Иванов Вячеслав

Домашнее задание №1

Отчёт

Для исследования в качестве тематического корпуса был взят газетный корпус НКРЯ (http://www.ruscorpora.ru/search-paper.html). Данный корпус содержит 433 373 документа, 16 669 748 предложений, 228 521 421 слово и включает в себя тексты публицистических работ, созданных в период с 2000 по 2014 годы. В корпусе используется морфологическая разметка, а также частично снята омонимия.

В качестве контрастного корпуса был взят основной корпус НКРЯ (<http://www.ruscorpora.ru/search-main.html>), который включает в себя 115 645 документов, 23 803 881 предложение, 283 431 966 слов и состоит из произведений художественной литературы, изданных с 1800 по 2014 годы.

В качестве второй метрики был взят tf-idf по тематическому корпусу (https://ru.wikipedia.org/wiki/TF-IDF):







|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| wi | Тип (тематическое vs. общеупотребительное) | CountSpecC  (частота в Specific Corpora) | CountRefC (частота в контрастном корпусе  (Reference Corpora)) | LogLikelihood | Ранг | Вес 2 (tf-idf) | Ранг |
| корреспондент | тематическое | 285,9 | 57,8 | 83,96 | 3 | 0,0027 | 3 |
| агентство | тематическое | 249,9 | 27,2 | 125,36 | 2 | 0,0030 | 2 |
| событие | общеупотребительное | 268,1 | 182,9 | 0,17 | 4 | 0,0023 | 4 |
| дело | общеупотребительное | 1368,2 | 1571,1 | 228,5 | 1 | 0,0035 | 1 |

Обе метрики (LogLikelihood и tf-idf) показали одинаковый результат. Самую высокую позицию в ранжированном списке получило общеупотребительное слово‘дело’, а тематические слова ‘агентство’ и ‘корреспондент’ стоят на 2 и 3 позиции соответственно. Это показывает, что общеупотребительное (по нашему предположению) слово ‘дело’, оказавшееся в верху списка, на самом деле не является общеупотребительным, а скорее специфичным именно для газетного корпуса. Однако выделенные нами тематические слова всё же оказались характерными именно для газетного корпуса, так как тоже показали высокий результат. Результат для общеупотребительного слова ‘событие’ подтвердил наши ожидания.