Сонина Полина, бак-141

***Задание 1***

Я взяла в качестве коллекции подкорпус НКРЯ со снятой омонимией в формате xhtml.

Очистила тексты от метаданных, тегов, помет, знаков препинания.

533 документа, около 991100 словоупотреблений

## Задание 1.1

Текст: kra.txt (Алексей Краевский. Журналы и поклонники // ««Октябрь», №7», 2003)

1228 словоупотреблений

Ключевые слова:

|  |  |  |  |
| --- | --- | --- | --- |
| lex | частота по частотному словарю (ipm) | частота в коллекции (ipm) | частота в тексте (ipm) |
| журнал | 162,0 | 184,6 | 17101,0 |
| читатель | 114,7 | 129,1 | 11400,7 |
| автор | 262,1 | 266,4 | 8957,7 |
| критик | 31,3 | 67,6 | 7329,0 |
| литература | 156,0 | 203,8 | 6514,7 |
| номер | 201,2 | 151,3 | 3257,3 |

*1)Есть ли среди выбранных вами ключевых слов редкие слова?*

Нет редких слов.

*2)Есть ли среди выбранных вами слов слова, вошедшие в топ 500 по частоте?*

*Автор* – 398 место в частотном словаре

*3)К каким частям речи относятся выбранные вами слова, слов какой части речи больше?*

Существительные

*4)Какие слова встретились во всех или в большинстве документов? Каковы их грамматические характеристики?*

Слово *автор* встретилось в 91 документе из 533 (остальные в меньшем количестве)  
Существительное, м.р.

## Задание 1.2 (Усложнённый вариант)

Матрица по всей коллекции в файле: tf\_idf\_docfreq.xlsx

фрагмент:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| term | freq | Ndocs | idf | 00051091.txt | 00054288.txt | 00055988.txt | 00058141.txt | 00059906.txt | … |
| и | 33274 | 529 | 0,003271537 | 56 | 37 | 23 | 41 | 52 | … |
| в | 31477 | 529 | 0,003271537 | 40 | 37 | 19 | 24 | 77 | … |
| не | 17273 | 504 | 0,024296673 | 35 | 16 | 11 | 12 | 17 | … |
| этот | 14934 | 507 | 0,021719250 | 35 | 10 | 4 | 5 | 11 | … |
| что | 14880 | 497 | 0,030370820 | 33 | 11 | 5 | 7 | 14 | … |
| на | 13923 | 520 | 0,010723865 | 40 | 33 | 15 | 14 | 41 | … |
| … | … | … | … | … | … | … | … | … | … |

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

Таких нет.  
У двух слов обратный индекс 529 - ***и****,* ***в*** (всего 533 документа). Исключения – очень короткие тексты.

*Найдите тексты, удовлетворяющие запросу Word1&Word2&-Word3*

Например, слова: *ребёнок, друг, человек*

9 документов:   
1-7.txt, 10975741.txt, 331465.txt, 450910.txt, 517871.txt, dash10\_09.txt, interview5.txt, nachshk.txt, nezhny.txt

## Задание 1.3

Тестовые слова:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Слово (Лексема) | Fr (L)  (ipm) | Fr(Coll) Countcoll(wi)/N(Coll)  (ipm) | Count(wi) | tf |
| журнал | 162,0 | 184,6 | 21 | 0,0171 |
| читатель | 114,7 | 129,1 | 14 | 0,0114 |
| автор | 262,1 | 266,4 | 11 | 0,0090 |
| критик | 31,3 | 67,6 | 9 | 0,0073 |
| литература | 156,0 | 203,8 | 8 | 0,0065 |
| номер | 201,2 | 151,3 | 4 | 0,0033 |
| у | 4306,1 | 5046,92 | 1 | 0,0008 |
| бы | 2322,4 | 2589,04 | 4 | 0,0033 |
| пустышка | 1,7 | 3,03 | 1 | 0,0008 |
| эссеистика | нет данных | 1,01 | 1 | 0,0008 |

Тематический вес (tf.idf) слов из тестового списка:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Слово (Лексема) | Count(w) | LengthDoc | tf | Count(doc) | N | idf | tf\*idf |
| журнал | 21 | 1228 | 0,017101 | 74 | 533 | 0,857495 | 0,0147 |
| читатель | 14 | 1228 | 0,011401 | 41 | 533 | 1,113943 | 0,0127 |
| критик | 9 | 1228 | 0,007329 | 29 | 533 | 1,249606 | 0,0092 |
| автор | 11 | 1228 | 0,008958 | 91 | 533 | 0,767686 | 0,0069 |
| литература | 8 | 1228 | 0,006515 | 61 | 533 | 0,941397 | 0,0061 |
| номер | 4 | 1228 | 0,003257 | 76 | 533 | 0,845914 | 0,0028 |
| эссеистика | 1 | 1228 | 0,000814 | 1 | 533 | 2,726727 | 0,0022 |
| пустышка | 1 | 1228 | 0,000814 | 1 | 533 | 2,249606 | 0,0018 |
| бы | 4 | 1228 | 0,003257 | 327 | 533 | 0,212179 | 0,0007 |
| у | 1 | 1228 | 0,000814 | 403 | 533 | 0,121422 | 0,0001 |

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

Соответствуют

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

Все

*3) Какие слова попали вниз ранжированного списка? Каковы их характеристики с точки зрения грамматических характеристик, семантики;*

Частотные из топ-100 по частотному словарю.   
 Частицы, предлоги; не имеют особой смысловой, стилистической окраски.

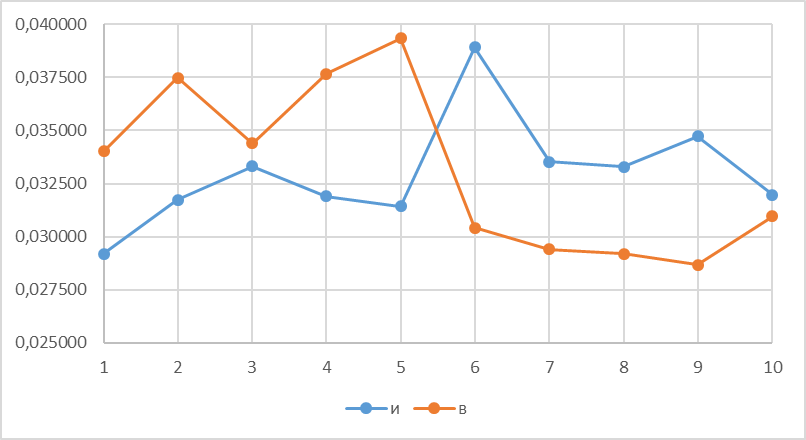
*4) Как, по-вашему, должен быть устроен список «стоп»-слов, данные о которых нет смысла включать в таблицу?*

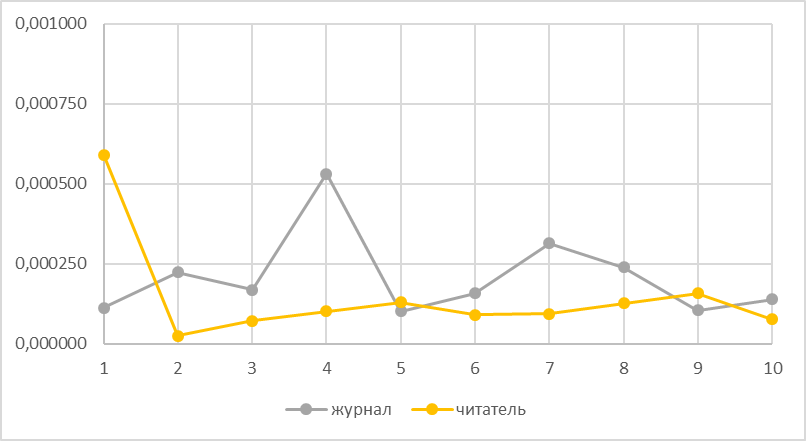
Например, самые частотные предлоги; частотные, нейтральные по стилю и смыслу частицы, союзы и местоимения.

## Задание 1.4

tf\*idf:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coll | Fr(*и*) | Fr(*в*) | Fr(*журнал*) | Fr(*читатель*) |
| Coll 1 | 0,029196 | 0,034034 | 0,000113 | 0,000591 |
| Coll 2 | 0,031727 | 0,037485 | 0,000224 | 0,000025 |
| Coll 3 | 0,033308 | 0,034399 | 0,000170 | 0,000073 |
| Coll 4 | 0,031896 | 0,037666 | 0,000532 | 0,000102 |
| Coll 5 | 0,031423 | 0,039342 | 0,000102 | 0,000130 |
| Coll 6 | 0,038925 | 0,030415 | 0,000158 | 0,000090 |
| Coll 7 | 0,033520 | 0,029416 | 0,000314 | 0,000093 |
| Coll 8 | 0,033301 | 0,029196 | 0,000239 | 0,000127 |
| Coll 9 | 0,034719 | 0,028685 | 0,000106 | 0,000158 |
| Coll 10 | 0,031973 | 0,030959 | 0,000140 | 0,000077 |





|  |  |  |
| --- | --- | --- |
| lex | ipm | частот. словарь |
| и | 33541,5 | 35801,8 |
| в | 31698,1 | 31374,2 |
| журнал | 184,6 | 162,0 |
| читатель | 129,1 | 114,7 |

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

Как очевидно, у частотных намного выше частота в коллекции. Слова *читатель* и *журнал* значительно тематичны для 1 и 4 подкорпусов соответственно.   
 Многие различия, возможно, вызваны тем, что в разных частях коллекции содержатся различные по тематике и стилю тексты (например, во второй половине коллекции больше интервью)