Міністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут ім. Ігоря Сікорського" Фізико-технічний інститут

Лабораторна робота № 1

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:

Студентки 3 курсу, групи ФБ-92 Шаповал Ольга Прохорська Олександра Перевірила: Селюх П.В.

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

- 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 1Н та 2Н за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1Н та 2Н на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1Н та 2Н на тому ж тексті, в якому вилучено всі пробіли.
- 2. За допомогою програми CoolPinkProgram оцінити значення 10(H), 20(H), 30(H).
- 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Спочатку був знайдений текст з яким відбувається подальша робота. Далі текст був відредагований та розбитий у двох різних .txt файлах. У одному з пробілами у іншому без. На цьому етапі виникли невеликі труднощі. Було тяжко видалити всі не потрібні символи. Спочатку ми намагались зробити це через replace на строку з вручну прописаними різними символами, але це не працювало коректно. Рішення було знайдене: було використано регулярні вирази та відповідну бібліотеку ге. Після цього текст редагувався коректно.

Код програми можна знайти на репозиторії в гітхаб (main.py)

Результати

Ентропія:

text_with_spaces:

Entropy for monogram: 4.2909824

Redundant: 0.14180352000000007

text_without_spaces:

Entropy for monogram: 4.49173059

Redundant: 0.09334828323562339

Bigrams for text_with_spaces and with intersection:

Entropy for bigram 3.9157157783714567

Redundant: 0.21685684432570862

Bigrams for text_with_spaces and without intersection:

Entropy for bigram 3.9156052154323326

Redundant: 0.21687895691353343

Bigrams for text_without_spaces and with intersection:

Entropy for bigram 4.159820757213352

Redundant: 0.16034397981122583

Bigrams for text_without_spaces and without intersection:

Entropy for bigram 4.159920757074965

Redundant: 0.16032379493050097

Найчастіша поява літер/біграм

| Монограми з пробілами | Монограми без пробілів |
|----------------------------|----------------------------|
| ' ': 0.20959958012850902, | 'o': 0.1121013474214812, |
| 'o': 0.08860495207009862, | 'e': 0.08523773749515316, |
| 'e': 0.067371943505065, | 'a': 0.07616324156649865, |
| 'a': 0.06019945811293432, | 'н': 0.06612301279565723, |
| 'н': 0.052263657076855444, | 'T': 0.0623218786351299, |
| 'T': 0.049259239040386774, | 'и': 0.06199229352462195, |
| 'и': 0.0489987348306579, | 'c': 0.054879556029468785, |
| 'c': 0.043376824128053146, | 'л': 0.047124612252811164, |
| 'л': 0.037247313310903155, | 'p': 0.04226686700271423, |
| 'p': 0.033407749425597794, | 'B': 0.039292119038386975, |
| 'B': 0.031056507385581668, | 'м': 0.03691474408685537, |
| 'м': 0.029177429225699124, | 'к': 0.03305787126793331, |
| 'к': 0.026128955330232184, | 'y': 0.031269387359441646, |
| 'y': 0.024715336898026966, | 'д': 0.03044905971306708, |
| 'д': 0.02406694958190032, | 'π': 0.027522780147343932, |
| 'π': 0.02175401698449138, | 'я': 0.02483399573478092, |
| 'я': 0.019628800655857658, | 'ь': 0.02082808259015122, |

'ы': 0.01790180302442807, 'ь': 0.016462525224373616, 'ы': 0.014149592626964675, '6': 0.017429236138038, '6': 0.013776075561544598, '4': 0.0165059131446297, '3': 0.01564317564947654, 'u': 0.013046280679877677, '3': 0.012364372601469742, 'r': 0.015615306320279177, 'r': 0.01234234467197061, 'ж': 0.011840829778984102, 'ж': 0.009358996828935888, 'й': 0.011570618456766189, 'm': 0.009240500193873595, 'й': 0.009145421686400819, 'ш': 0.007303695233060283, 'x': 0.00920293718495541, 'x': 0.007274005415039713, 'ю': 0.006088842574641334, 'a': 0.0051703664210934475, 'ю': 0.00481262372752792, '9': 0.004086659790121718, 'φ': 0.004745056223342381, 'φ': 0.003750494431243649, 'щ': 0.003707832493214424, 'III': 0.002930672359449838, 'ц': 0.002958995734780923 'ц': 0.0023387914711687926

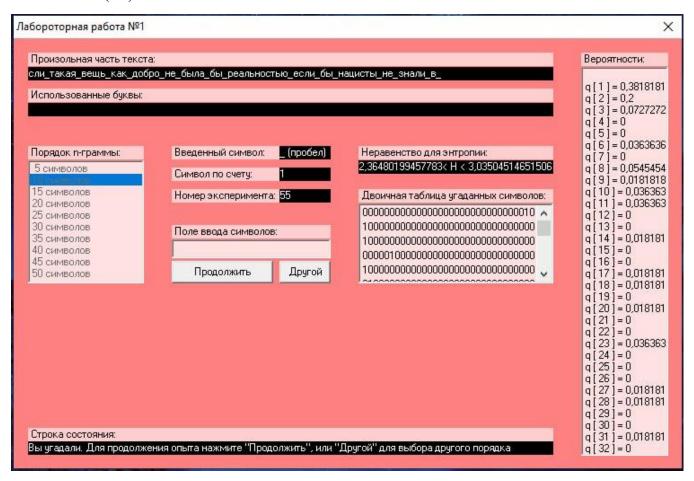
Біграми

| Частоти | Частоти не | Частоти | Частоти не |
|--------------------|--------------------|--------------------|-----------------------|
| перехресних | перехресних | перехресних | перехресних біграм |
| біграм у тексті з | біграм у тексті з | біграм у тексті | у тексті без пробілів |
| пробілами | пробілами | без пробілів | |
| ' ': 0.0572036049 | ' ': 0.0569277772 | 'то': 0.0153499418 | 'то': 0.0153717526 |
| 'o ': 0.0219484164 | 'o ': 0.021895741 | 'но': 0.0137044397 | 'но': 0.0138813494 |
| 'e ': 0.0177726911 | 'e ': 0.0179594495 | 'ст': 0.0131034316 | 'ст': 0.0130937379 |
| ' c': 0.0154540143 | ' c': 0.0154980702 | 'не': 0.0121716266 | 'по': 0.0121801086 |
| ' п': 0.0153275933 | ' п': 0.0154061276 | 'по': 0.0120407619 | 'не': 0.0121170997 |
| ' н': 0.0151255112 | 'н': 0.0151590319 | 'ен': 0.0110980516 | 'ен': 0.0109514347 |
| 'я ': 0.0139982569 | 'я ': 0.0141093542 | 'на': 0.0107551377 | 'на': 0.0109102365 |
| 'и ': 0.0135529101 | 'и ': 0.0135615297 | 'oc': 0.0103177104 | 'oc': 0.0102244087 |
| 'a ': 0.0135146007 | 'a ': 0.0135059811 | 'ко': 0.0097288193 | 'ко': 0.0098560489 |
| ' в': 0.0125923017 | ' в': 0.0125942172 | 'ов': 0.0091908201 | 'ов': 0.0092332299 |

В таблиці наведено 10 найчастіших біграм. Усі інші частоти біграм можна подивитися в текстовому файлі 1.txt у репозиторії.

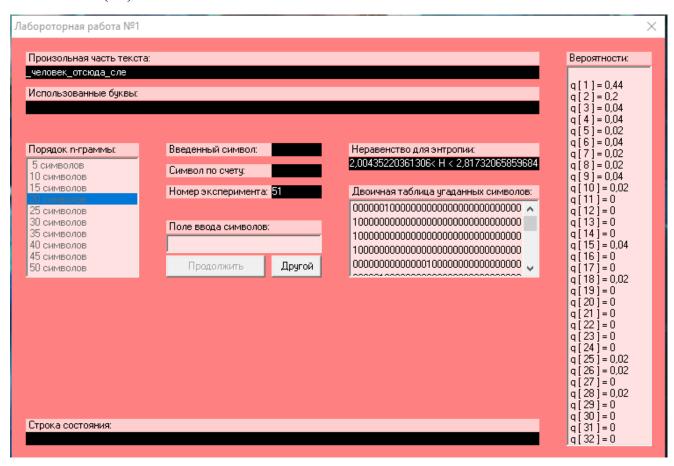
Cool pink program

Значення H(10):



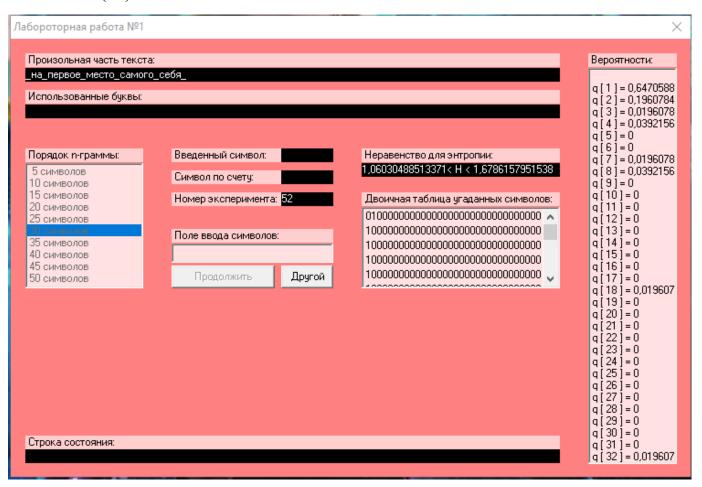
R = 0.46001528589999996

Значення Н(20):



R = 0.517832713779316

Значення Н(30):



R = 0.726107931971249

Висновки: в ході лабораторної роботи ми вивчили що таке ентропія на символі джерела та надлишковість, дослідили та порівняли різні джерела відкритого тексту для наближеного визначення ентропії (з пробілом та без пробілів), а також набули практичних навичок щодо оцінки ентропії на символі джерела.

Під час аналізу тексту ми експериментально перевірили той факт, що в російському алфавіті найчастіше зустрічається пробіл, що свідчить про те, що при шифруванні потрібного його видаляти, щоб зловмисник мав прикласти більше зусиль для зламу. Найбільш вживані монограми в тексті — це "o", "e", "a", а рідше за все зустрічаються: "щ", "ц". Серед біграм у тексті з пробілом частіше всього зустрічаються: "o_", "e_", "e_", "_c", а у тексті без пробілів: "то", "ст", "но".