Качество данных IQBuzz и рекомендации по их использованию

Методическая записка, выполненная в рамках проекта РНФ № 15-18-00091

О.Кольцова, М.Апишев, О.Нагорный

2016

Формирование выборки.

Задача формирования выборки была получить тексты, посвященные разным этническим группам, за период один год. Для этого был составлен словарь из 146 этнонимов + 4563 биграмм (включая биграммы типа "киргизская девочка"). Период закачки: 01.01.2011 по 01.11.2015. Региональный фильтр: Россия. Закачка осуществлялась через публичный API в виде ряда файлов xml, которые затем были конвертированы в формат json, т.к. из этого формата удобнее формировать базы данных. Единая база данных MongoDB была сформирована из этой выборки.

Состав исходной коллекции

Разбивка по источникам получилась: 80.6% - ВКонтакте (4766761); 6.7% - Twitter (394060); 3% - Google+ (175213); 1.8% - liveinternet (107211); 1% - ursa-tm.ru (58294); 0.4% - блоги «Эха Москвы» (20729). Все прочие источники были объединены в один общий.

Исходный объём словаря: 8326323

Исходное количество документов: 5915196

Исходная длина коллекции: 3535290923 слов

Исходное количество авторов постов: 912129

Предобработка коллекции

Вся информация рассчитана на коллекции, которая была получена после парсинга базы и лемматизации текстов. Соответственно, если какие-то записи имели нечитаемый формат, они были отброшены сразу и нигде в статистиках не фигурируют. Таких документов оказалось ещё 153999 (сверх описанного выше).

В коллекции в тегах, расположенных рядом с ключевыми словами поиска (этнонимами), находилось много нечитаемых слов. После того, как все слова из тегов были выгружены (их было 6008), лемматизированы тем же инструментом, что и коллекция (pymystem3), они были профильтрованы по порогу встречаемости 5, почищены от лишних символов, определенных вручную, и дополнены этнонимами из предподготовленного списка.

Итоговый набор этнонимов для моделирования включил в себя 588 этнонимов (и постсоветстких, и международных). Вся работа велась только с униграммами.

Предобработка коллекции с точки зрения геотегов и меток времени.

Геотеги в IQBuzz не имеют стандартного формата и соответствуют формату тех источников, из которых они берутся. Поэтому для них характерно разнообразие. Унификация геотегов производилась путем сравнения их с базой геотегов ВКонтакте. В результате было распознано 97% геотегов. Далее, все посты из малочисленных источников получили общие источник и геотег; все посты с некириллическим геотегом получили общий геотег; из меток времени была извлечена только дата в одном формате. Города на основе базы ВКонакте были приписаны субъектам РФ; все тексты, имеющие нероссийский геотег, были сгруппированы в один геотег. Всего сверх 85 российских геотегов получилось 13 нероссийских или сгруппированных из нераспознанных, они включили %%% текстов.

Фильтрация словаря (удаление стоп-слов).

Были отброшены слова:

а) встречающиеся в коллекции меньше 150 раз

б) встречающиеся в коллекции чаще 1 млн раз

в) с длиной меньше 4 символов

г) с длиной более 30 символов

д) содержащие что-то, кроме кириллицы.

Корректно считанные в самом начале документы в дальнейшем не удалялись, только преобразовывались. Удалены могли быть слишком короткие документы из только редких слов, которые в результате фильтрации словаря стали пустыми.

Статистика итоговой коллекции

Итоговый объём словаря: 75363 + 10819 отфильтрованных по коллекции биграмм с этнонимами

Итоговое число геотегов: 98

Итоговое число меток времени: 715

Итоговое количество документов: 5911992

Итоговая длина коллекции: 1550824224

Средняя длина документа: 262 слова

Более детальная статистика по длинам:

1-10 - 13%

10-100 - 32%

100-200 - 12%

200-500 - 23%

500-1000 - 15%

1000+ - 3%

Ручное тестирование.

Проводилось на коллекции 7181 текст, каждый из которых был прочитан тремя независимыми кодировщиками. Единогласно были признаны понятными только 4947 текст. Непонятные тексты представляют собой, как правило, тексты на других языках, чаще всего кириллических, либо отрывки из тредов комментариев без начала и конца. Далее, было выявлено 6383 текста, в которых хотя бы один кодировщик усмотрел наличие этнонима (этнонимов), и 4121 текста, единогласно понятных, в которых все кодировщики усмотрели хотя бы по одному этнониму. Таким образом, существует погрешность IQBuzz в поиске по ключевым словам.

Выводы и рекомендации по использованию.

Данные IqBuzz, насколько можно судить по проведенному исследованию, могут быть использованы для получения выборок достаточно высокого качества, однако достижение этой задачи требует большого количества технических компетенций, компьютерных мощностей и времени. Сырые данные IQBuzz не подходят для использования в исследованиях и практически не посильны для подготовки социальными исследователями.

Исходные данные подлежат множественной чистке по множеству параметров: метки места (геотеги), времени, наличие ключевого слова, язык и др. В результате коллекция может сильно уменьшиться; доля ужатия зависит от характера коллекции. Существенный объем работы связан с разными форматами данных, обусловленными множественностью источников. Если, как в нашем случае, подавляющее большинство текстов приходится на Вконтакте, можно рекомендовать ограничиться только этим источником. Поскольку данные социальных сетей в принципе довольно грязные и содержат большую долю мало осмысленных текстов, не очень перспективной оказывается задача выделения пропорций тех или иных типов текстов или тем по общей коллекции. Можно рекомендовать отсечение очень коротких текстов. Следует понимать, что IQBuzz и другие агрегаторы данных со временем меняют алгоритмы сбора данных. Поэтому мониторинг изменения чего-либо по этим данным на длинных промежутках времени затруднен. На коротких промежутках это возможно, однако рекомендуется консультироваться с представителями агрегаторов на предмет недавних изменений в сборе данных.