DataSet

Задание

В данной лабораторной работе требуется сформировать набор данных. Для этого необходимо распарсить какой-нибудь ресурс.

- Выберите веб-сайт, который вы будете парсить. Сайт должен содержать список каких-нибудь однотипных объектов в неструктурированном виде. Если вы будете парсить готовую таблицу или json, то за это будут снижены баллы.
- Наборы данных должны быть уникальны. Если вы парсите один и тот же сайт с другим студентом, то у вас должны различаться подкатегории объектов на этом сайте. Для это существует специальная таблица, в которую необходимо предварительно записаться: https://docs.google.com/spreadsheets/d/12yeZwQpYavaW8J-2kYcmWwVRGbtuMb2YlRcKALFnA_g/
- Набор данных должен содержать как минимум 2 категориальных и 2 числовых признака. Всего должно быть не менее 6 признаков.
- Набор данных должен содержать не менее 500 строк (объектов).
- Набор данных может содержать другие типы данных: текст, картинки, аудио, видео, ряды и т.д. Они могут пригодиться в соответствующих лабораторных работах. Если вам не хватает текстовых и категориальных признаков, то необходимо их извлечь в рамках данной лабораторной работы.
- На стадии парсинга запрещено отбрасывать объекты или признаки с пропусками, отбрасывать аномальные объекты, заменять аномальные или пропущенные значения, сливать несколько разных значений категории в одно, пытаться нормализовать значения.
- На стадии парсинга необходимо унифицировать единицы измерения и «очищать» числовые значения от форматирования, например: превращать «1 234 567 м.» или «1,234.567 км.» в «1234567». Единицы измерения нужно сохранить в названии признака.
- На стадии парсинга необходимо унифицировать одинаковые значения одной категории, например: превращать «Cat», «CAT» или «кот» в «cat».
- Набор данных необходимо сохранить в сыром виде в tsv формате. Затем преобразовать в arff формат с определением типов признаков и описанием.
- Набор данных необходимо предобработать: выбрать целевой категориальный признак, заполнить пропуски, преобразовать не целевые категории в числа, нормализовать набор данных. Это необходимо делать после сохранения в tsv и arff форматы. Данные после преобразования можно тоже сохранить, например в csv формат.
- Код и наборы данных необходимо загрузить в github: https://classroom.github.com/a/i5ZS6Ifr

Вы можете использовать любые вспомогательные библиотеки. Например, requests для краулинга и lxml для парсинга в python.

Рекомендации

- Выбирайте сайт с умом. На этих данных вы будете обучать алгоритмы, которые реализуете в следующих лабораторных работах.
- Желательно разделить процесс парсинга на две стадии: получение html-кода страниц и последующий их разбор.
- Не рекомендуется скачивать подряд все html-страницы, необходимо ограничивать число запросов в секунду до 3-х или меньше. Также рекомендуется использовать прокси и указывать User-Agent, Cookie и другие заголовки, чтобы избежать или максимально отсрочить бан.
- Не рекомендуется парсить по минимуму. Если сайт содержит больше объектов или признаков, то их тоже желательно включить в набор данных. Но сильно много объектов (больше 100 000) тоже не хорошо.
- Если нужно проматывать сайт для дозагрузки объектов, то это можно сделать программно. Откройте консоль браузера через Inspect code, во вкладке Консоль напишите js код для промотки. Найти такой код в интернете легко.
- Лучше всего сортировать объекты по популярности, если на сайте имеется такая функция. Если вы будете парсить сайты с фильмами, то не стоит брать слишком новые или ещё не вышедшие фильмы, так как для них будет меньше информации.
- Данные других типов хранятся в отдельной папке, в датасете хранятся пути к файлам. Например, если у каждой записи датасета есть картинка, то хранить в таблице стоит путь к ней. Саму картинку хранить файлом в папке /pics/. Простой текст можно хранить внутри набора данных.