0.002	0.011	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	0.011	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	
р	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	
с	0.001	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	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
т	0.001	0.009	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.004	0.000	0.000	0.000	0.000	0.001	0.009	0.000	0.000	0.000	0.000	0.001	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	
у	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	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	0.002	0.000	0.000	0.000	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	0.000	0.000	
х	0.005	0.000	0.001	0.000	0.001	0.002	0.000	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	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	
ц	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	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	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	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	0.002	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.001	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	
щ	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	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	
ы	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	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	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	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	
я	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

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	я	е		
а	599	3152	3886	1726	4145	3671	2529	4765	1880	1595	8508	10375	6942	11916	11380	4501	4045	8933	10541	903	563	1768	1328	2073	1779	504	0	0	0	625	2393	2935	54	
б	1249	53	185	1	35	2487	22	43	1734	0	7	184	1854	77	555	5451	0	7	1384	432	18	2053	10	74	3	86	39	774	335	3088	56	154	178	63
в	11826	485	86	627	818	6408	85	551	5538	5	1788	2303	1038	3149	10044	116	2157	5460	1287	1238	187	117	215	508	645	23	9	4834	424	385	21	496	75	
г	788	67	18	3	383	185	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
д	7383	94	1261	83	583	798	232	153	5232	1	458	1220	395	10773	5611	735	1545	1018	203	2537	23	50	72	131	104	1	0	50	828	482	54	36	585	173
е	820	2153	5122	4753	734	3191	10671	2862	1779	4504	4056	7916	7443	10440	2380	4727	9968	10586	16681	815	381	1229	487	2661	1184	1046	0	0	50	238	649	208	10	
ж	1328	195	21	11	1578	4401	133	3	1838	0	355	8	21	1808	52	24	19	1	240	740	31	1	5	20	9	1	0	1	40	3	9	1	61	
з	788	67	18	3	383	185	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
и	1473	1450	6016	1373	4143	5101	673	5089	5261	2486	6108	5899	5924	10799	3001	3821	4075	774	10521	788	461	3300	1677	3552	247	12	0	0	558	1229	1366	9	9	
й	3671	2529	4765	1880	1595	8508	10375	6942	11916	11380	4501	4045	8933	10541	903	563	1768	1328	2073	1779	504	0	0	0	0	0	0	0	0	0	0	0	0	
к	8508	10375	6942	11916	11380	4501	4045	8933	10541	903	563	1768	1328	2073	1779	504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
л	903	563	1768	1328	2073	1779	504	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	4501	4045	8933	10541	903	563	1768	1328	2073	1779	504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
н	4045	8933	10541	903	563	1768	1328	2073	1779	504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
о	8933	10541	903	563	1768	1328	2073	1779	504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
п	903	563	1768	1328	2073	1779	504	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	
р	504	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	
с	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	0	
т	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	0	
у	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	0	
ф	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	0	
х	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	0	
ц	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	0	
ч	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	0	
ш	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	0	
щ	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	0	
ъ	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	0	
ы	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	0	
ь	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	0	
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	0	
я	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	0	

Рис. 3: к-ть біграм

[illegible]

Рис. 4: ймовірність появи біграм

4 Обчислення ентропій H_1 та H_2

- Обчислення H_1 :

```
1 fn compute_h1(letter_frequencies: &HashMap<char, i64>) -> f64 {
2     let mut h1 = 0.0;
3     let probabilities = count_letters_probabilities(&
4         letter_frequencies);
5
6     for (_key, _value) in probabilities {
7         h1 += _value * f64::log2(_value);
8     }
9
10    h1 = -h1;
11    h1
12 }
```

- Обчислення H_2 :

```
1 fn compute_h2(bigram_frequencies: &HashMap<String, i64>) -> f64 {
2     let mut h2 = 0.0;
3     let probabilities = count_bigram_probabilities(&
4         bigram_frequencies);
5     for (_key, _value) in probabilities {
6         h2 += _value * f64::log2(_value);
7     }
8     h2 = -h2/2.0;
9
10    h2
11 }
```

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:

Лабораторная работа №1

Произвольная часть текста:
чество_в_

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

Неравенство для энтропии:
 $2.94983312141789 < H < 3.45521058262596$

Двоичная таблица угаданных символов:

00010000000000000000000000000000
10000000000000000000000000000000
00000000000000000000000000000000
00000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

$q[1] = 0.3$
$q[2] = 0.1$
$q[3] = 0.08$
$q[4] = 0.1$
$q[5] = 0$
$q[6] = 0.04$
$q[7] = 0$
$q[8] = 0$
$q[9] = 0.02$
$q[10] = 0.04$
$q[11] = 0$
$q[12] = 0.04$
$q[13] = 0.02$
$q[14] = 0.02$
$q[15] = 0.02$
$q[16] = 0.1$
$q[17] = 0$
$q[18] = 0$
$q[19] = 0.02$
$q[20] = 0.02$
$q[21] = 0$
$q[22] = 0$
$q[23] = 0$
$q[24] = 0$
$q[25] = 0.04$
$q[26] = 0$
$q[27] = 0.02$
$q[28] = 0$
$q[29] = 0$
$q[30] = 0$
$q[31] = 0.02$
$q[32] = 0$

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

Рис. 5: значения $H^{(10)}$

Лабораторная работа №1

Произвольная часть текста:
потому_что_в_разные_века_различные_цивилизации_придерживались_совершенно_не

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: 50

Неравенство для энтропии:
 $2.40463996775542 < H < 2.97176867487454$

Двоичная таблица угаданных символов:

01000000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000
00010000000000000000000000000000

Вероятности:

$q[1] = 0.38$
$q[2] = 0.18$
$q[3] = 0.1$
$q[4] = 0.04$
$q[5] = 0$
$q[6] = 0.04$
$q[7] = 0$
$q[8] = 0.02$
$q[9] = 0.02$
$q[10] = 0$
$q[11] = 0.02$
$q[12] = 0$
$q[13] = 0$
$q[14] = 0$
$q[15] = 0$
$q[16] = 0$
$q[17] = 0$
$q[18] = 0$
$q[19] = 0.04$
$q[20] = 0$
$q[21] = 0$
$q[22] = 0.02$
$q[23] = 0$
$q[24] = 0.02$
$q[25] = 0$
$q[26] = 0$
$q[27] = 0.06$
$q[28] = 0.04$
$q[29] = 0$
$q[30] = 0.02$
$q[31] = 0$
$q[32] = 0$

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

Рис. 6: значения $H^{(20)}$

