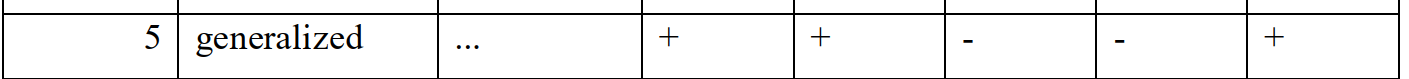
|  |  |  |
| --- | --- | --- |
|  | | |
| Федеральное государственное бюджетное  образовательное учреждение высшего образования «Новосибирский государственный технический университет» | | |
|  | | |
| Кафедра теоретической и прикладной информатики | | |
| Лабораторная работа № 2 | | |
| по дисциплине «Статистический анализ нечисловых данных» | | |
|  | | |
|  | | |
|  | Группа ПМ-13 |  |
| Бригада 5 | зиянуров артём |
| Вариант 5 | вострецова екатерина |
|  |  |
|  |  |
| Преподаватель | Тимофеева анастасия юрьевна |
|  |  |
|  |  |
| Новосибирск, 2024 | | |

# Данные

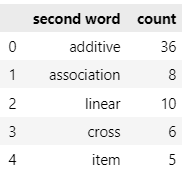
Исходный набор данных содержит текст учебника:  
Hastie T., Tibshirani R., Friedman J. H. The elements of statistical learning: data mining,  
inference, and prediction. – New York : Springer, 2009. – 745 c.  
Для него выполнена предварительная обработка:  
⎯ текст разбит на страницы,  
⎯ удалены переносы,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

# Ход работы

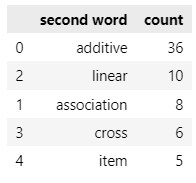
1. Из всего текста отобрать комбинации смежных слов (**стоящие рядом в одном  
   предложении**), содержащие слово соответствующее Вашему варианту (см. таблицу 1).  
   *Варианты по бригадам те же самые, что и для лабораторной работы №1.*Если слово в таблице 1 указано как word1, то оно должно быть только на первой позиции.  
   На второй позиции может быть любое слово, которое вместе с word1 встречается в тексте  
   **не менее 5 раз**.  
   Если слово в таблице 1 указано как word2, то оно должно быть только на второй позиции.  
   На первой позиции может быть любое слово, которое вместе с word2 встречается в тексте  
   **не менее 5 раз**



Слова, которые образуют нужные словосочетания находятся в столбце “second word”, а их количество в столбце “count”



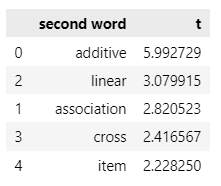
1. Проранжировать биграммы по частоте встречаемости. Сделать предварительные  
   выводы о возможных коллокациях.



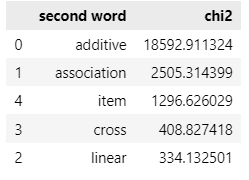
Предварительный вывод: По этой таблице видно, что в тексте чаще всего встречается словосочетание generalized additive (обобщенная добавка), а также generalized linear (обобщённый линейный). Кроме того, вероятно, данные словосочетания являются названиями понятий в этой предметной области, потому, возможно, они все являются коллокациями.

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

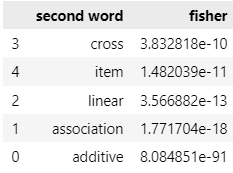
Ранжирование по критерию стьюдента:



По критерию хи-квадрат:



По критерию фишера:

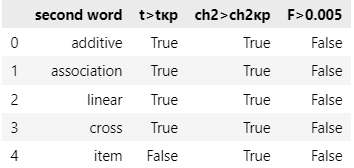


1. Проверить гипотезу об отсутствии взаимосвязи между словами в соответствие с  
   критериями Вашего варианта (см. таблицу 1), для показателя MI использовать  
   перестановочный критерий.

Критические значения для уровня значимости 0.99

tкр = 2.3263689962598413 (количество степеней свободы = Количество биграмм – 1 = 176550 – 1 = 176549)

χкр = 6.6348966010212145 (при количестве степеней свободы = 1)



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

Вывод: по вычисленным статистикам можно видеть, что критерий Фишера НЕ отвергается во всех случаях, однако хи-квадрат наоборот во всех случаях отвергает гипотезу об отсутствии взаимосвязи, критерий Стьюдента отвергает гипотезу лишь в одном случае, когда второе слово – item. Отсюда можно сделать вывод о том, что данные словосочетания связаны друг с другом. Возможно, словосочетание generalized linear описывает какое-то существительное(так как оба слова являются прилагательными), остальные же словосочетания, вероятно, описываются какое-то понятие из данной области, поэтому данные словосочетания(кроме слова item) являются коллокациями. нулевая гипотеза отвергается при всех словах, кроме слова item. Отсюда можно сделать вывод о связи слов и возможных коллокациях.