**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:

- 0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- 1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1 M б), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.
- 2. За допомогою програми CoolPinkProgram оцінити значення *H10*, *H20*, *H30*.
- 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

## Хід роботи:

- 1. Сформували текст, з яким ми працюватимемо. Щоб отримати достатній обсяг тексту, ми зібрали в тексті такі книги, як «Пригоди Незнайка» Носова, Коран та «1984» Оруела. На цьому етапі труднощів не було.
- 2. Написали код, який проводить «фільтрацію» тексту: робить всі літери стрічними, прибирає зайві символи, зайві пробіли, та створили за допомогою цього коду «чисту» версію нашого тексту, готову для аналізу. З труднощів тут було опанування необхідних можливостей «регулярних виразів»
- 3. Написали код, який рахує літери, біграми та обчислює відповідні частоти. На цьому етапі труднощів не було.
- 4. Написали код, який переводить частоти до читабельного вигляду. На цьому етапі труднощі були суто механічні та алгоритмічні.
- 5. Зібрали наші результати в протокол і зробили висновки. Труднощів не було

### Отримані дані:

# Таблиця частот літер з пробілом

# Таблиця частот літер без пробілу

_		
_	0.1646	
0	0.091	
е	0.0698	
a	0.0684	
И	0.0652	
Н	0.0545	
Т	0.0542	
Л	0.0429	
С	0.0419	
В	0.0359	
р	0.0336	
к	0.0293	
М	0.0265	
д 0.0251		
У	0.023	
П	0.0225	
ы 0.0176		
3	0.0155	
Ь	0.0153	
Я	0.015	
б	0.0146	
Г	0.014	
Ч	ч 0.0117	
х	0.0095	
й	0.0091	
ж	0.0079	
Ш	0.0074	
Ю	0.0047	
Щ	0.0033	
ц	0.0024	
Э	э 0.0023	
ф	ф 0.0007	
ë	ë 0.0004	
ъ	0.0002	

H1 = 4.36455, R = 0.14210	H1	= 4	.364	55,	R =	0.	1421	0
---------------------------	----	-----	------	-----	-----	----	------	---

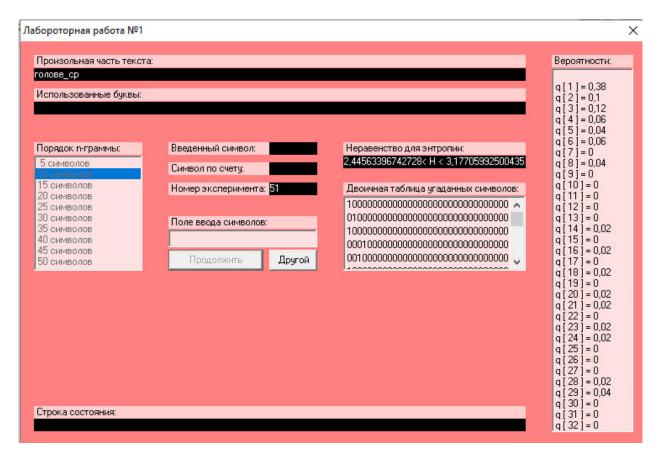
0	0.1089		
е	e 0.0835		
a	0.0819		
И	0.078		
Н	0.0652		
Т	0.0648		
Л	0.0514		
С	0.0501		
В	0.0429		
р	0.0402		
К	0.035		
M	м 0.0317		
Д	д 0.03		
У			
П	0.027		
Ы	0.021		
3	0.0186		
Ь	ь 0.0184		
Я	0.018		
б	0.0174		
Γ	г 0.0168		
Ч	0.014		
Х	0.0114		
й	0.0109		
ж	0.0095		
Ш	0.0089		
Ю	0.0056		
щ	0.004		
Ц	0.0029		
Э	0.0028		
ф	0.0009		
ë	0.0005		
Ъ	0.0002		

H1 = 4.45208, R = 0.11742

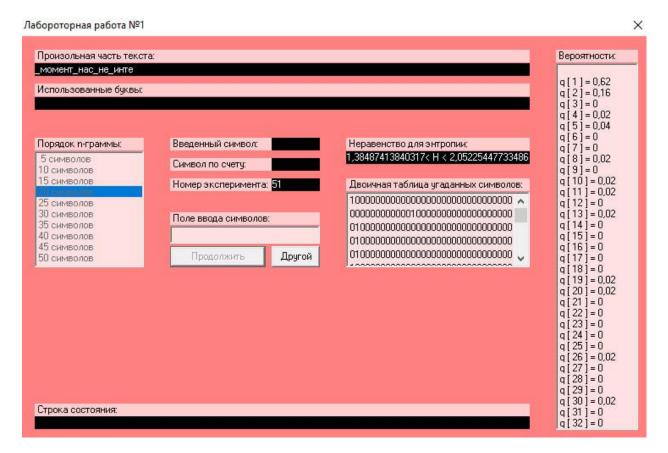
	Ентропія (Н)	Надлишковість (R)
Н2 (з перетинами з пробілами)	7.90085	0.18624
Н2 (без перетинів з пробілами)	7.89978	0.18222
Н2 (з перетинами без пробілів)	8.25920	0.16205
Н2 (без перетинів без пробілів)	8.25764	0.15912
H(10)	2.44563 < H < 3.17705	0.36459 < R < 0.51087
H(20)	1.38487 < H < 2.05225	0.58955 < R < 0.72303
H(30)	1.29114 < H < 2.04913	0.59017 < R < 0.74177

Повна таблична інформація по біграмам залишена у файлі data.xlxs

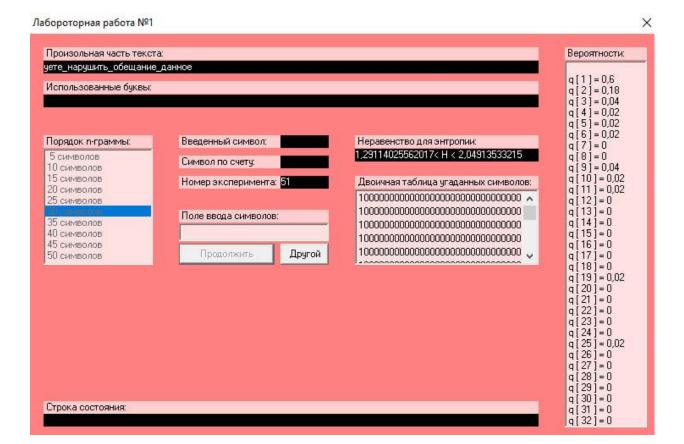
#### H<sub>10</sub>



### H20



H30



#### Оцінки та висновки:

Надлишковість тексту дуже сильно залежить від того, яким саме чином ми його аналізували. Аналізуючи частоти символів ми отримали надлишковість порядку 12-14%, частоти біграм дали нам надлишковість 16-19%. Коли ж ми почали розраховувати умовну ентропію, значення її надлишковості стрімко зросли, особливо на 20 і 30 символах. І це природньо, бо маючи певну послідовність символів, ми зменшуємо ентропію, а з нею і росте надлишковість.

Можемо зробити висновок, що російська мова доволі надлишкова і в теорії можна скоротити слова (а з ними і речення, і тексти) більше, ніж на половину, не втрачаючи змісту.