|  |  |  |
| --- | --- | --- |
|  | | |
| Федеральное государственное бюджетное  образовательное учреждение высшего образования «Новосибирский государственный технический университет» | | |
|  | | |
| Кафедра теоретической и прикладной информатики | | |
| Лабораторная работа №2 | | |
| по дисциплине «Статистический анализ нечисловых данных» | | |
| Место для ввода текста. | | |
|  | | |
|  | Бригада 11 | Мещанинов Григорий |
| Группа ПМ-14 | Хамитова екатерина |
| Вариант 11 |  |
|  |  |
|  |  |
|  |  |
| Преподаватель | Тимофеева анастасия юрьевна |
|  |  |
| Новосибирск, 2024 | | |

**Данные**

Исходный набор данных содержит текст учебника: Hastie T., Tibshirani R., Friedman J. H. The elements of statistical learning: data mining, inference, and prediction. – New York : Springer, 2009. – 745 c. Для него выполнена предварительная обработка:

* текст разбит на страницы,
* удалены переносы,
* текст разбит на предложения,
* удалены рисунки, формулы, цифры, пунктуация, предложения с одним словом,
* каждое предложение разбито на слова, удалены однобуквенные слова,
* небуквенные символы,
* все символы переведены в нижний регистр.

Результаты представлены в файле text\_sent\_1.txt. Каждая строка соответствует одному предложению. Слово с апострофом считается как одно слово (например, what’s).

**Задание 1.**

Из всего текста отобрать комбинации смежных слов (стоящие рядом в одном предложении), содержащие слово, соответствующее Вашему варианту (nearest). Если слово в таблице 1 указано как word1, то оно должно быть только на первой позиции.

На второй позиции может быть любое слово, которое вместе с word1 встречается в тексте не менее 5 раз.

Биграммы:

|  |  |  |
| --- | --- | --- |
| W1 | W2 | Частота |
| nearest | neighbor | 133 |
| nearest | neighbors | 82 |
| nearest | shrunken | 22 |
| nearest | neighborhood | 9 |
| nearest | centroid | 7 |

**Задание 2.**

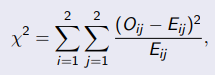
Проранжировать биграммы по частоте встречаемости. Сделать предварительные выводы о возможных коллокациях.

Предварительный вывод: чаще всего встречается биграмма «**nearest neighbor**» (ближайший сосед), на втором месте биграмма с почти таким же смыслом (ближайшие соседи). Первое, что пришло в голову, если не учитывать тот факт, что текст взят из книги о статистике и машинном обучении, это то, что текст о каком-то бытовом рассказе. Однако, подозрительно часто бы тогда употреблялось данное словосочетание (и вряд ли бы это было коллокацией). А если все же взять во внимание тематику текста, то «ближайший сосед» вполне можно считать коллокацией.

**Задание 3.**

Проранжировать биграммы, отобранные в п. 2, в соответствие с критериями Вашего варианта (в таблице 2 нужные критерии отмечены знаком +). 

Хи-2 Пирсона





|  |  |  |
| --- | --- | --- |
| W1 | W2 | Chi2 |
| nearest | neighbor | 85636.412677 |
| nearest | neighbors | 42433.333737 |
| nearest | shrunken | 9421.347758 |
| nearest | neighborhood | 1004.176095 |
| nearest | centroid | 798.255258 |

Критерий отношения правдоподобия

Критерий Фишера

**Задание 4.**

Проверить гипотезу об отсутствии взаимосвязи между словами в соответствие с критериями Вашего варианта (см. таблицу 1).

**Задание 5.**

Исходя из результатов пп. 1-4, сделать выводы о том, какие биграммы можно считать коллокациями.