|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Слово (Лексема) | Fr (L) | Fr(Coll) Countcoll(wi)/N(Coll) | Count(wi) | tf |
|  |  |  |  |  |
| город | 573.4 | 7.602 | 1969 | 0.000074 |
| год | 3727.5 | 13.019 | 3372 | 0.00014 |
| музей | 87.0 | 6.135 | 1589 | 0.000071 |
| Испания | 25.7 | 0.613 | 159 | 0.00021 |
| ресторан | 79.3 | 2.046 | 530 | 0.000042 |
| Мадрид | 4.1 | 0.536 | 139 | 0.00021 |
| виза | 15.7 | 0.833 | 216 | 0.000213 |
| быть | 12160.7 | 16.984 | 4399 | 0.000071 |
| копия | 33.7 | 0.0115 | 3 | 0.000073 |
| кража | 9.2 | 0.0115 | 3 | 0.000071 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count(w) | LengthDoc | tf | Count  (doc) | N | idf | tf\*idf |
| город | 7 | 1900 | 0.0036 | 259 | 259 | 0.80051 | 0.000056 |
| год | 4 | 1900 | 0.0021 | 259 | 259 | 0.38391 | 0.000046 |
| музей | 1 | 1900 | 0.0005 | 259 | 259 | 0.4737 | 0.000035 |
| Испания | 23 | 1900 | 0.0121 | 259 | 5 | 1.3341 | 0.00028 |
| ресторан | 8 | 1900 | 0.0042 | 259 | 259 | 0.41329 | 0.000100 |
| Мадрид | 16 | 1900 | 0.0084 | 259 | 5 | 1.26717 | 0.00027 |
| виза | 6 | 1900 | 0.0031 | 259 | 259 | 1.20917 | 0.00025 |
| быть | 7 | 1900 | 0.0036 | 259 | 259 | 0.38391 | 0.000027 |
| копия | 1 | 1900 | 0.0005 | 259 | 2 | 1.0330 | 0.0000734 |
| кража | 1 | 1900 | 0.0005 | 259 | 2 | 2.11226 | 0.00015 |

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

Да, в основном это слова ключевые или просто очень частотные в языке.

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

Нет, всего лишь половина слов ключевых попала на вершину

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

В низ списка попали служебные части речи (предлоги, союзы, частицы)

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

Служебные слова (несоставные предлоги и тд), местоимения (в том числе и указательные), наречия места и времени.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Fr(w1) быть | Fr(w2)  который | Fr(w3)  день | Fr(w4)  Испания |
| Coll 1 | 0.0083 | 0.0033 | 0.0043 | 0.0 |
| Coll 2 | 0.0016 | 0.0055 | 0.0 | 0.0 |
| Coll 3 | 0.0027 | 0.0029 | 0.0053 | 0.0 |
| Coll 4 | 0.0016 | 0.0037 | 0.0064 | 0.0 |
| Coll 5 | 0.0011 | 0.0011 | 0.0011 | 0.0 |
| Coll 6 | 0.0038 | 0.0069 | 0.0075 | 0.0 |
| Coll 7 | 0.0020 | 0.0011 | 0.0064 | 0.019 |
| Coll 8 | 0.0023 | 0.0025 | 0.0021 | 0.0 |
| Coll 9 | 0.0016 | 0.0018 | 0.013 | 0.0 |
| sColl 10 | 0.0025 | 0.0015 | 0.015 | 0.0 |

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

Значения частотных слов в корпусе и слов в р.я. во многом пересекаются, поэтому нельзя сказать, что они сильно отличаются.ы