**Задание 1.1**

*(ipm рассчитывалось для всех словоформ)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Лексема** | **ipm по словарю** | **ipm по коллекции** | **ipm по тексту** |
| Велосипед | 23,1 | 578,0 | 13133,2 |
| Ездить | 93,6 | 261,9 | 7504,7 |
| Машина | 490,4 | 1120,0 | 10318,9 |
| Разогнать | 6,8 | 9,0 | 938,1 |
| Тормоз | 14,7 | 36,1 | 3752,3 |
| Бежать | 143,6 | 668,4 | 1876,2 |

1. Есть ли среди выбранных вами ключевых слов редкие слова?

Да, *разогнать* можно отнести к редким.

1. Есть ли среди выбранных вами слов слова, вошедшие в топ 500 по частоте?

Да, слово *машина* на 187 месте в частотном списке лемм (<http://dict.ruslang.ru/freq.php>)

1. К каким частям речи относятся выбранные вами слова, слов какой части речи больше?

Три существительных, три глагола.

1. Какие слова встретились во всех или в большинстве документов? Каковы их грамматические характеристики

Союзы, предлоги, местоимения, частицы. Топ-5: *и, я, не, а, в.*

**Задание 1.2**

Матрица term\*document

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | Fomina\_text\_1 | Fomina\_text\_2 | Fomina\_text\_3 | Fomina\_text\_4 | Fomina\_text\_5 | Fomina\_text\_1\_ fragment\_research | | велосипед | 0 | 8 | 2 | 0 | 11 | 8 | | ездить | 0 | 2 | 0 | 0 | 0 | 2 | | машина | 1 | 5 | 0 | 0 | 6 | 4 | | разогнать | 0 | 1 | 0 | 0 | 0 | 1 | | тормоз | 0 | 4 | 0 | 0 | 0 | 4 | | бежать | 1 | 2 | 0 | 1 | 1 | 1 | | и | 489 | 849 | 558 | 504 | 649 | 58 | | в | 156 | 258 | 169 | 176 | 178 | 14 | | не | 132 | 230 | 174 | 163 | 195 | 21 | |  |  |  |  |  |  |

1. Назовите те слова, у которых мощность обратного индекса (количество документов, в которых слово встречается) равна количеству документов в коллекции

В данном случае таких слов нет, т.к. я строила матрицу исходя из того, что term = лемма (исходная форма слова). Если бы учитывались все словоформы, то ответом на этот вопрос стало бы слово *бежать. (Это я поняла из дальнейших заданий)*

**Задание 1.3**

|  |  |  |  |
| --- | --- | --- | --- |
| **Слово (Лексема)** | **Fr (L)** | **tf** | **df** |
| велосипед | 23,1 | 0,013133208 | 5 |
| тормоз | 14,7 | 0,003752345 | 1 |
| разогнать | 6,8 | 0,000938086 | 1 |
| гребешок | 2,6 | 0,000938086 | 1 |
| скотч | 2 | 0,000938086 | 1 |
| машина | 490,4 | 0,010318949 | 9 |
| бежать | 143,6 | 0,001876173 | 11 |
| ездить | 93,6 | 0,00750469 | 9 |
| и | 35801,8 | 0,054409006 | 11 |
| не | 18028 | 0,019699812 | 11 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Слово (Лексема)** | **Count(w)** | **LengthDoc** | **tf** | **df** | **N** | **idf** | **tf\*idf** |
| велосипед | 14 | 1066 | 0,01313 | 5 | 11 | 0,68484536 | 0,008994217 |
| тормоз | 4 | 1066 | 0,00375 | 1 | 11 | 2,08278537 | 0,00781533 |
| разогнать | 1 | 1066 | 0,00094 | 1 | 11 | 2,08278537 | 0,001953832 |
| гребешок | 1 | 1066 | 0,00094 | 1 | 11 | 2,08278537 | 0,001953832 |
| скотч | 1 | 1066 | 0,00094 | 1 | 11 | 2,08278537 | 0,001953832 |
| машина | 11 | 1066 | 0,01032 | 9 | 11 | 0,17430035 | 0,001798596 |
| бежать | 2 | 1066 | 0,00188 | 11 | 30 | 0,87145714 | 0,001635004 |
| ездить | 8 | 1066 | 0,0075 | 9 | 11 | 0,17430035 | 0,00130807 |
| и | 58 | 1066 | 0,05441 | 11 | 11 | 0 | 0 |
| не | 21 | 1066 | 0,0197 | 11 | 11 | 0 | 0 |

1. Соответствуют ли те слова, которые попали вверх списка, упорядоченного по убыванию tf.idf, Вашей интуиции?

Да, соответствуют.

1. Все ли ключевые слова попали в верхнюю часть списка (в первые шесть слов), ранжированного по tf.idf?

Нет, *бежать* не попало.

1. Какие слова попали вниз ранжированного списка? Каковы их характеристики с точки зрения грамматических характеристик, семантики;

В конец списка попали самые частотные слова, которые есть во всех документах коллекции. Союз, частица и глагол *бежать*.

1. Как, по-вашему, должен быть устроен список «стоп»-слов, данные о которых нет смысла включать в таблицу?

Мне кажется, в этот список должны входить как раз самые частотные слова: союзы, предлоги, частицы, некоторые местоимения.

**Задание 1.4**

*NB! В моей коллекции текстов было всего 11 файлов, поэтому я не стала дробить их на 10 и оставила как есть. Надеюсь, это не очень страшно.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **велосипед** | **тормоз** | **и** | **я** |
| **text\_1** | 0,00000 | 0,00000 | 0,05776 | 0,03662 |
| **text\_2** | 0,00106 | 0,00030 | 0,06442 | 0,03384 |
| **text\_3** | 0,00021 | 0,00000 | 0,05802 | 0,04242 |
| **text\_4** | 0,00000 | 0,00000 | 0,05344 | 0,03997 |
| **text\_5** | 0,00133 | 0,00000 | 0,06178 | 0,03208 |
| **text\_6** | 0,00000 | 0,00000 | 0,06472 | 0,03053 |
| **text\_7** | 0,00011 | 0,00000 | 0,05106 | 0,03337 |
| **text\_8** | 0,00000 | 0,00000 | 0,03654 | 0,02436 |
| **text\_9** | 0,00327 | 0,00000 | 0,03582 | 0,01168 |
| **text\_10** | 0,00000 | 0,00000 | 0,03508 | 0,01405 |
| **text\_11** | 0,00000 | 0,00000 | 0,04220 | 0,01972 |

*Из-за очень большого диапазона значений я сделала отдельную диаграмму для велосипеда и тормоза.*

1. Отличаются ли диаграммы для самых частотных в языке слов и для слов с высоким tf.idf в Вашем списке, если отличаются, то чем?

Да, очень сильно отличаются. Значения для *и, я* выше из-за того, что они гораздо более частотны.

|  |  |  |
| --- | --- | --- |
| **Лексема** | **ipm по коллекции** | **ipm по словарю** |
| велосипед | 578,0 | 23,1 |
| тормоз | 36,1 | 14,7 |
| и | 51436,5 | 35801,8 |
| я | 28938,1 | 12684,4 |