DataSet

Формирование набора данных

В данной лабораторной работе требуется сформировать собственный набор данных. Для этого можно использовать один из трёх вариантов получения данных.

Использование готовых табличных или структурированных данных

- Данные должны быть «сырыми».
- Данные не должны быть уже типизированы.
- Не используйте популярные наборы данных, такие как ирисы или титаник.
- В идеале использовать таблицы вида: https://en.wikipedia.org/wiki/Comparison_of_file_systems

Парсинг интернет ресурса

- Сайт должен содержать список каких-нибудь однотипных объектов в неструктурированном виде.
- Вы можете использовать любые вспомогательные библиотеки. Например, requests для краулинга и lxml для парсинга в python.

Ручное формирование набора данных

- Запрещено использовать информацию, которую можно автоматически спарсить как в предыдущем варианте.

Общие требования

- Наборы данных должны быть уникальны. Например, если вы парсите один и тот же сайт с другим студентом, то у вас должны различаться подкатегории объектов на этом сайте. Для этого существует специальная таблица, в которую необходимо предварительно записаться: https://docs.google.com/spreadsheets/d/15pgPh-bVkIVkvJUnNVPTrmTgqg6dfL8k3OVaOrAshvQ/
  При этом нельзя, чтобы ваши наборы данных «совпадали» по признакам. Например, наборы данных с одного сайта с квартирами в Москве и в Санкт-Петербурге считаются одинаковыми, но с продажей и арендой квартир — разными.
- Набор данных должен содержать как минимум 2 категориальных и 2 числовых признака. Если вы парсите веб-сайты, то всего должно быть не менее 10 признаков. Если вы используете готовые табличные данные, то не менее 20 признаков.
- Набор данных должен содержать не менее 20 объектов (строк). Если вы парсите веб-сайты, то всего должно быть не менее 200 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 сбора данных 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 сбора данных 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 сбора данных необходимо унифицировать одинаковые значения одной категории, например: превращать «Cat», «CAT» или «кот» в «cat».
- Если вы собираете текстовые данные, то обрабатывайте спец.символы и переносы строк.

Задание

- Соберите набор данных. Сформируйте из него таблицу. Сохраните сырые данные в tsv формате в файле **data.tsv**.
- Типизируйте данные. Сохраните их в файлах **data.arff** и **data.json** в соответствующих форматах. JSON-файл должен следовать той же схеме, что и набор данных из примера: https://drive.google.com/file/d/1GKKtOZ41AHs3uZljMI9Gq4CgTREMpLZt/
- После типизации и сохранения типизированных данных их необходимо предобработать: выбрать целевой категориальный признак, заполнить пропуски, преобразовать нецелевые категории в числа, нормализовать набор данных. Сохраните полученные данные в файле **data.csv**.
- Код и наборы данных необходимо загрузить в github: https://classroom.github.com/a/WFDks129

Рекомендации

- Выбирайте источник данных с умом. На этих данных вы будете обучать алгоритмы, которые реализуете в следующих лабораторных работах.
- Не рекомендуется парсить «по минимуму». Если источник содержит больше объектов или признаков, то их тоже желательно включить в набор данных. Но сильно много объектов (больше 100 000) тоже нехорошо.
- 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.
- 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.