Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное   
учреждение высшего образования

«Новосибирский государственный технический университет»



Кафедра теоретической и прикладной информатики

Расчетно-графическая работа по курсу

**«СТАТИСТИЧЕСКИЙ АНАЛИЗ НЕЧИСЛОВЫХ ДАННЫХ»**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| https://psv4.userapi.com/c848132/u275429460/docs/d16/7a48be4e8485/Logo1.png?extra=xygsIQQwad0cqCKIxugjRIy9kd3JmRBfEKfWxcAvrfIKQ1CgI-Mr8IGFc60Gm-UtH3C0t8N-ws0mwmk1iKBnRdnqgFaqqJOX8cT3bxo_ADKOGBCCZ_VicmoU5cJzeVveeKnmjlfBib7DlanMmJKw6Dk2pw | Факультет: | ПМИ |  |  |
| Группа: | ПМИ-72 |  |  |
| Вариант: | 15 |  |  |
| Студент: | Антонов С. |  |  |
| Преподаватели: | Тимофеева А. Ю. |  |  |
|  |  |  |  |

Новосибирск

2020

1. Цель работы

Выявить комбинации с данным словом из корпуса Брауна. Провести расчеты и на их основании сделать выводы, какие комбинации можно назвать коллокациями.

1. Задание

1) Из всего массива данных отобрать строки с комбинациями слов, содержащими слово соответствующее варианту:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Вариант | Слово | Часть речи | t-test | chi2 | LHR | MI |
| 15 | home | N | + | + | + | + |

Слово может быть как на первой позиции (word1), так и на второй (word2).

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

2) Из массива, полученного в п.1, отобрать комбинации слов, соответствующих любому из речевых шаблонов, указанных в таблице. При этом нужно учитывать часть речи слова, соответствующего варианту. Проранжировать биграммы по частоте встречаемости. Сделать предварительные выводы о возможных коллокациях.

*Таблица 1. – Шаблоны использования частей речи в биграммах*

|  |  |  |  |
| --- | --- | --- | --- |
| Часть речи | | Обозначение | |
| Слово 1 (word 1) | Слово 2 (word 2) | Pos1 | Pos2 |
| Прилагательное | Существительное | J - adjective | N – noun |
| Существительное | Существительное | N - noun | N – noun |
| Глагол | Предлог | V - verb | I – preposition |
| Глагол | Существительное | V - verb | N – noun |

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

4) Проверить гипотезу об отсутствии взаимосвязи между словами в соответствие с критериями Вашего варианта (см. таблицу 2), кроме показателя MI (для него проверять гипотезу о значимости не нужно).

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

1. Ход работы

Из всего массива данных были отобраны строки с комбинациями слов, содержащими слово «number».

*Таблица 2. - Строки, содержащие слово «home»*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | word1 | pos1 | word2 | pos2 | O11 | O12 | O21 | O22 |
| 446 | a | D | home | N | 16 | 22755 | 433 | 886564 |
| 2125 | and | C | home | N | 8 | 27492 | 441 | 881827 |
| 3179 | at | I | home | N | 66 | 4964 | 383 | 904355 |
| 4433 | came | V | home | N | 10 | 587 | 439 | 908732 |
| 4947 | come | V | home | N | 7 | 374 | 442 | 908945 |
| 6059 | drove | V | home | N | 5 | 49 | 444 | 909270 |
| 7122 | for | I | home | N | 12 | 8858 | 437 | 900461 |
| 7323 | four | C | home | N | 5 | 339 | 444 | 908980 |
| 7406 | from | I | home | N | 16 | 4005 | 433 | 905314 |
| 7626 | go | V | home | N | 19 | 540 | 430 | 908779 |
| 8789 | home | N | . | . | 83 | 361 | 45390 | 863934 |
| 8790 | home | N | and | C | 17 | 427 | 20564 | 888760 |
| 8791 | home | N | at | I | 10 | 434 | 4620 | 904704 |
| 8792 | home | N | entertainment | N | 5 | 439 | 23 | 909301 |
| 8793 | home | N | for | I | 17 | 427 | 7994 | 901330 |
| 8794 | home | N | from | I | 9 | 435 | 4002 | 905322 |
| 8795 | home | N | in | I | 34 | 410 | 17953 | 891371 |
| 8796 | home | N | is | V | 5 | 439 | 9409 | 899915 |
| 8797 | home | N | of | I | 15 | 429 | 35484 | 873840 |
| 8798 | home | N | on | I | 10 | 434 | 5785 | 903539 |
| 8799 | home | N | or | C | 7 | 437 | 3131 | 906193 |
| 8800 | home | N | rule | N | 7 | 437 | 53 | 909271 |
| 8801 | home | N | run | N | 7 | 437 | 42 | 909282 |
| 8802 | home | N | runs | N | 25 | 419 | 13 | 909311 |
| 8803 | home | N | the | D | 6 | 438 | 58338 | 850986 |
| 8804 | home | N | to | T | 20 | 424 | 24779 | 884545 |
| 8805 | home | N | town | N | 5 | 439 | 199 | 909125 |
| 8806 | home | N | with | I | 8 | 436 | 6269 | 903055 |
| 12199 | new | J | home | N | 5 | 1050 | 444 | 908269 |
| 13029 | of | I | home | N | 8 | 33950 | 441 | 875369 |
| 13655 | on | I | home | N | 5 | 5802 | 444 | 903517 |
| 14196 | own | J | home | N | 6 | 716 | 443 | 908603 |
| 18797 | the | D | home | N | 36 | 69141 | 413 | 840178 |
| 18798 | the | D | home | Y | 5 | 69172 | 10 | 840581 |
| 23069 | went | V | home | N | 7 | 487 | 442 | 908832 |

Проранжируем строки по частоте встречаемости биграммы (О11)

*Таблица 3. - Строки, содержащие слово «home» проранжированные по частоте*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | word1 | pos1 | word2 | pos2 | O11 | O12 | O21 | O22 |
| 8789 | home | N | . | . | 83 | 361 | 45390 | 863934 |
| 3179 | at | I | home | N | 66 | 4964 | 383 | 904355 |
| 18797 | the | D | home | N | 36 | 69141 | 413 | 840178 |
| 8795 | home | N | in | I | 34 | 410 | 17953 | 891371 |
| 8802 | home | N | runs | N | 25 | 419 | 13 | 909311 |
| 8804 | home | N | to | T | 20 | 424 | 24779 | 884545 |
| 7626 | go | V | home | N | 19 | 540 | 430 | 908779 |
| 8790 | home | N | and | C | 17 | 427 | 20564 | 888760 |
| 8793 | home | N | for | I | 17 | 427 | 7994 | 901330 |
| 446 | a | D | home | N | 16 | 22755 | 433 | 886564 |
| 7406 | from | I | home | N | 16 | 4005 | 433 | 905314 |
| 8797 | home | N | of | I | 15 | 429 | 35484 | 873840 |
| 7122 | for | I | home | N | 12 | 8858 | 437 | 900461 |
| 4433 | came | V | home | N | 10 | 587 | 439 | 908732 |
| 8791 | home | N | at | I | 10 | 434 | 4620 | 904704 |
| 8798 | home | N | on | I | 10 | 434 | 5785 | 903539 |
| 8794 | home | N | from | I | 9 | 435 | 4002 | 905322 |
| 2125 | and | C | home | N | 8 | 27492 | 441 | 881827 |
| 8806 | home | N | with | I | 8 | 436 | 6269 | 903055 |
| 13029 | of | I | home | N | 8 | 33950 | 441 | 875369 |
| 4947 | come | V | home | N | 7 | 374 | 442 | 908945 |
| 8799 | home | N | or | C | 7 | 437 | 3131 | 906193 |
| 8800 | home | N | rule | N | 7 | 437 | 53 | 909271 |
| 8801 | home | N | run | N | 7 | 437 | 42 | 909282 |
| 23069 | went | V | home | N | 7 | 487 | 442 | 908832 |
| 8803 | home | N | the | D | 6 | 438 | 58338 | 850986 |
| 14196 | own | J | home | N | 6 | 716 | 443 | 908603 |
| 6059 | drove | V | home | N | 5 | 49 | 444 | 909270 |
| 7323 | four | C | home | N | 5 | 339 | 444 | 908980 |
| 8792 | home | N | entertainment | N | 5 | 439 | 23 | 909301 |
| 8796 | home | N | is | V | 5 | 439 | 9409 | 899915 |
| 8805 | home | N | town | N | 5 | 439 | 199 | 909125 |
| 12199 | new | J | home | N | 5 | 1050 | 444 | 908269 |
| 13655 | on | I | home | N | 5 | 5802 | 444 | 903517 |
| 18798 | the | D | home | Y | 5 | 69172 | 10 | 840581 |

Выберем те записи, которые подходят под шаблон из таблицы. В нашем случае подходят два шаблона.

*Таблица 4. – Таблица шаблонов подходящая нашему варианту*

|  |  |  |  |
| --- | --- | --- | --- |
| Часть речи | | Обозначение | |
| Слово 1 (word1) | Слово 2 (word2) | pos1 | pos2 |
| Прилагательное | Существительное | **J**  adjective | **N**  noun |
| Существительное | Существительное | **N**  noun | **N**  noun |
| Глагол | Существительное | **V**  verb | **N**  noun |

Результат:

*Таблица 5. – Строки подходящие по шаблону*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | word1 | pos1 | word2 | pos2 | O11 | O12 | O21 | O22 |
| 8802 | home | N | runs | N | 25 | 419 | 13 | 909311 |
| 7626 | go | V | home | N | 19 | 540 | 430 | 908779 |
| 4433 | came | V | home | N | 10 | 587 | 439 | 908732 |
| 4947 | come | V | home | N | 7 | 374 | 442 | 908945 |
| 8800 | home | N | rule | N | 7 | 437 | 53 | 909271 |
| 8801 | home | N | run | N | 7 | 437 | 42 | 909282 |
| 23069 | went | V | home | N | 7 | 487 | 442 | 908832 |
| 14196 | own | J | home | N | 6 | 716 | 443 | 908603 |
| 6059 | drove | V | home | N | 5 | 49 | 444 | 909270 |
| 8792 | home | N | entertainment | N | 5 | 439 | 23 | 909301 |
| 8805 | home | N | town | N | 5 | 439 | 199 | 909125 |
| 12199 | new | J | home | N | 5 | 1050 | 444 | 908269 |

Судя по таблице, наибольшая частота встречаемости у коллокаций *home runs,go home и came home*

1. **t-критерий проверки неравенства**

Для каждой биграммы была подсчитана t-статистика по формуле

Здесь – частота биграммы из слов w1 и w2

– частота встречаемости слова wi, N – общее число биграмм

В данной работе N для всех одинакова и равна

Для примера возьмем home runs

*Таблица 6. – Строки проранжированные по t-критерию*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Word1 | Word2 | O11 | O12 | O21 | O22 | N | t-test |
| home | runs | 25 | 419 | 13 | 909311 | 909768 | 4,998871236 |
| go | home | 19 | 540 | 430 | 908779 | 909768 | 4,300390093 |
| came | home | 10 | 587 | 439 | 908732 | 909768 | 3,072722611 |
| come | home | 7 | 374 | 442 | 908945 | 909768 | 2,57708377 |
| home | rule | 7 | 437 | 53 | 909271 | 909768 | 2,636139181 |
| home | run | 7 | 437 | 42 | 909282 | 909768 | 2,638136264 |
| went | home | 7 | 487 | 442 | 908832 | 909768 | 2,55633355 |
| own | home | 6 | 716 | 443 | 908603 | 909768 | 2,307162757 |
| drove | home | 5 | 49 | 444 | 909270 | 909768 | 2,225379521 |
| home | entertainment | 5 | 439 | 23 | 909301 | 909768 | 2,231110738 |
| home | town | 5 | 439 | 199 | 909125 | 909768 | 2,193130059 |
| new | home | 5 | 1050 | 444 | 908269 | 909768 | 2,006904102 |

Упорядочение практически не поменялось по сравнению с частотным анализом. Чем больше частота вхождения сочетания слова, тем больше значение критерия. В лидерах также коллокации *home runs ,go home и came home*

**Критерий Хи-квадрат Пирсона**

Общая формула критерия выглядит так:

Более подробно формула выглядит так:

Для примера рассчитаем критерий Хи-квадрат для home runs:

*Таблица 7. – Строки проранжированные по критерию Хи-квадрат*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Word1 | Word2 | O11 | O12 | O21 | O22 | N | Хи-квадрат |
| home | runs | 25 | 419 | 13 | 909311 | 909768 | 33668,95827 |
| go | home | 19 | 540 | 430 | 908779 | 909768 | 1272,203094 |
| came | home | 10 | 587 | 439 | 908732 | 909768 | 320,0610752 |
| come | home | 7 | 374 | 442 | 908945 | 909768 | 247,0019343 |
| home | rule | 7 | 437 | 53 | 909271 | 909768 | 1660,321098 |
| home | run | 7 | 437 | 42 | 909282 | 909768 | 2036,154272 |
| went | home | 7 | 487 | 442 | 908832 | 909768 | 187,4182992 |
| own | home | 6 | 716 | 443 | 908603 | 909768 | 89,5013255 |
| drove | home | 5 | 49 | 444 | 909270 | 909768 | 928,5999217 |
| home | entertainment | 5 | 439 | 23 | 909301 | 909768 | 1820,446527 |
| home | town | 5 | 439 | 199 | 909125 | 909768 | 241,3777123 |
| new | home | 5 | 1050 | 444 | 908269 | 909768 | 38,59890478 |

Упорядочение сильно поменялось по сравнению с частотным анализом и t-критерию. По данному критерию, коллокациями могут быть биграммы home runs и *home entertainment* .

1. **Критерий отношения правдоподобия**

Для каждой биграммы была подсчитана статистика отношения правдоподобия по формуле

,

*,*

*,*

*c12 = C(w1,w2), c1 = C(w1), c2 = C(w2),*

*, , .*

Для примера рассчитаем для large number

*Таблица 8. – Строки проранжированные по отношению правдоподобия*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | O11 | O12 | O21 | O22 | N | С1 | С2 | LR |
| home | runs | 25 | 419 | 13 | 909311 | 909768 | 444 | 38 | 333,8780816 |
| go | home | 19 | 540 | 430 | 908779 | 909768 | 559 | 449 | 124,8032194 |
| came | home | 10 | 587 | 439 | 908732 | 909768 | 597 | 449 | 51,45135509 |
| come | home | 7 | 374 | 442 | 908945 | 909768 | 381 | 449 | 37,24108482 |
| home | rule | 7 | 437 | 53 | 909271 | 909768 | 444 | 60 | 63,68584367 |
| home | run | 7 | 437 | 42 | 909282 | 909768 | 444 | 49 | 66,71153632 |
| went | home | 7 | 487 | 442 | 908832 | 909768 | 494 | 449 | 33,68491605 |
| own | home | 6 | 716 | 443 | 908603 | 909768 | 722 | 449 | 22,71211782 |
| drove | home | 5 | 49 | 444 | 909270 | 909768 | 54 | 449 | 42,92541189 |
| home | entertainment | 5 | 439 | 23 | 909301 | 909768 | 444 | 28 | 50,05355788 |
| home | town | 5 | 439 | 199 | 909125 | 909768 | 444 | 204 | 29,53655623 |
| new | home | 5 | 1050 | 444 | 908269 | 909768 | 1055 | 449 | 13,7259573 |

Упорядочение соответствует тому, что было получено в ходе частотного анализа. В лидерах коллокации *home runs, go home* и *came home*.

1. **Точечная взаимная информация**

Рассчитаем точечную взаимную информацию по формуле:

Пример расчета точечной взаимной информации для биграммы *home runs*:

*Таблица 9. – Строки проранжированные по точечной взаимной информации*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | O11 | O12 | O21 | O22 | N | С1 | С2 | PMI |
| home | runs | 25 | 419 | 13 | 909311 | 909768 | 444 | 38 | 10,39665197 |
| go | home | 19 | 540 | 430 | 908779 | 909768 | 559 | 449 | 6,105790571 |
| came | home | 10 | 587 | 439 | 908732 | 909768 | 597 | 449 | 5,084908504 |
| come | home | 7 | 374 | 442 | 908945 | 909768 | 381 | 449 | 5,218275265 |
| home | rule | 7 | 437 | 53 | 909271 | 909768 | 444 | 60 | 7,901187625 |
| home | run | 7 | 437 | 42 | 909282 | 909768 | 444 | 49 | 8,193368376 |
| went | home | 7 | 487 | 442 | 908832 | 909768 | 494 | 449 | 4,843555221 |
| own | home | 6 | 716 | 443 | 908603 | 909768 | 722 | 449 | 4,073675004 |
| drove | home | 5 | 49 | 444 | 909270 | 909768 | 54 | 449 | 7,551608123 |
| home | entertainment | 5 | 439 | 23 | 909301 | 909768 | 444 | 28 | 8,515296471 |
| home | town | 5 | 439 | 199 | 909125 | 909768 | 444 | 204 | 5,650226051 |
| new | home | 5 | 1050 | 444 | 908269 | 909768 | 1055 | 449 | 3,263468341 |

По данному критерию, ранжировка коллокаций напоминает ранжировку Хи-квадрат. В лидерах биграммы *home runs, go home* и *came home*.

1. **Средняя взаимная информация**

Для каждой биграммы была подсчитана средняя взаимная информация по формуле

Рассчитаем пример для *home runs*

*Таблица 10. – Строки проранжированные по средней взаимной информации*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | O11 | O12 | O21 | O22 | N | С1 | С2 | MI |
| home | runs | 25 | 419 | 13 | 909311 | 909768 | 444 | 38 | 0,000264729 |
| go | home | 19 | 540 | 430 | 908779 | 909768 | 559 | 449 | 9,89554E-05 |
| came | home | 10 | 587 | 439 | 908732 | 909768 | 597 | 449 | 4,07954E-05 |
| come | home | 7 | 374 | 442 | 908945 | 909768 | 381 | 449 | 2,95281E-05 |
| home | rule | 7 | 437 | 53 | 909271 | 909768 | 444 | 60 | 5,0496E-05 |
| home | run | 7 | 437 | 42 | 909282 | 909768 | 444 | 49 | 5,2895E-05 |
| went | home | 7 | 487 | 442 | 908832 | 909768 | 494 | 449 | 2,67085E-05 |
| own | home | 6 | 716 | 443 | 908603 | 909768 | 722 | 449 | 1,80083E-05 |
| drove | home | 5 | 49 | 444 | 909270 | 909768 | 54 | 449 | 3,40352E-05 |
| home | entertainment | 5 | 439 | 23 | 909301 | 909768 | 444 | 28 | 3,96871E-05 |
| home | town | 5 | 439 | 199 | 909125 | 909768 | 444 | 204 | 2,34193E-05 |
| new | home | 5 | 1050 | 444 | 908269 | 909768 | 1055 | 449 | 1,08832E-05 |

Упорядочение соответствует тому, что было получено в ходе частотного анализа. В лидерах коллокации *home runs, go home*  и *home run*.

1. Проверка критериев

Используемый уровень значимости α = 0,0005, степень свободы: (N – 1)

Квантиль распределения Стьюдента равен 3,480768949.

Квантиль χ2-распределения равен 12,11566515.

Квантиль отношения распределения равен 12,11566515.

Для каждого критерия приведена сводная таблица

*Таблица 11. – Сводная таблица для t-критерия*

|  |  |  |
| --- | --- | --- |
| word1 | word2 | t-test |
| home | runs | 4,998871 |
| go | home | 4,30039 |
| came | home | 3,072723 |
| come | home | 2,577084 |
| home | rule | 2,636139 |
| home | run | 2,638136 |
| went | home | 2,556334 |
| own | home | 2,307163 |
| drove | home | 2,22538 |
| home | entertainment | 2,231111 |
| home | town | 2,19313 |
| new | home | 2,006904 |

Можно выделить две биграммы – *home runs, go home.* Остальные биграммы имеют низкую частоту вхождения. Это показывает, что критическое значение больше, сравниваемого значения, поэтому нулевая гипотеза принимается.

*Таблица 12. – Сводная таблица для хи-квадрат*

|  |  |  |
| --- | --- | --- |
| word1 | word2 | Хи-квадрат |
| home | runs | 33668,96 |
| go | home | 1272,203 |
| came | home | 320,0611 |
| come | home | 247,0019 |
| home | rule | 1660,321 |
| home | run | 2036,154 |
| went | home | 187,4183 |
| own | home | 89,50133 |
| drove | home | 928,5999 |
| home | entertainment | 1820,447 |
| home | town | 241,3777 |
| new | home | 38,5989 |

Для всех биграмм статистика хи-квадрат превышает критическое значение. Нулевая гипотеза отвергается.

*Таблица 13. – Сводная таблица для LHR*

|  |  |  |
| --- | --- | --- |
| word1 | word2 | LR |
| home | runs | 333,8781 |
| go | home | 124,8032 |
| came | home | 51,45136 |
| come | home | 37,24108 |
| home | rule | 63,68584 |
| home | run | 66,71154 |
| went | home | 33,68492 |
| own | home | 22,71212 |
| drove | home | 42,92541 |
| home | entertainment | 50,05356 |
| home | town | 29,53656 |
| new | home | 13,72596 |

Для всех биграмм статистика хи-квадрат превышает критическое значение. Для них гипотеза об отсутствии взаимосвязи отвергается.

1. Выводы

На основании всех анализов, коллокацией можно считать две пары – *home runs, go home*.

Наиболее информативной оказалась проверка гипотезы на основе t-критерия. Она позволила выделить две биграммы, образующие коллокацию. Остальные проверки гипотезы оказались малоинформативными, показав практически все значения.