**Зарифян Маша БКЛ-132**

Корпуса:

1. **Specific Corpora:** Нехудожественные тексты жанра “публицистика” НКРЯ (репортажи, хроники, информационные сообщения. Объём корпуса: 3 610 473 слова.

специфичные слова: *корреспондент, депутат*

1. **Reference Corpora:** Художественные тексты жанра “детектив, боевик” (подкорпус НКРЯ). Объём подкорпуса: 6 904 330 слов

Общие слова: *информация, событие*

|  |  |  |  |
| --- | --- | --- | --- |
| **специфичные слова** | **частота в Specific Corpora** | **частота в Reference Corpora** | **Всего** |
| *депутат* | 1572 | 611 | 2183 |
| *корреспондент* | 6188 | 221 | 6409 |

|  |  |  |  |
| --- | --- | --- | --- |
| **общие слова** | **частота в подкорпусе НКРЯ “детектив, боевик” (худож.)** | **частота в подкорпусе НКРЯ “публицистика” (нехудож.)** | **Всего:** |
| *информация* | 1729 | 369 | 2098 |
| *событие* | 827 | 2062 | 2889 |

**LogLikelihood: G2 = 2(a\*ln(a/E1) + b\*ln(b/E2)), где**

**Е1 = с\*(a+b)/(c+d); E2 = d\*(a+b)/(c+d)**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Депутат*** | Подкорпус | Другие тексты | Всего |
| Частота | 1572 | 611 | 2183 |
| Размер | 3610473 | 6904330 | 10 514 803 |

Е1 = 749,5777676; Е2 = 1433,422232

G2 = 1286,399629 = 1286,40

|  |  |  |  |
| --- | --- | --- | --- |
| ***Корреспондент*** | Подкорпус | Другие тексты | Всего |
| Частота | 6188 | 221 | 6409 |
| Размер | 3610473 | 6904330 | 10514803 |

E1 = 2200,66143; E2 = 4208,338565

G2 = 11492,55273 = 11492,55

|  |  |  |  |
| --- | --- | --- | --- |
| ***Информация*** | Подкорпус 1 | Подкорпус 2 | Всего |
| Частота | 1729 | 369 | 2098 |
| Размер | 6904330 | 3610473 | 10514803 |

Е1 = 1377,608724; Е2 = 720,3912764

G2 = 291,9164991 = 291, 92

|  |  |  |  |
| --- | --- | --- | --- |
| ***Событие*** | Подкорпус 1 | Подкорпус 2 | Всего |
| Частота | 827 | 2062 | 2889 |
| Размер | 6904330 | 3610473 | 10514803 |

Е1 = 1897,00267; Е2 = 991,9973296

G2 = 1644,383913 = 1644,38

**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

*Депутат:*

|  |  |  |
| --- | --- | --- |
| **Ws** | **Ts** | **FrSw** |
| 1572 | 3610473 | 0,0004354 |
| **Wr** | **Tr** | **FrRw** |
| 611 | 6904330 | 8,84952E-05 |

**Weirdness = 4,92**

*Корреспондент:*

|  |  |  |
| --- | --- | --- |
| **Ws** | **Ts** | **FrSw** |
| 6188 | 3610473 | 0,001713903 |
| **Wr** | **Tr** | **FrRw** |
| 221 | 6904330 | 3,20089E-05 |

**Weirdness = 53,5**

*Информация:*

|  |  |  |
| --- | --- | --- |
| **Ws** | **Ts** | **FrSw** |
| 369 | 3610473 | 0,000102203 |
| **Wr** | **Tr** | **FrRw** |
| 1729 | 6904330 | 0,000250423 |

**Weirdness = 0,4**

*Событие:*

|  |  |  |
| --- | --- | --- |
| **Ws** | **Ts** | **FrSw** |
| 2602 | 3610473 | 0,000720681 |
| **Wr** | **Tr** | **FrRw** |
| 827 | 6904330 | 0,00011978 |

**Weirdness = 6,01**

**ИТОГ:**

**LogLikelihood:**

*депутат*: 1286,40

*корреспондент*: 11492,55

*информация:*  291, 92

*событие:* 1644,38

**Weirdness:**

*депутат*: 4,92

*корреспондент*: 53,5

*информация:* 0,4

*событие:* 6,01

Сравнивая два способа вычислений, мы видим, что по результатам обоих подсчётов, наиболее специфичным является слово *корреспондент*. *Информация*, как и следовало ожидать, является наименее специфичным словом. При этом, слово *депутат,* которое на первый взгляд показалось специфичным, в итоге имеет коэффициент специфичности меньший, чем общеупотребительное слово *событие*. Возможно, результаты вычислений были бы чуть более точными, если бы размер коллекций текстов был примерно одинаковым, а не одна в два раза больше другой, как в нашем случае.

Как мне показалось, Weirdness является наиболее удобным и более наглядным способом вычисления.