Иванов Вячеслав, БКЛ141

Домашнее задание по АвтОбрЕЯ №1

# Задание 1.1

**Добавить частоты ipa сюда**

1) Среди выбранных слов есть редкие слова, такие, как "ветеран", "поликлиника" и "инвалидность".

2) Слов, вошедших по частоте в топ 500 не оказалось (самое близкое - это "очередь", занявшее 511 место)

3) Все слова являются существительными

4) Во всех документах встретились слова "и" (союз), "в" (предлог), "не" (частица), "на" (предлог), "что" (союз), я (местоимение).

# Задание 1.2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | bazhov\_gornyj\_master | Chernyshevskij\_chto\_delat | Gilyarovskij\_moi\_skitanija | 1\_krasavin | 107\_dragunskij\_volshebnaja\_sila\_iskusstva |
| очередь | 0 | 0 | 1 | 9 | 0 |
| комиссия | 0 | 0 | 0 | 5 | 0 |
| коридор | 0 | 0 | 0 | 1 | 5 |
| ветеран | 0 | 2 | 0 | 2 | 0 |
| поликлиника | 0 | 0 | 0 | 5 | 0 |
| инвалидность | 0 | 0 | 0 | 4 | 0 |

# Задание 1.3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count(w) | LengthDoc | tf | Count  (doc) | N | idf | tf\*idf |
| поликлиника | 5 | 1350 | 0.003703 | 1 | 200 | 4.602059 | 0,017041 |
| комиссия | 5 | 1350 | 0.003703 | 2 | 200 | 4.0 | 0,014812 |
| очередь | 9 | 1350 | 0.006666 | 18 | 200 | 2.091514 | 0,013942 |
| инвалидность | 4 | 1350 | 0.002962 | 1 | 200 | 4.602059 | 0,013631 |
| ветеран | 2 | 1350 | 0.001481 | 1 | 200 | 4.602059 | 0,006525 |
| фойе | 1 | 1350 | 0.000740 | 3 | 200 | 3.647817 | 0,002699 |
| сказать | 8 | 1350 | 0.005925 | 135 | 200 | 0.341392 | 0,002022 |
| коридор | 1 | 1350 | 0.000740 | 22 | 200 | 1.917214 | 0,001418 |
| лежачий | 1 | 1350 | 0.000740 | 1 | 200 | 4.602059 | 0,001402 |
| и | 556 | 1350 | 0.411851 | 200 | 200 | 0.0 | 0 |

1. Слова, попавшие в верх списка, упорядоченного по tf.idf, соответствуют моей интуиции.
2. Все слова, кроме одного («коридор») попали в верх списка.
3. В самый низ списка попали причастие «лежачий» и союз «и». Это говорит о том, что их вес в коллекции незначителен.
4. Список стоп-слов должен включать в себя все служебные части речи, часто встречающиеся в текста.

# Задание 1.4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | и | в | поликлиника | комиссия |
| Coll 1 | 0,604 | 0,460 | 0,003 | 0,003 |
| Coll 2 | 0,688 | 0,560 | 0 | 0 |
| Coll 3 | 0,791 | 0,552 | 0 | 0 |
| Coll 4 | 0,547 | 0,438 | 0 | 0 |
| Coll 5 | 0,503 | 0,289 | 0 | 0,0007 |
| Coll 6 | 0,582 | 0,237 | 0 | 0 |
| Coll 7 | 0,543 | 0,278 | 0 | 0 |
| Coll 8 | 0,763 | 0,498 | 0 | 0 |
| Coll 9 | 0,523 | 0,246 | 0 | 0 |
| Coll 10 | 0,443 | 0,265 | 0 | 0 |
| ipm по всему корпусу | 41451 | 26560 | 34 | 34 |

Значения ipm отличаются от данных в корпусе.

Диаграммы для самых частотных слов и слов с высоким tf.idf отличаются намного прежде всего потому, что эти слова имеют меньшую частотность в корпусе и его подкорпусах, а порой и вовсе не встречаются в них.