УДК 004.622

Разработка механизма автоматического формирования качественных датасетов из исходного множества данных

Акрамовский Р. Н.
(Владимирский государственный университет им. А. Г. и Н. Г. Столетовых, г. Владимир)
Научный руководитель – Воронова Н. М.
(Владимирский государственный университет им. А. Г. и Н. Г. Столетовых, г. Владимир)

В докладе представлен опыт практической разработки механизма формирования срезов данных, удовлетворяющих заданным критериям качества на примере формирования датасетов из исходного множества погодных данных.

Введение

Качество данных охватывает целый ряд аспектов. В связи с этим, обычно выделяют уровни качества, при которых одни аспекты оказываются более значимыми, чем другие. Важность каждого аспекта определяется задачами, которые должны быть решены в ходе анализа этих данных. Некорректность в данных может привести к тому, что исходный датасет окажется частично или полностью непригодным к аналитической обработке или, хуже того, будет казаться достоверным, но приведет к неправильным выводам.

Основная часть

Качество данных (data quality) — степень, с которой набор характеристик, присущих данным, отвечает конкретным требованиям с точки зрения их применения. Неправильно выстроенные уровни качества данных непосредственно влияют на успех проекта: можно либо задать слишком высокий уровень и не достигнуть его, либо установить слишком низкий уровень и тогда будет потерян смысл системы аналитики.

На практике наиболее важными критериями оценки исходных данных является два критерия – отсутствие пропусков и целостность (соответствие значений всех данных определенному непротиворечивому набору правил). Если данные хранятся в БД, то проверка на целостность сводиться к проверке на отсутствие дубликатов и противоречий в связях, где под дубликатами понимаются записи в БД с разными уникальными идентификаторами (id), но полностью идентичными значениями остальных полей (включая связные данные из других таблиц), а под противоречиями – нарушение логики в данных (в том числе в связных данных из других таблиц).

Имеется множество прогнозных и текущих погодных данных для разных локаций (данные размещены в БД). Требуется получить из этого множества различные срезы - датасеты (например, прогнозы по температуре на 5 дней для нескольких локаций или прогнозы по температуре на 1,2,3,5 и 10 дней для одной локации и т.д.). Полученные датасеты не должны содержать пропуски, противоречия и дубликаты и должны включать только реальные записи (автозаполнение пропусков не допускается).

Анализ качества исходного множества данных показал существенное количество ошибок (для анализа был взят срез данных по одной локации – город Владимир с интервалом дат с июня 2017 года по ноябрь 2020 года), а именно:

1) дубликаты - на одну и ту же дату, с одним временем суток, локацией и погодной характеристикой, есть несколько разных записей (для одной локации было выявлено 27534 таких записи, что составляет 10.73% от общего количества записей)
2) пропуски - отсутствуют прогнозные данные (для одной локации было выявлено 72605 записей, что составляет 17.62% от общего количества записей)

3) противоречия - присутствуют прогнозы на прошедшие даты (для одной локации была выявлена 301 запись, что составляет 0.1% от общего количества записей)

Поэтому необходимо было разработать механизм, который позволит автоматически получать нужные срезы данных требуемого качества (без противоречий, пропусков и дубликатов).

Выводы

В результате был реализован механизм для автоматического формирования качественных срезов данных: разработаны соответствующие Java- и Python- приложения. Java-приложение выгружает данные по каждой локации в .csv файл. Далее Python-приложение вырезает из выгруженных данных все интервалы данных, в которых есть записи, содержащие пропуски и противоречия.

Акрамовский Р.Н. (автор) Подпись

Воронова Н.М. (научный руководитель) Подпись