**Московский авиационный институт  
(национальный исследовательский университет)**

Факультет прикладной математики и физики  
Кафедра вычислительной математики и программирования

**Лабораторная работа №5  
по курсу «Обработка естественно-языковых текстов»**

Поиск коллокаций

|  |  |  |
| --- | --- | --- |
|  | **Студент:** | Рожлейс И.А. |
| **Группа:** | М8О-108М-18 |
| **Преподаватель:** | Калинин А.Л. |
| **Дата:** | 27.09.2019 |

Москва, 2019

**Цель работы**

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 использованных алгоритмов.

В отчёте должны быть приведены количество найденных коллокаций, оценка точности метода, примеры найденных коллокаций и ошибочно найденных словосочетаний.

**Ход работы**

Решение проблемы в данной работе было проведено в два этапа. Первый этап заключается в создании некоторого рода хранилища куда помещается статистика частоты использования строковых значений: в данном случае монограмм, или токенов, и биграмм – словосочетаний. В качестве хранилища выступают объекты классов Collector и BigramCollector наследующий его. Объект класса Collector, как более общий класс хранит токены. Класс BigramCollector более специализирован для хранения биграмм, однако полностью реализует функциональность класса Collector.

Данные хранятся внутри отображения стандартной библиотеки шаблонов std::map. Ключом является строковое значение, а значение отображения представляет собой количество таких строк внутри корпуса документов.

Документы из корпуса обрабатывается последовательно. Перед обработкой текст документа нормализуется. Нормализация текста заключается в удалении особых символов UTF-8 таких как EN\_DASH, EM\_DASH, MINUS\_SIGN и т.д. Разделителем токенов считается любой символ из " ,.!?@$#&%\_/\*-+|<>(){}[]:;=`\\\"\'\x0A\x0D", а регистр токена понижается.

Корпус состоит из ~65 тыс. документов из которых 30 тыс. документов были обработаны. Всего было определено 7,4 млн. токенов; из них 0,3 млн. уникальных. Также было определено 4,0 млн. биграмм (словосочетаний); из них 1,9 млн. уникальных.

На **табл. 1** представлены 30 наиболее часто встречаемых токенов. Биграммы, содержащие хотя бы один токен из этой таблицы не обрабатываются. Это сделано с целью удаления эффекта зашумлённости вывода программы, когда наиболее часто встречаемые токены наиболее часто образуют словосочетания.

|  |  |  |
| --- | --- | --- |
|  | **Количество** | **Токен** |
| **1** | 253808 | в |
| **2** | 200186 | и |
| **3** | 86624 | категория |
| **4** | 86199 | на |
| **5** | 84176 | с |
| **6** | 52600 | вид |
| **7** | 50099 | примечания |
| **8** | 49098 | из |
| **9** | 47050 | год |
| **10** | 43862 | а |
| **11** | 41590 | автор |
| **12** | 34892 | заглавие |
| **13** | 33545 | от |
| **14** | 33502 | по |
| **15** | 32709 | животные |
| **16** | 32036 | у |
| **17** | 31587 | род |
| **18** | 30830 | до |
| **19** | 30522 | или |
| **20** | 27591 | к |
| **21** | 26327 | издание |
| **22** | 26211 | году |
| **23** | 25499 | см |
| **24** | 24794 | ссылка |
| **25** | 24494 | не |
| **26** | 24286 | издательство |
| **27** | 23185 | место |
| **28** | 22701 | статья |
| **29** | 22089 | как |
| **30** | 21689 | для |

**Табл. 1.** Наиболее часто употребляемые токены

В целом данный результат достаточно верно согласуется с данными предоставленными НКРЯ (см. Источники), хотя и имеют ряд специфических включений слов. Например, «категория», «примечание», «автор», «заглавие» и т.д. Наличие данных слов в списке наиболее употребляемых слов объясняется особым форматированием текста документов, где каждый документ имеет несколько секций, которые имеют соответствующее название.

На **табл. 2** и **табл. 3** представлены наиболее часто встречаемые токены (за исключением токенов из **табл. 1**) и наиболее часто встречаемые биграммы.

|  |  |  |
| --- | --- | --- |
|  | **Количество** | **Токен** |
| **1** | 20393 | книга |
| **2** | 19900 | мм |
| **3** | 19890 | насекомые |
| **4** | 19618 | рода |
| **5** | 18888 | м |
| **6** | 18026 | также |
| **7** | 17870 | видов |
| **8** | 17573 | язык |
| **9** | 17416 | том |
| **10** | 17402 | таксон |
| **11** | 17350 | что |
| **12** | 16917 | страницы |
| **13** | 16742 | тип |
| **14** | 16614 | описание |
| **15** | 15689 | ссылки |
| **16** | 15027 | при |
| **17** | 13649 | литература |
| **18** | 13439 | растения |
| **19** | 12960 | описанные |
| **20** | 12924 | около |
| **21** | 12826 | длина |
| **22** | 12738 | номер |
| **23** | 12710 | был |
| **24** | 12315 | ответственный |
| **25** | 11943 | часть |
| **26** | 11745 | более |
| **27** | 11502 | файл |
| **28** | 10280 | их |
| **29** | 10100 | журнал |
| **30** | 9933 | вида |

**Табл. 2.** Часто употребляемые токены

|  |  |  |
| --- | --- | --- |
|  | **Количество** | **Биграмма** |
| **1** | 8766 | тип журнал |
| **2** | 4224 | заголовок добавлен |
| **3** | 4224 | добавлен ботом |
| **4** | 4112 | впервые описан |
| **5** | 3838 | был впервые |
| **6** | 3501 | северной америки |
| **7** | 3455 | длина тела |
| **8** | 3093 | южной америки |
| **9** | 3001 | выпуск том |
| **10** | 2961 | определитель насекомых |
| **11** | 2857 | литература книга |
| **12** | 2847 | номер страницы |
| **13** | 2380 | честь людей |
| **14** | 2357 | описание мелкие |
| **15** | 2349 | насекомые азии |
| **16** | 2337 | заголовок курсивом |
| **17** | 2124 | образ жизни |
| **18** | 2071 | книга определитель |
| **19** | 2018 | том числе |
| **20** | 1949 | во время |
| **21** | 1911 | новая зеландия |
| **22** | 1876 | том страницы |
| **23** | 1830 | северная америка |
| **24** | 1823 | южная америка |
| **25** | 1820 | словарь названий |
| **26** | 1798 | названий животных |
| **27** | 1796 | пятиязычный словарь |
| **28** | 1789 | млн лет |
| **29** | 1775 | страницы страниц |
| **30** | 1763 | книга пятиязычный |

**Табл. 3.** Часто употребляемые биграммы

После сбора статистики встречаемости токенов и биграмм программа реализует статистическую выборку биграмм по двум алгоритмам: t-критерий Стьюдента и метод максимального правдоподобия.

t-критерий Стьюдента показывает какую значимость имеют отклонения от средней величины выборки:

Метод максимального правдоподобия (ММП) в свою очередь позволяет сравнить правдоподобие двух гипотез:

На **табл. 4** представлен результат работы обоих алгоритмов. В таблице представлены 30 лучших результатов по оценкам алгоритмов.

|  |  |  |
| --- | --- | --- |
|  | **t-критерий Стьюдента** | **ММП** |
| **1** | *естественных врагов* | лесов рыбы |
| **2** | *горный массив* | архей файл |
| **3** | двум изолированным | распространён лишь |
| **4** | фактор гомеобокса | *насекомых эффект* |
| **5** | гинура греческого | это верхние |
| **6** | здесь квалификация | *образует один* |
| **7** | κατα здесь | получить что |
| **8** | здесь перечислено | ящерица что |
| **9** | *широкопирамидальная крона* | клюва что |
| **10** | группировок являясь | же общий |
| **11** | пигидиальной пластинкой | насекомые комары |
| **12** | дарвина вьюрков | отсутствуют расположены |
| **13** | тромбином фактор | растениях но |
| **14** | здесь подозрительные | же организма |
| **15** | высокую незаконную | длине была |
| **16** | результатам выходило | видом африка |
| **17** | вигна отнесены | это хромосомы |
| **18** | *ответный фактор* | *растений занимает* |
| **19** | лейкемию фактор | растений такое |
| **20** | *здесь размещалась* | даже представители |
| **21** | рыбами попугаями | также апоптоз |
| **22** | *архейный фактор* | также палеонтология |
| **23** | *тяжёлых суглинках* | *насекомые нашли* |
| **24** | *постепенному развитию* | *очень декоративные* |
| **25** | тяжёлых врождённых | … виды |
| **26** | лекарственный дягиль | тысяч были |
| **27** | произраставшим здесь | ряд все |
| **28** | фактор трансфактор | жизни рядом |
| **29** | проявлял высокую | образования были |
| **30** | высокую трихомонада | часть главный |

**Табл. 4.** Результат работы статистических алгоритмов

**Заключение**

Была проведена обработка 30 тыс. документов; или 7,4 млн. токенов (из них 2,9 млн. уникальных); или 4,1 млн. словосочетаний (из них 1,9 млн. уникальных). Всего алгоритм t-критерия Стьюдента определил 0,5 млн. биграмм, а алгоритм ММП в свою очередь 1,9 млн. биграмм. Исходя из выборки полученных результатов можно сделать вывод что статистический алгоритм t-критерия Стьюдента (27%) справился в данном случае лучше, чем алгоритм ММП (17%).

**Источники**

[1] Маннинг, К. Введение в информационный поиск // Вильямс, 2014. – 528 с.

[2] Введение в обработку текстов [Электронный ресурс] // URL: <http://tpc.at.ispras.ru/wp-content/uploads/2011/10/lecture4-20142.pdf>

[3] How can I sort a std::map first by value, then by key? [Электронный ресурс] // URL: <https://stackoverflow.com/questions/19842035/how-can-i-sort-a-stdmap-first-by-value-then-by-key>

[4] Список частотности по НКРЯ [Электронный ресурс] // URL: [https://‌ru.wiktionary.org/wiki/Приложение:Список\_частотности\_по\_НКРЯ](https://‌ru.wiktionary.org/‌wiki/Приложение:Список_частотности_по_НКРЯ)