Московский авиационный институт
Факультет прикладной математики и физики

Лабораторная работа №5

по курсу:
«Обработка естественно-языковых текстов»
по теме:
«Поиск коллокаций»
2 семестр

Студент: Ахмед С. Х.
Преподаватель: Калинин А. Л.
Группа: 8О-106М

Москва, 2019 г

Постановка задачи

Необходимо найти коллокации в имеющемся корпусе, использованного для построения поисковой системы (или его случайному подмножеству достаточного размера). Для поиска коллокаций необходимо использовать как минимум два статистических алгоритма из рассмотренных на лекциях. Сравнить выделенные коллокации между собой, пояснить различия с точки зрения использованых алгоритмов. В отчёте должны быть приведены количество найденных коллокаций, оценка точности метода, примеры найденных коллокаций и ошибочно найденных словосочетаний.

Ход работы

Для поиска коллокаций было решено использовать метод поиска по биграммам(частотность) и расчет взаимной информации. Было решено для данной лабораторной работы воспользоваться половиной своего корпуса для осуществления работы. Поиск по биграммам заключается в том, что будет подсчитано количество всех словосочетаний и наиболее частовстречаемые будут признаны коллокациями. Поиск коллокаций по частным словосочетаниям признается не самым эффективным способом, но все равно его стоит проверить и оценить его эффективность.

	0	1
0	во время	22320
1	см также	21333
2	jpg thumb	18598
3	из за	16078
4	том числе	14801
5	при этом	14465
6	категория фильмы	14423
7	один из	14044
8	то есть	12569
9	так как	12469
10	на территории	11585
11	одним из	10734
12	примечания литература	10713
13	из них	10534
14	том что	10072
15	thumb px	9763
16	кроме того	9691
17	российской федерации	9663
18	таким образом	9570
19	несмотря на	9372
20	примечания ссылки	9275
21	настоящее время	8950
22	января года	8652
23	может быть	8520
24	то что	8345
25	не только	8234
26	декабря года	8057
27	xix века	7777
28	до года	7609
29	не менее	7510
..	...	...
70	of the	4729
71	того как	4683
72	по данным	4651
73	после смерти	4640
74	состоит из	4590
75	представляет собой	4582
76	тем что	4549
77	по год	4537
78	на то	4503
79	зависимости от	4456
80	году был	4439
81	на год	4365
82	из самых	4355
83	после чего	4273
84	но не	4235
85	под названием	4221
86	того что	4200
87	могут быть	4099
88	великой отечественной	4072
89	алфавиту категория	4008

Полученный список заполнен случайными фразами, которые нельзя отнести к коллокациям. Однако можно найти коллокации и тут: “великой отечественной”, “российской федерации”,”таким образом”,”несмотря на”.

При расчете взаимной информации обнаружился недостаток метода: если словосочетание встречается 1 раз, то величина взаимной информации будет максимальной. Начало списка найденных коллокаций (у всех словосочетаний одинаковый уровень взаимной информации):

```
''айыл аймагы 1 22.384427007026
''возвращенье сознанья 1 22.384427007026
''глаз осьминога 1 22.384427007026
''дети сбыслава 1 22.384427007026
''дж э.боулта 1 22.384427007026
''живой 'нейтральный 1 22.384427007026
''загрубели обветрели 1 22.384427007026
''зарядити зарядитися 1 22.384427007026
''игорь лосиевский 1 22.384427007026
''киборги меклар 1 22.384427007026
''курт лонгджон 1 22.384427007026
''марина кудимова 1 22.384427007026
''марки помгола 1 22.384427007026
''менандр 'мена́ндр 1 22.384427007026
''меня восхищают 1 22.384427007026
''морис родригес 1 22.384427007026
''морфизм с-пучков 1 22.384427007026
''мёртвый 'защёлка 1 22.384427007026
''несплошные синусоидные 1 22.384427007026
''никита пустосвят 1 22.384427007026
''оби-хасами фигурно 1 22.384427007026
''окончатые висцеральные 1 22.384427007026
''округ куин-эннс 1 22.384427007026
''олег мраморнов 1 22.384427007026
''олимпиада çёнтермех 1 22.384427007026
''отряжати отряживати 1 22.384427007026
''паразиты иткул 1 22.384427007026
''писатель н.помяловский 1 22.384427007026
''погибла гофолия 1 22.384427007026
''проспект перемоги 1 22.384427007026
''разряда розряда 1 22.384427007026
''разрядити розрядити 1 22.384427007026
''разрядный розрядный 1 22.384427007026
''разряжати розряжати 1 22.384427007026
''рид ротчайлд 1 22.384427007026
''рыбообразные трилариане́ц 1 22.384427007026
''рядецъ рядца 1 22.384427007026
''рядовичь рядовникъ 1 22.384427007026
''рядовое оговоренная 1 22.384427007026
''синонимы хризопал 1 22.384427007026
''союз лівих 1 22.384427007026
''срядитися срядитися 1 22.384427007026
''сьогодення софіївки 1 22.384427007026
''ф̀лойд гондолли 1 22.384427007026
''яңа әлифба/jаңа 1 22.384427007026
''єремія́ ярема 1 22.384427007026
''інокентій анненський 1 22.384427007026
''історичні етюди 1 22.384427007026
'-уд- -юд- 1 22.384427007026
'-ъ -ь 1 22.384427007026
'-ыни кънѧгыни 1 22.384427007026
'11д613 рд-215м 1 22.384427007026
'11д614 рд-216м 1 22.384427007026
'15-я 'ван 1 22.384427007026
'16-я 'чарльз 1 22.384427007026
'17-я 'барри 1 22.384427007026
'19-я 'гарольд 1 22.384427007026
'20-я 'эдмунд 1 22.384427007026
'26-я 'фрэнк 1 22.384427007026
'27-я 'эдмонд 1 22.384427007026
```

'арка́дная платфо́рма 1 22.384427007026
'ассирия ашшур-раби 1 22.384427007026
'астрапей молний́ный 1 22.384427007026
'асхаб аль-бадр 1 22.384427007026
'атаргатис атаргатида 1 22.384427007026
'афины архипп 1 22.384427007026
'ахом сукапхаа 1 22.384427007026
'балбачан 'болбачан 1 22.384427007026
'балласт голл 1 22.384427007026
'василе́й базиле́вс 1 22.384427007026
'басмала 'тасмия 1 22.384427007026
'бассейнов 'федеративная 1 22.384427007026
'батманов зейнудин 1 22.384427007026
'беатри́с мариа́на 1 22.384427007026
'бейскбаев бахтурас 1 22.384427007026
'бенжи гэйтер 1 22.384427007026
'бергамот 'бессемянка 1 22.384427007026
'бессемянка 'вильямс 1 22.384427007026
'биайнили 'ва́нское 1 22.384427007026
'биджиев солтан-хамид 1 22.384427007026
'биопсия хориона 1 22.384427007026
'бли́жняя испа́ния 1 22.384427007026
'боваль 'двое 1 22.384427007026
'болеслав кривоустый 1 22.384427007026
'бортово́й самопи́сец 1 22.384427007026
'боя́рская ду́ма 1 22.384427007026
'брацигово бяга 1 22.384427007026
'бу́рой ржа́вчиной 1 22.384427007026
'бульвар уилшир 1 22.384427007026
'ва внеоборотные 1 22.384427007026
'ва́нское ца́рство 1 22.384427007026
'вагела вирдхавала 1 22.384427007026
'вал1 вал27 1 22.384427007026
'вале́рия и́горевна 1 22.384427007026
'валенти́н козьми́ч 1 22.384427007026
'валу́евский циркуля́р 1 22.384427007026
'валы 'клапана 1 22.384427007026
'ван хефлин 1 22.384427007026
'варбелов 'дёлиц 1 22.384427007026
'варвар 'колдун 1 22.384427007026
'василия автомоновича 1 22.384427007026
'ведущие 'муз-тв 1 22.384427007026
'вели́кая сфе́ра 1 22.384427007026
'венад падманабха 1 22.384427007026
'вернер клиннер 1 22.384427007026
'виконта годериха 1 22.384427007026
'вилчек 'вилкзек 1 22.384427007026
'вильям дергам 1 22.384427007026
'вильямс 'дюшес 1 22.384427007026
'вио́ла виоли́нья 1 22.384427007026
'вклад 'спасение 1 22.384427007026
'вогодого недега 1 22.384427007026
'вогу́лы вогуличи 1 22.384427007026
'волк заподозривший 1 22.384427007026
'вор вре́мени 1 22.384427007026
'ворскла 'сокол 1 22.384427007026
'впо-504 апс-м 1 22.384427007026
'вывод выведем 1 22.384427007026
'га́лицкие русофи́лы 1 22.384427007026
'гаври́лович кро́тов 1 22.384427007026
'газ-3102 нижневальный 1 22.384427007026
'газ-61-40 'газ-21 1 22.384427007026
'гао фададьо 1 22.384427007026
'гаркловский вовкулак 1 22.384427007026

Среди них можно увидеть имена собственные, какие-то случайные номерные названия. Однако можно увидеть и коллокации: “глаз осьминога”,”аркадная платорфма”,”великая сфера”,”вклад спасение” и так далее. Стоит заметить, что высокочастотные термины данный алгоритм обрабатывает уверенно:

Out[83]:
```
[(('во', 'и'), -8.18545645045836),
 (('в', 'того'), -7.736918287880556),
 (('в', 'этого'), -7.636003364663512),
 (('в', 'войны'), -7.498189327935901),
 (('году', 'года'), -7.3662222908921855),
 (('в', 'население'), -7.287264429851739),
 (('по', 'что'), -7.2667361588292465),
 (('во', 'на'), -7.214526406198978),
 (('в', 'мира'), -7.2087248830919),
 (('для', 'года'), -7.184932244602425),
 (('что', 'года'), -7.086161445268107),
 (('в', 'литература'), -7.073631568156934),
 (('к', 'из'), -7.050639902527568),
 (('из', 'к'), -7.050639902527568),
 (('не', 'году'), -7.044262729663458),
 (('году', 'году'), -6.977720314987774),
 (('в', 'территории'), -6.959724227121967),
 (('в', 'над'), -6.95917284357202),
 (('с', 'время'), -6.930771389916941),
 (('а.', 'на'), -6.880578016157941),
 (('во', 'в'), -6.8138029830019455),
 (('в', 'имеет'), -6.7977021644263615),
 (('по', 'из'), -6.717472384862493),
 (('категория', 'и'), -6.646564469796591),
 (('о', 'по'), -6.622175421233248),
 (('от', 'году'), -6.621811543469942),
 (('в', 'даже'), -6.6144546401445155),
 (('не', 'и'), -6.555370559770601),
 (('с', 'же'), -6.522108876624639),
 (('как', 'году'), -6.515512564050564),
 (('согласно', 'в'), -6.471759288925689),
 (('при', 'по'), -6.463543132091942),
 (('под', 'с'), -6.452478861930185),
 (('о', 'года'), -6.441600707672109),
 (('в', 'одним'), -6.435748753159995),
 (('перед', 'в'), -6.434162502268041),
 (('в', 'деревня'), -6.41096369012584),
 (('со', 'в'), -6.3955596630442955),
 (('года', 'году'), -6.3662222908921855),
 (('в', 'имя'), -6.352540339212396),
 (('не', 'также'), -6.3303252463186155),
 (('де', 'в'), -6.293401701511399),
 (('в', 'де'), -6.293401701511399),
 (('был', 'что'), -6.252300903118702),
 (('в', 'количество'), -6.2353826381535455),
 (('о', 'из'), -6.223774112176869),
 (('из', 'о'), -6.223774112176869),
 (('в', 'франции'), -6.223493483368127),
 (('на', 'во'), -6.214526406198978),
 (('от', 'по'), -6.190888232935492),
 (('примечания', 'с'), -6.188647166747064),
 (('за', 'году'), -6.186435781954163),
 (('на', 'у'), -6.177769823782896),
 (('на', 'он'), -6.177509094192089),
 (('таким', 'в'), -6.175881626404884),
 (('к', 'также'), -6.166027038773553),
 (('в', 'между'), -6.144861608604636),
 (('в', 'всех'), -6.132525257210499),
 (('в', 'владимир'), -6.131546827169004),
```

Как мы видим это коллокациями не является, что верно и отображено в оценке MI теста.

Выводы

Оба примененных метода имеют свои недостатки: при простом подсчете частот большая вероятность ошибочного определения коллокации из-за того, что существует большое число слов, которые просто часто употребляются друг с другом. Также на частотность тех или иных словосочетаний влияет стиль текстов: новости и художественные произведения имеют разные наборы слов и свои устойчивые словосочетания.

Метод подсчета взаимной информации показал неэффективность в очень редких словосочетаниях: если каждое слово встретилось по 1 разу и только друг с другом (как это часто бывает с редкими именами и фамилиями), то по этому алгоритму, найденное словосочетание будет коллокацией.