HW1 Alexeeva Maria БКЛ132

корпус 1: детективный жанр художественного корпуса НКРЯ = 6 904 330

корпус 2: газетный корпус НКРЯ = 228 521 421

специфичные слова (для корпуса 1): детектив, сыщик

общие: закон, авария

|  |  |  |  |
| --- | --- | --- | --- |
|  | корпус 1 | корпус 2 | весь корпус (1+2) |
| частота | а | b | a+b |
| размер | c | d | c+d |

**LogLikelihood:**

G2 = 2(a\*ln(a/E1)+b\*ln(b/E2))

E1=c\*((a+b)/(c+d))

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

**Weirdness:**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **детектив** | корпус1 | корпус2 | корпус целиком |
| частота | 256 | 2 773 | 3 029 |
| размер | 6904330 | 228521421 | 235425751 |
|  | 88,83147014 | 2940,16853 | 217,272512 |
|  | E1 | E2 | LogLikelihood |

Weirdness (детектив) = 3,05559275

|  |  |  |  |
| --- | --- | --- | --- |
| **сыщик** | корпус1 | корпус2 | корпус целиком |
| частота | 458 | 2 873 | 3 331 |
| размер | 6904330 | 2,3E+08 | 2,35E+08 |
|  | 97,68822 | 3233,312 | 736,41 |
|  | E1 | E2 | LogLikelihood |

Weirdness (сыщик) = 5,276369815

|  |  |  |  |
| --- | --- | --- | --- |
| **закон** | корпус1 | корпус2 | корпус целиком |
| частота | 1 008 | 124 841 | 125 849 |
| размер | 6904330 | 2,3E+08 | 2,35E+08 |
|  | 3690,773 | 122158,2 | 2807,536 |
|  | E1 | E2 | LogLikelihood |

Weirdness (закон) = 0,267244434

|  |  |  |  |
| --- | --- | --- | --- |
| **авария** | корпус1 | корпус2 | корпус целиком |
| частота | 176 | 22 122 | 22 298 |
| размер | 6904330 | 228521421 | 2,35E+08 |
|  | 653,933351 | 21644,0666 | 636,2279 |
|  | E1 | E2 | LogLikelihood |

Weirdness (авария) = 0,263325947

|  |  |  |
| --- | --- | --- |
| **слово** | **LogLikelihood** | **Weirdness** |
| ***детектив*** | 217,272512 | 3,05559275 |
| ***сыщик*** | 736,41 | 5,276369815 |
|  |  |  |
| ***закон*** | 2807,536 | 0,267244434 |
| ***авария*** | 636,2279 | 0,263325947 |

Выводы:

Методы показывают разные результаты. С увеличением количества вхождений слова в корпус, автоматически увеличивается и LogLikelihood (видно, из таблиц, т.к. получилось, что частота вхождений выбранных мною специфичных слов для корпуса 1 – детективного оказалась больше в корпусе 2 – газетном). Но это можно объяснить тем, что второй корпус намного больше, чем первый. Однако метод LogLikelihood не учитывает это, а вот метод Weirdness это учитывает, соответственно его значения для этих слов меньше, чем для специфичных.