Отчет о ДЗ1

0. Корпус текстов

Корпус "произведения Пушкина" состоит из 12 текстов и 110374 словоформ.

В нем следующие произведения:

- Евгений Онегин (Е)

- Капитанская дочка (К)

- Борис Годунов (Б)

- Дубровский (Д)

- Метель (М)

- Руслан и Людмила (РиЛ)

- Сказка о царе Салтане

- Сказка о рыбаке и рыбке

- Сказка о попе и работнике его Балде

- Сказка о мертвой царевне и о семи богатырях

- Сказка о золотом петушке

- Песнь о вещем Олеге

1. Выбранный текст

- Евгений Онегин отрывок в 2000 слов

1.1. Ключевые слова в выбранном тексте

любить; 503.1; 851.6; 21354.2

любовь; 323.9; 1132.5; 26562.5

Евгений; 72.7; 416.8; 23437.5

думать; 755.5; 625.1; 10416.7

герой; 184.0; 398.6; 7812.5

иль; 7.7; 788.2; 30208.3

В топ-500 вошли: любить, любовь, думать

Среди довольно редких слов: иль (по словарю), герой (для текстов корпуса)

Среди выбранных слов: 2 глагола, 3 существительных (больше всего) и 1 союз

Почти во всех текстах (кроме 3-х, 4-х) встретились любить, думать и любовь - 2 глагола и 1 существительное, которое однокоренное к 1 из

тих глаголов (самостоятельные части речи).

1.2. term\*document

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| term document | Дубровский | Кап дочка | Салтан | Метель | Р и Л |
| любить | 14 | 7 | 1 | 3 | 10 |
| любовь | 3 | 10 | 1 | 4 | 37 |
| Евгений | 0 | 0 | 0 | 0 | 0 |
| думать | 13 | 32 | 0 | 2 | 5 |
| герой | 3 | 0 | 0 | 1 | 23 |
| иль | 1 | 0 | 4 | 0 | 7 |
| человек | 57 | 74 | 3 | 6 | 1 |
| время | 39 | 40 | 2 | 7 | 9 |
| рука | 38 | 61 | 0 | 4 | 39 |

при таком порядке word\_i все 5 текстов подходят под запрос Word1&Word2&¬Word3 (любить&любовь&не(Евгений))

1.3.

Возьмем того же Онегина, но не отрывок, а весь текст

- любить

- любовь

- Татьяна

- Евгений

- серодце

- Онегин

Два высокочастотных слова в русском языке:

- и

- в

Два редких слова, которые встречаются в Онегине по 1 разу, а во всех текстах не больше 3-х раз:

- жеребец

- недостаток

Таблица: (я сюда вставила и Count(wi) на всякий случай)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Слово (Лексема) | Fr (L) | Fr(Coll) Countcoll(wi)/N(Coll) | Count(wi) | tf |
| любить | 503.1 | 0,00085165 | 41 | 0,001688841 |
| любовь | 323.9 | 0,001132513 | 51 | 0,002100754 |
| Татьяна | 58.6 | 0,000933191 | 103 | 0,004242699 |
| Евгений | 72.7 | 0,000416765 | 46 | 0,001894798 |
| сердце | 245.1 | 0,00084259 | 70 | 0,000823825 |
| Онегин | 9 | 0,000797289 | 88 | 0,00362483 |
| и | 35801.8 | 0,041739903 | 1053 | 0,043374387 |
| в | 31374.2 | 0,025504195 | 637 | 0,026238827 |
| жеребец | 6.2 | 2,71803E-05 | 1 | 4,11913E-05 |
| недостаток | 62.1 | 2,71803E-05 | 1 | 4,11913E-05 |

Таблица tf\*idf:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count(w) | LengthDoc | tf | df | N | idf | tf\*idf |
| Татьяна | 103 | 24277 | 0,004242699 | 1 | 12 | 1.079 | 0.004578 |
| Онегин | 88 | 24277 | 0,00362483 | 1 | 12 | 1.079 | 0.003911 |
| Евгений | 46 | 24277 | 0,001894798 | 1 | 12 | 1.079 | 0.002044 |
| любовь | 51 | 24277 | 0,002100754 | 7 | 12 | 0.468 | 0.000983 |
| любить | 41 | 24277 | 0,001688841 | 9 | 12 | 0.255 | 0.000431 |
| сердце | 70 | 24277 | 0,000823825 | 9 | 12 | 0.255 | 0.000210 |
| жеребец | 1 | 24277 | 4,11913E-05 | 2 | 12 | 1.556 | 0.000064 |
| недостаток | 1 | 24277 | 4,11913E-05 | 2 | 12 | 1.556 | 0.000064 |
| и | 1053 | 24277 | 0,043374387 | 12 | 12 | 0 | 0 |
| в | 637 | 24277 | 0,026238827 | 12 | 12 | 0 | 0 |

Да, в верх таблицы попали те слова, которые я посчитала ключевыми.

Да, все ключевые попали в верхние шесть.

Вниз попали редкие слова (у них специфическая семантика, встречающаяся в определенных контекстах) и самые частотные слова (служебные).

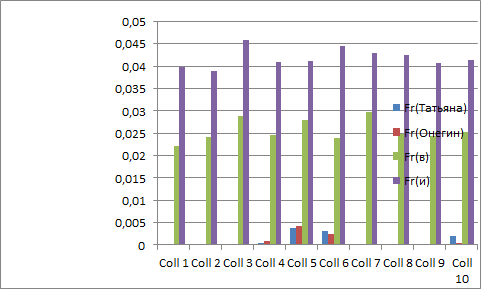
В "стоп"-словах должны быть самые частотные слова РЯ: служебные слова (предлоги, союзы, частицы), может быть, личные местоимения 3 лица, глагол "быть" (хотя это скорее для английского).

1.4.

Делим на 10 частей примерно по 11000 словоформ в тексте.

2 слова с самым высоким tf\*idf: Татьяна и Онегин, которые встречаются только в тексте Евгений Онегин

2 высокочастотных слова в языке: "в" и "и"



Самые частотные слова есть во всех coll1, а с высоким tf\*idf только в небоьшом количестве коллекций (в которые вошли части Евгения Онегина), еще во всех коллекциях стабильно «в» больше, чем «и», а там, где есть «Татьяна» и «Онегин» их соотношение непредсказуемо.