Больгина Татьяна, 4 курс

ДЗ1

1. **Specific Corpora** Газетный корпус региональной и зарубежной прессы, объем - 13 291 019 слов, <http://ruscorpora.ru/search-regional.html>
2. **Reference Corpora** Основной корпус русского языка, объем - 265 401 717 слов, <http://ruscorpora.ru/search-main.html>
3. Метод 2: **Wierdness,** (wi) = frs(wi)/frr(wi) = (Ws / Ts) / (Wr / Tr), источник - презентация

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| wi | Тип (специфичное vs. общеупотребитель-ное) | CountSpecC  (частота в Specific Corpora) | CountRefC (частота в контрастном корпусе  (Reference Corpora)) | LogLikelihood | Ранг | Вес 2 | Ранг |
| эквилибр | специфичное | 2 | 14 | 1.78 | 3 | 2,852642901 | 1 |
| кресало | специфичное | 4 | 87 | 0.00 | 4 | 0,918091968 | 2 |
| знать | общеупотребитель-ное | 8 937 | 466 439 | 40.06 | 1 | 0,382597697 | 4 |
| говорить | общеупотребитель-ное | 16 169 | 553 066 | 6.74 | 2 | 0,583783276 | 3 |

1. Метод 2 weirdness оказался наиболее точным, так как специфичные слова получили более высокий вес и ранг соответственно. Метод LogLikelihood дал практически обратные результаты, где специфичные слова имеют меньший вес, чем общеупотребительные.

* *Wi-* слово *i;*

**\*Расчет метода 2**

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 коллекци

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **word** | **Ws** | **Ts** | **Ws / Ts** | **Wr** | **Tr** | **Wr / Tr** | **(Ws / Ts) / (Wr / Tr)** |
| 1 | 2 | 13291019 | 1,5048E-07 | 14 | 265401717 | 5,27502E-08 | 2,852642901 |
| 2 | 4 | 13291019 | 3,0096E-07 | 87 | 265401717 | 3,27805E-07 | 0,918091968 |
| 3 | 8937 | 13291019 | 0,00067241 | 466439 | 265401717 | 0,001757483 | 0,382597697 |
| 4 | 16169 | 13291019 | 0,00121654 | 553066 | 265401717 | 0,002083883 | 0,583783276 |