Отчет по ДЗ№1 «Тематический вес для подколлекции»

Мельник Анастасия

ФиКЛ 4 курс 2018

Цель: для некоторой тематической/жанрово-специфичной коллекции (**Specific Corpora**) оценить степень специфичности слов из корпуса по отношению к контрастной коллекции текстов (**Reference corpus**) двумя методами: **LogLikelihood** и **weirdness**.

Задача: Найти 10 наиболее и 10 наименее специфичных (тематически значимых) для тематического корпуса слов с помощью статистик LogLikelihood и weirdness.

Материал:

Тематический корпус – православные тексты о семье на современном русском языке (~ 10 тысяч слов).

Источник текстов: веб-страница lib.pravmir.ru.

Контрастный корпус – часть произведения Л.Н.Толстого «Анна Каренина» (~ 10 тысяч слов).

Предобработка:

* удаление пунктуации
* приведение к нижнему регистру
* токенизация
* лемматизация
* удаление стоп-слов

Таблица 1. Статистические характеристики первых 10 и последних 10 keywords согласно LogLikelihood метрике[[1]](#footnote-1).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| wi | Тип | CountSpecC  (частота в Specific Corpora) | CountRefC (частота в контрастном корпусе  (Reference Corpora)) | LogLikelihood | Ранг (из 3837) | Weirdness | Ранг (из 540) |
| брак | тематическое | 137 | 1 | 13.57 | 1 | 140.05 | 1 |
| любовь | тематическое | 57 | 5 | 7.25 | 2 | 11.65 | 3 |
| муж | тематическое | 64 | 11 | 6.55 | 3 | 5.95 | 27 |
| бог | тематическое | 61 | 11 | 6.29 | 4 | 5.67 | 28 |
| ребёнок | тематическое | 82 | 22 | 6.20 | 5 | 3.81 | 58 |
| должный | тематическое | 53 | 9 | 5.98 | 6 | 6.02 | 26 |
| жена | тематическое | 73 | 27 | 4.81 | 7 | 2.76 | 99 |
| дух | тематическое | 22 | 2 | 4.47 | 8 | 11.24 | 5 |
| супруг | тематическое | 20 | 2 | 4.19 | 9 | 10.22 | 8 |
| семейство | тематическое | 20 | 2 | 4.19 | 10 | 10.22 | 10 |
| присутствие | общеупотреб | 1 | 14 | -3.63 | 3818 | 0.073 | 536 |
| сделать | общеупотреб | 4 | 23 | -3.79 | 3819 | 0.178 | 523 |
| дверь | общеупотреб | 1 | 15 | -3.80 | 3820 | 0.068 | 538 |
| понимать | общеупотреб | 1 | 15 | -3.80 | 3821 | 0.068 | 537 |
| ах | общеупотреб | 1 | 16 | -3.96 | 3822 | 0.064 | 539 |
| место | общеупотреб | 1 | 17 | -4.11 | 3833 | 0.060 | 540 |
| лицо | общеупотреб | 9 | 38 | -4.32 | 3834 | 0.242 | 510 |
| глаз | общеупотреб | 2 | 22 | -4.37 | 3835 | 0.093 | 533 |
| брат | общеупотреб | 4 | 34 | -5.15 | 3836 | 0.120 | 531 |
| рука | общеупотреб | 3 | 35 | -5.57 | 3837 | 0.088 | 534 |

В топ-10 сверху попадают слова характерные для тематического корпуса, но не общеупотребительные, а в топ-10 снизу – общеупотребительные слова, не характерные для тематического корпуса.

Замечание: Статистики Loglikelihood и Weirdness довольно неплохо коррелируют друг с другом, но слова с более высокими обеими частотами, получают более низкий ранг для метрики weirdness, чем для метрики loglikelihood.

Эксперимент:

Вручную выделенные ключевые слова: брак, Бог, жена, муж, Христос, вечность, любовь, таинство, грех.

50% из топа-10 были выделены вручную, 50% не вошло в топ-10 из выделенных вручную.

Используемые статистики:

1. LogLikelihood (loglikelihood.llr)
2. Weirdness

[Лукашевский]

Пусть w – слово. Тогда

,

где

Ws - частотность слова в коллекции предметной области;

Ts – совокупная частотность слов в коллекции предметной области;

Wg -  частотность слова в контрастной коллекции;

Tg – совокупная частотность слов в контрастной коллекции.

1. Не рассматриваются слова, которые встретились только в одном из корпусов. [↑](#footnote-ref-1)