**Лабораторна робота №1**

**Експериментальна оцінка ентропії на символ джерела відкритого тексту**

**Виконали:**

Борщевський Олександр(ФБ-03)

Ржевський Андрій(ФБ-03)

**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**

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H(1) та H(2) за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1() та H(2) на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H(1) та H(2) на тому ж тексті, в якому вилучено всі пробіли.

2. За допомогою програми CoolPinkProgram оцінити значення H(10), H(20), H(30)

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела

**Хід роботи**

Початковою задачею було очищення тексту. Проблем на цьому етапу не виникло, усі не текстові символи видалили, прописні літери замінили на аналогічні стрічні, послідовність пробілів і знак переносу рядків замінили на пробіл. Після очищення тексту можна приступати до наступних задач

Поточною задачею є обчислення частот появи літер. Для цього спочатку рахується кількість випадків появи літери , а потім загальна кількість символів. Частотою появи літери є відношення кількості появи до загальної кількості літер. Тепер можна переходити до обчислення H(1).

Для обчислення H(1) ми всі значення частот літер записуємо у список у спадковому порядку. Після цього застосовуємо формулу

де n – кількість літер алфавіту, p(i) – частота появи літери в тексті

Для обчислення H(2) треба порахувати частоту біграм. Частота біграм – відношення кількості появ певної біграми до загальної кількості біграм у тексті. H(2) обчислюється за формулою

де p(i,j) – частота появи біграми

Для обчислення першої частини формули ми записуємо частоти появи усіх біграм до файлу. Якщо біграма має частоту 0, то вона не записується. Друга частина формули, фактично, дорівнює H(1).

**Результати обчислень**

Повні таблиці містяться у відповідних .csv файлах, та зібрані в одне ціле в xlsx файлі.

Таблиця частот літер

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Літера | Кількість  (без пробілу) | Частота  (без пробілу) | Кількість  (з пробілом) | Частота  (з пробілом) |
| \_ | - | - | 214338 | 0.1545 |
| а | 88265 | 0.0752 | 88265 | 0.0636 |
| б | 19234 | 0.0163 | 19234 | 0.0138 |
| в | 50540 | 0.0430 | 50540 | 0.0364 |
| г | 20815 | 0.0177 | 20815 | 0.0150 |
| д | 35518 | 0.0302 | 35518 | 0.0256 |
| е | 102088 | 0.0870 | 102088 | 0.0735 |
| ё | 18 | 0.000015 | 18 | 0.000013 |
| ж | 11336 | 0.0096 | 11336 | 0.0081 |
| з | 20752 | 0.0176 | 20752 | 0.0149 |
| и | 83040 | 0.0707 | 83040 | 0.0598 |
| й | 13313 | 0.0113 | 13313 | 0.0095 |
| к | 37233 | 0.0317 | 37233 | 0.0268 |
| л | 55455 | 0.0472 | 55455 | 0.0399 |
| м | 39972 | 0.0340 | 39972 | 0.0288 |
| н | 79071 | 0.0674 | 79071 | 0.0569 |
| о | 131090 | 0.1117 | 131090 | 0.0944 |
| п | 32359 | 0.0275 | 32359 | 0.0233 |
| р | 55457 | 0.0472 | 55457 | 0.0399 |
| с | 63733 | 0.0543 | 63733 | 0.0459 |
| т | 71576 | 0.0610 | 71576 | 0.0515 |
| у | 31075 | 0.0264 | 31075 | 0.0224 |
| ф | 2528 | 0.0021 | 2528 | 0.0018 |
| х | 11700 | 0.0099 | 11700 | 0.0084 |
| ц | 3708 | 0.0031 | 3708 | 0.0026 |
| ч | 16922 | 0.0144 | 16922 | 0.0121 |
| ш | 9888 | 0.0084 | 9888 | 0.0071 |
| щ | 4744 | 0.0040 | 4744 | 0.0034 |
| ъ | 226 | 0.00019 | 226 | 0.00016 |
| ы | 25216 | 0.0214 | 25216 | 0.0181 |
| ь | 21912 | 0.0186 | 21912 | 0.0157 |
| э | 3260 | 0.0027 | 3260 | 0.0023 |
| ю | 6724 | 0.0057 | 6724 | 0.0048 |
| я | 24117 | 0.0205 | 24117 | 0.0173 |

Таблиці частот біграм

Таблиці містять 15 найпопулярніших біграм, бо їх дуже багато

|  |  |  |
| --- | --- | --- |
| Без перетину, без пробілу | | |
| Біграма | Кількість | Частота |
| то | 8828 | 0.015053 |
| ст | 7914 | 0.013494 |
| но | 7455 | 0.012712 |
| не | 7079 | 0.012071 |
| на | 6953 | 0.011856 |
| по | 6449 | 0.010996 |
| ен | 6082 | 0.010371 |
| ли | 5903 | 0.010065 |
| ос | 5855 | 0.009983 |
| ов | 5840 | 0.009958 |
| ни | 5729 | 0.009769 |
| ко | 5677 | 0.009680 |
| ра | 5581 | 0.009516 |
| ер | 5507 | 0.009390 |
| ал | 5413 | 0.009230 |

|  |  |  |
| --- | --- | --- |
| Без перетину з пробілом | | |
| Біграма | Кількість | Частота |
| о\_ | 13705 | 8828 |
| и\_ | 12704 | 7914 |
| е\_ | 12402 | 7455 |
| \_п | 11271 | 7079 |
| \_н | 10774 | 6953 |
| \_с | 10625 | 6449 |
| \_в | 10085 | 6082 |
| а\_ | 9804 | 5903 |
| то | 8573 | 5855 |
| ст | 7867 | 5840 |
| но | 7342 | 5729 |
| я\_ | 7331 | 5677 |
| \_о | 7279 | 5581 |
| не | 6900 | 5507 |
| на | 6817 | 5413 |

|  |  |  |
| --- | --- | --- |
| З перетином без пробілу | | |
| Біграма | Кількість | Частота |
| то | 17621 | 0.015023 |
| ст | 15814 | 0.013483 |
| но | 14911 | 0.012713 |
| не | 13841 | 0.011800 |
| на | 13742 | 0.011716 |
| по | 12937 | 0.011030 |
| ен | 12154 | 0.010362 |
| ли | 11648 | 0.009931 |
| ос | 11640 | 0.009924 |
| ов | 11575 | 0.009868 |
| ко | 11446 | 0.009758 |
| ни | 11330 | 0.009659 |
| ра | 11256 | 0.009596 |
| ер | 11048 | 0.009419 |
| ал | 10944 | 0.009330 |

|  |  |  |
| --- | --- | --- |
| З перетином з пробілом | | |
| Біграма | Кількість | Частота |
| о\_ | 27507 | 0.019828 |
| и\_ | 25519 | 0.018395 |
| е\_ | 24780 | 0.017863 |
| \_п | 22416 | 0.016158 |
| \_н | 21323 | 0.015371007668564 |
| \_с | 21159 | 0.015252785783386 |
| \_в | 20116 | 0.014500923428262 |
| а\_ | 19366 | 0.01396027456312 |
| то | 17197 | 0.012396718045129 |
| ст | 15517 | 0.011185664587211 |
| но | 14647 | 0.010558511903646 |
| я\_ | 14555 | 0.010492192309522 |
| \_о | 14513 | 0.010461915973074 |
| не | 13771 | 0.00992703402916 |
| на | 13698 | 0.009874410872953 |

Надлишковість відкритого тексту обчислюється за формулою

5*.*087для тексту з пробілами, 5.044 для тексту без пробілів

|  |  |  |
| --- | --- | --- |
| Модель відкритого тексту | Ентропія | Надлишковість |
| H(1) пробілів немає | 4.451553 | 0.1175 |
| H(1) пробіли є | 4.387398 | 0.1376 |
| H(2) пробілів немає, пари перетинаються | 3.839537 | 0.23885 |
| H(2) пробілів немає, пари не перетинаються | 3.839695 | 0.23881 |
| H(2) пробіли є, пари перетинаються | 3.599388 | 0.29224 |
| H(2) пробіли є, пари не перетинаються | 3.597969 | 0.29277 |

|  |  |
| --- | --- |
| H(10) |  |
| H(20) |  |
| H(30) |  |

Висновок: Під час виконання лабораторної роботи ми навчилися екпериментально визначати частоти літер і біграм у тексті і на основі цих значень обчислювати ентропію і надлишковіть у різних моделях відкритого тексту. Також, за допомогою спеціальної програми приблизно обчислили значення H(10), H(20), H(30)