**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 1Мб), 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 |
| о | 0.091 |
| е | 0.0698 |
| а | 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 |
| з | 0.0155 |
| ь | 0.0153 |
| я | 0.015 |
| б | 0.0146 |
| г | 0.014 |
| 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 | 0.0023 |
| ф | 0.0007 |
| ё | 0.0004 |
| ъ | 0.0002 |

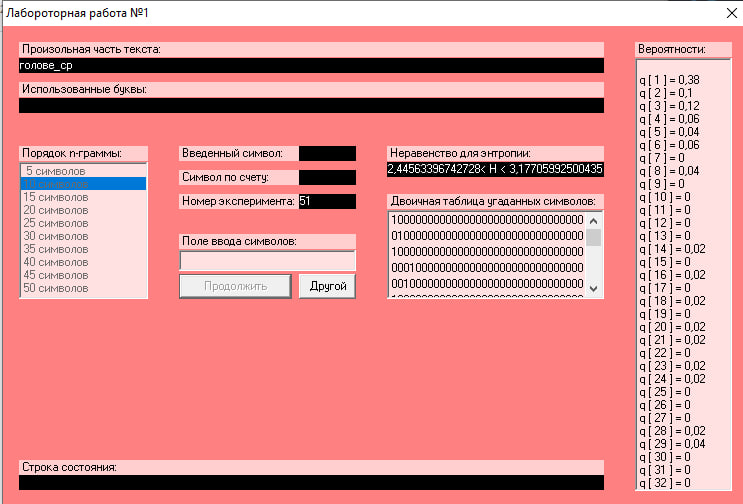
|  |  |
| --- | --- |
| о | 0.1089 |
| е | 0.0835 |
| а | 0.0819 |
| и | 0.078 |
| н | 0.0652 |
| т | 0.0648 |
| л | 0.0514 |
| с | 0.0501 |
| в | 0.0429 |
| р | 0.0402 |
| к | 0.035 |
| м | 0.0317 |
| д | 0.03 |
| у | 0.0275 |
| п | 0.027 |
| ы | 0.021 |
| з | 0.0186 |
| ь | 0.0184 |
| я | 0.018 |
| б | 0.0174 |
| 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 |
| ё | 0.0005 |
| ъ | 0.0002 |

H1 = 4.36455, R = 0.14210 H1 = 4.45208, R = 0.11742

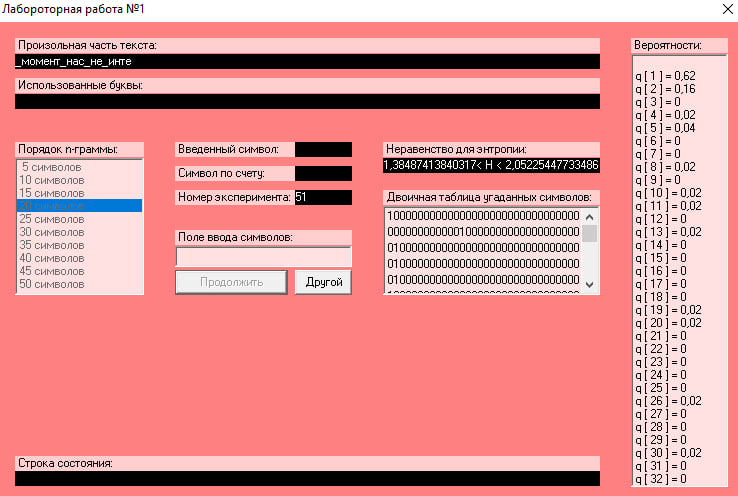
|  |  |  |
| --- | --- | --- |
|  | Ентропія (H) | Надлишковість (R) |
| H2 (з перетинами з пробілами) | 3.95043 | 0.18624 |
| H2 (без перетинів з пробілами) | 3.94989 | 0.18222 |
| H2 (з перетинами без пробілів) | 4.1296 | 0.16205 |
| H2 (без перетинів без пробілів) | 4.12882 | 0.15912 |
| H(10) | 2.44563 < 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

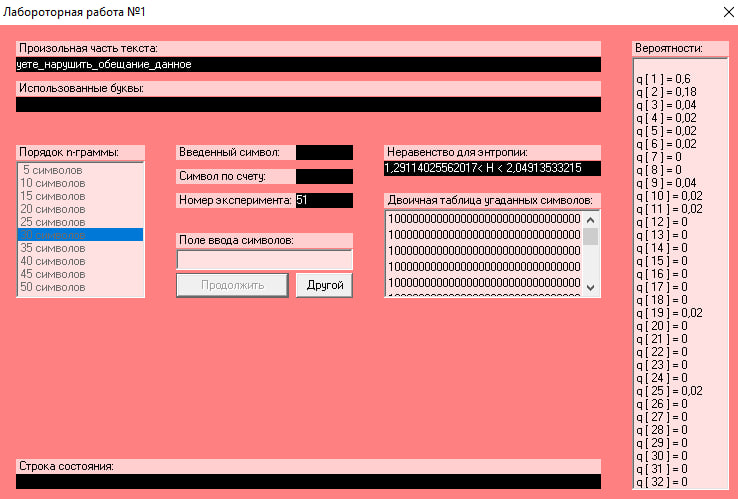
H10



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.