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:**

Студенти 3 курсу,

Групи ФБ-04

Дмитренко Даніїл

Сербіненко Олексій

**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**

1. Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.
2. Написати програми для підрахунку частот букв і частот біграм в тексті, а також

підрахунку 1H та 2H за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1H та 2H на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1H та 2H на тому ж тексті, в якому вилучено всі пробіли.

1. За допомогою програми CoolPinkProgram оцінити значення 10(H), 20(H), 30(H).
2. 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. Рішення було знайдене: було використано регулярні вирази та відповідну бібліотеку re. Після цього текст редагувався коректно.

Потім ми визначили скільки разів зустрічаються монограми та біграми, визначили частоти та ентропії.

Також ми використали бібліотеку Pandas для того щоб автоматично заповнювати файли Excel.

**Результати**

**Текст з пробілами:**

**монограми:**

Ентропія : 4,37299591618975

Надлішковість: 0,125400816762049

**біграми з перетином:**

Ентропія: 3.960490518055981

0.20790189638880385

**біграми без перетину:**

Ентропія: 3.959026608747424

Надлішковість: 0.20819467825051519

**Текст без пробілів:**

**монограми:**

Ентропія для монограм: 4,45924557855973

Надлішковість: 0,099905

**біграми з перетином:**

Ентропія: 4.135534629112443

Надлішковість: 0.17289307417751143

**біграми без перетину:**

Ентропія: 4.134324331849451

Надлішковість 0.17313513363010968

Найчастіша поява монограм

|  |  |
| --- | --- |
| **Буква** | **Частота** |
|  | 0,16339557 |
| о | 0,09534618 |
| а | 0,07135748 |
| е | 0,06975928 |
| н | 0,05487715 |
| и | 0,05423128 |
| т | 0,0486333 |
| с | 0,04469079 |
| л | 0,04244673 |
| в | 0,03794858 |
| р | 0,03596659 |
| к | 0,02803145 |
| м | 0,02546374 |
| д | 0,02544799 |
| у | 0,02266833 |
| п | 0,02092693 |
| я | 0,01864276 |
| ь | 0,01739256 |
| г | 0,01646314 |
| ы | 0,01596478 |
| б | 0,0151442 |
| з | 0,0143895 |
| ч | 0,01234879 |
| й | 0,00968799 |
| ж | 0,00894188 |
| ш | 0,00831463 |
| х | 0,00731218 |
| ю | 0,00531873 |
| э | 0,00257201 |
| ц | 0,00248179 |
| щ | 0,00245315 |
| ф | 0,00138052 |

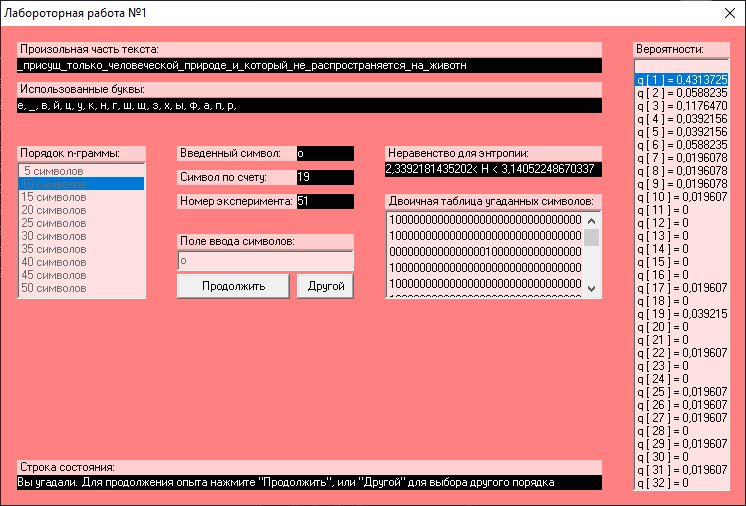
монограми без пробілу монограми з пробілом

|  |  |
| --- | --- |
| **Буква** | **Частота** |
| о | 0,113968 |
| а | 0,085294 |
| е | 0,083384 |
| н | 0,065595 |
| и | 0,064823 |
| т | 0,058132 |
| с | 0,053419 |
| л | 0,050737 |
| в | 0,04536 |
| р | 0,042991 |
| к | 0,033506 |
| м | 0,030437 |
| д | 0,030418 |
| у | 0,027096 |
| п | 0,025014 |
| я | 0,022284 |
| ь | 0,020789 |
| г | 0,019679 |
| ы | 0,019083 |
| б | 0,018102 |
| з | 0,0172 |
| ч | 0,014761 |
| й | 0,01158 |
| ж | 0,010688 |
| ш | 0,009939 |
| х | 0,00874 |
| ю | 0,006358 |
| э | 0,003074 |
| ц | 0,002967 |
| щ | 0,002932 |
| ф | 0,00165 |

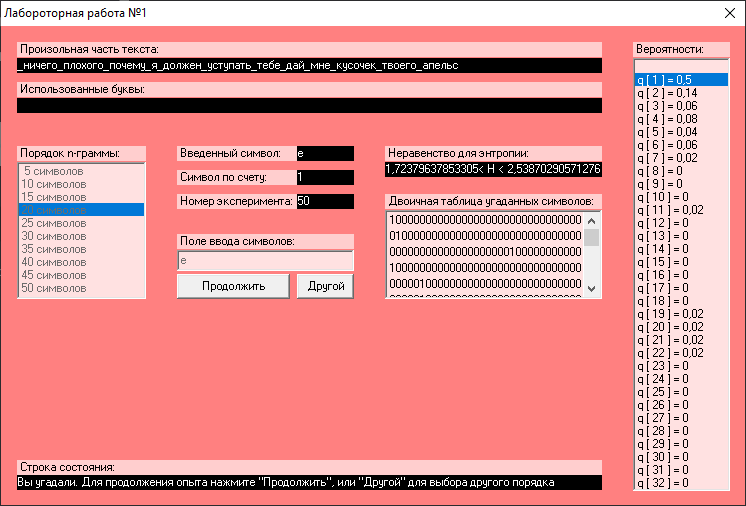
**Біграми**

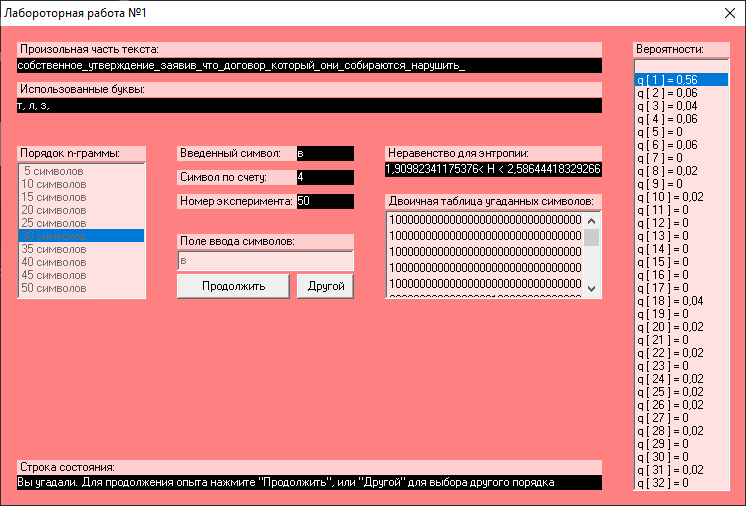
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
| |  |  | | --- | --- | | **Біграми без перетину без пробілу** | **Частота** | | то | 0,016919 | | на | 0,013499 | | ст | 0,013441 | | ов | 0,011318 | | не | 0,011246 | | ал | 0,011202 | | го | 0,010859 | | ос | 0,010582 | | он | 0,010486 | | |  |  | | --- | --- | | **Біграми без перетину з пробілом** | **Частота** | | о | 0,021123 | | и | 0,018382 | | а | 0,017993 | | е | 0,017082 | | с | 0,01707 | | н | 0,016099 | | п | 0,015108 | | в | 0,014536 | | то | 0,014246 | |  |
|  |  |  |
|  |  |  |
|  |  |  |
| |  |  | | --- | --- | | **Біграми з перетином без проб** | **Частота** | | то | 0,017195 | | ст | 0,013362 | | на | 0,01335 | | ов | 0,011416 | | не | 0,011241 | | ал | 0,011214 | | го | 0,010791 | | он | 0,01061 | | ос | 0,010531 | | |  |  | | --- | --- | | **Біграми з перетином з проб** | **Частота** | | о | 0,021283512 | | и | 0,018458027 | | а | 0,018118625 | | е | 0,017098987 | | с | 0,016839781 | | н | 0,016033522 | | п | 0,014840603 | | в | 0,014496905 | | то | 0,014143182 | |  |

У таблицях наведені по 9 біграм найчастіших біграм



R= 0.452025936977643

R = 0.573750071575419



R = 0.550373240495358

Висновки: засвоїли поняття ентропії на символ джерела та його надлишковості, вивчили та порівняли різні моделі джерела відкритого тексту для наближеного визначення ентропії, набули практичних навичок щодо оцінки ентропії на символ джерела. Після аналізу тексту визначили що найчастіші у використанні букви рос. алфавіту «о», «а», «е».