Кобзева Настя

группа БКЛ-132

Домашнее задание №1

**Определение степени специфичности слов**

1) SpecC: газетный корпус НКРЯ. Объем корпуса: 433 373 документа, 16 669 748 предложений, 228 521 421 слово.

2) RefC: устный корпус НКРЯ. Объем корпуса: 3 665 документов, 1 662 905 предложений, 11 349 008 слов.

3) Еще один метод, который можно использовать для выявления специфичной лексики – вычисление значения странности (weirdness). Weirdness(wi) = frs(wi)/frr(wi) = (Ws/Ts)/(Wr/Tr), где:

* frs(wi) – относительная частота слова в коллекции текстов определенной тематической области
* frr(wi) – относительная частота слова в контрастной коллекции
* Ws – абсолютная частота wi в тематической коллекции
* Ts – количество слов в тематической коллекции

4) Значение LogLikelihood (LL) было посчитано при помощи [вот этого калькулятора](http://ucrel.lancs.ac.uk/llwizard).

Таблица 1. Степень специфичности слов

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| wi | Тип | Count(SpecC) | Count (RefC) | LL | Ранг | Weirdness | Ранг |
| президент | специфичное | 233 891 | 3 650 | 7219.38 | 1 | 3,182377 | 1 |
| правительство | специфичное | 144 424 | 2 566 | 3796.36 | 2 | 2,795205 | 2 |
| поездка | общеупотребительное | 20 290 | 370 | 514.64 | 4 | 2,723399 | 3 |
| жизнь | общеупотребительное | 180 876 | 12 125 | 937.86 | 3 | 0,74085 | 4 |

5) Комментарий. В целом, более специфичные слова получили бо́льшие веса при вычислении LL и Weirdness. Характерные для новостных текстов слова ‘президент’ и ‘правительство’ заняли первые два места в ранжировании по данными параметрам.