ДЗ №1

Специфичное для детектива, боевика – *пистолет*, *убийца*

Общеупотребительные – *говорить, красивый*

1. SpecC: НКРЯ (<http://ruscorpora.ru>), жанр текста – детская, тип текста – сказка, 1 166 113 словоупотреблений.
2. RefC: НКРЯ (<http://ruscorpora.ru>), жанр текста – детектив, боевик, 6 904 330 словоупотреблений.
3. <https://drive.google.com/open?id=1wiBzoJiPeNGox7Vo0pkPkk8_aDgmqWDoAahr0Zdi9jg>

Weirdness (wi) = frs(wi)/frr(wi) = (Ws / Ts) / (Wr / Tr)

frs(wi) – относительная частота слова в коллекции текстов определенной тематической области

frr(wi) – относительная частота слова в контрастной коллекции (reference corpus)

Ws – абсолютная частота wi в тематической коллекции

Ts – количество слов в тематической коллекции

Wr – абсолютная частота wi в контрастной коллекции

Tr - количество слов в контрастной коллекции

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| wi | Тип (специфичное vs. общеупотребитель-ное) | CountSpecC  (частота в Specific Corpora) | CountRefC (частота в контрастном корпусе  (Reference Corpora)) | LogLikelihood | Ранг | Weirdness | Ранг |
| пистолет | специфичное | 1 477 | 100 | 102.76 | 3 | 2.49 | 2 |
| убийца | специфичное | 1 470 | 8 | 390.31 | 1 | 31.03 | 1 |
| говорить | общеупотребительное | 13 333 | 2 614 | 46.99 | 4 | 0.86 | 3 |
| красивый | общеупотребительное | 1 583 | 522 | 155.69 | 2 | 0.51 | 4 |

1. Комментарий:

По weirdness специфичные слова получили высокую позицию в ранжированном списке, а общеупотребительные – низкую.

По loglikelihood одно из специфичных слов получило высокую позицию в ранжированном списке (которое имело бОльший разброс в абсолютных частотах), а другое – низкую. Аналогично и для общеупотребительных слов.

Таким образом, лучше с выделением тематических слов справилась статистика weirdness.