## ДЗ №1 Специфичность лексики

1) характеристики специфичного корпуса:

(1) признаки специфичного корпуса - жанр / тематику (любую другую характеристику, по которой Вы отбирали корпус);

*Специфичный корпус – публицистический подкорпус НКРЯ*

(2) источник текстов;

*НКРЯ*

(3) объем корпуса;

*108 739 638 слов*

(4) параметры предобработки: сырой / лемматизированный / с морфологической аннотацией …;

*С морфологической аннотацией*

2) характеристики контрастного корпуса (жанр/тематику…, объем, источник, доп. параметры)

*Подкорпус НКРЯ художественной литературы,*

*источник текстов – НКРЯ,*

*объем - 122 411 245 слов,*

*с морфологической аннотацией*

3) Для второго метода укажите формулу, источник - откуда Вы взяли формулу

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 слова в контрастной коллекции (reference corpus)

Tr – количество слов в контрастной коллекции (reference corpus)

4) Заполните таблицу 1:

Таблица 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| wi | Тип (тематическое vs. общеупотребительное) | CountSpecC  (частота в Specific Corpora Публицистический корпус) | CountRefC (частота в контрастном корпусе  (Reference Corpora) Корпус худ литры) | LogLikelihood | Ранг | Вес 2  (Weirdness) | Ранг |
| выборы | тематическое | 243.4 ipm | 20.5 ipm | 860.08 | 2 | 13.37 | 1 |
| политика | тематическое | 328 ipm | 35 ipm | 1130.38 | 1 | 10.5 | 2 |
| думать | общеупотребительное | 553.8 ipm | 1127.4 ipm | 458.38 | 3 | 0.55 | 4 |
| душа | общеупотребительное | 307.3 ipm | 447.8 ipm | 366.63 | 4 | 0.77 | 3 |

5) Напишите комментарии: насколько точен оказался результат применения весов

(тематическое слово – получило высокую позицию в ранжированном списке, а общеупотребительное - низкую)

*Как можно заметить по рангам, то высокие позиции заняли в обоих случаях тематические слова, а общеупотребительные заняли низкие позиции. Это говорит о том, что результат применения весов достаточно точен, хотя разные веса по-разному определяют места внутри тематических и внутри общеупотребительных. Второй вес показал наиболее тематическим словом – слово «выборы», а первый вес – слово «политика».*