Урок 25. Виды датасетов, источники данных и большие данные

7 класс · Урок 25 из 34 · 45 минут

Виды датасетов, источники данных и большие данные: урок для 7 класса о форматах, происхождении, разнообразии и пригодности набора для задачи.

В папке могут лежать записи, рисунки и звук, но количество файлов мало говорит об их пользе. Разберём виды датасетов и соберём небольшой каталог материалов: что в нём есть, откуда это взялось и каких данных не хватает.

▼ Содержание урока

1. Цель урока
2. Ход урока
3. Этап 1. Вход в тему: одна папка
4. Этап 2. Форма данных и происхождение
5. Этап 3. Когда данных становится слишком много
6. Этап 4. Каталог, которому можно доверять
7. Этап 5. Итоги урока

▼ 01. Цель урока

Научиться оценивать датасет по виду данных, происхождению и условиям применения, а не только по размеру.

К концу урока ученики смогут:

- Различать табличные, текстовые, графические, звуковые наборы и временные ряды.
- Объяснять, откуда берутся данные и зачем описывать их происхождение.
- Объяснять трудности объёма, скорости и разнообразия больших данных.
- Описывать доступные материалы, находить пробелы и планировать пополнение набора.

▼ 02. Ход урока

Этап	Мин	Учитель	Ученики
1. Вход в тему: одна папка	5	Собирает названия реальных видов материалов, не просит показывать файлы.	Объясняют, какая структура важна для их примеров.
2. Форма данных и происхождение	10	Разбирает виды данных и сведения об источнике.	Объясняют, какую структуру сохраняет каждый вид.
3. Когда данных становится слишком много	10	Связывает трудности big data с разнообразием речи.	Различают размер набора и покрытие условий.
4. Каталог, которому можно доверять	14	Помогает собрать каталог и проверить происхождение материалов.	Находят материалы, описывают связи и пробелы коллекции.
5. Итоги урока	6	Обсуждает общие выводы и ответы на итоговые вопросы.	Формулируют знания и объясняют ответы.

03. Этап 1. Вход в тему: одна папка 5 минут

Говорим ученикам

Посмотрите, какие виды учебных материалов вы действительно использовали за последнюю неделю: заметки, рисунки, запись объяснения, результаты измерений. Назовите два вида и объясните, почему их неудобно хранить как один неразличимый столбец. Личные файлы открывать и показывать не нужно.

Датасет не обязан быть прямоугольной таблицей. Разные виды датасетов сохраняют разную структуру: порядок во времени, текст, форму или звук. Сегодня выберем способ организации, а затем разберёмся, почему происхождение и качество важнее одного количества записей.

04. Этап 2. Форма данных и происхождение 10 минут

Говорим ученикам

В табличном наборе записи имеют общие поля. В текстовом важны слова и контекст. В коллекции изображений нужны сами изображения и сведения об их содержании или условиях съёмки. В звуковом наборе может храниться запись вместе с расшифровкой. Временной ряд — последовательность наблюдений с отметками времени; без порядка он теряет часть смысла.

Эти виды могут сочетаться. В одном наблюдении бывают изображение, описание и время. Если разорвать их связь при копировании, программа получит несоответствующие друг другу части. Поэтому идентификаторы и описание структуры нужны не меньше, чем сами файлы.

Данные получают измерением, наблюдением, опросом, оцифровкой документов, работой сервисов или генерацией в моделируемой среде. Синтетические данные — созданные программой примеры. Они полезны, когда отражают нужные свойства, но их происхождение нужно сохранять: созданная картинка не подтверждает, что событие произошло.

Открытый доступ к файлу не означает, что разрешён любой способ использования. Сначала выясняют, кто собрал данные, каким методом, для какой цели и на каких условиях ими делится. Из набора убирают ненужные личные сведения. Название файла само по себе не подтверждает надёжность данных.



Кто собрал · Как получили · Условия использования

Данные поступают из разных источников и имеют разную форму. Для любого набора нужно знать, кто и как его собрал и на каких условиях разрешено использование.

05. Этап 3. Когда данных становится слишком много

10 минут

Говорим ученикам

Большие данные (big data) — наборы данных, для которых нужны особые способы хранения и обработки. Причиной может быть большой объём, высокая скорость поступления или разнообразие данных. Единой границы в количестве строк нет: задача и доступные средства тоже влияют на трудность. Непрерывный поток может быть сложным даже тогда, когда отдельная запись мала.

Много одинаковых записей не заменяют разнообразие. Если записать одного говорящего тысячи раз, набор вырастет, но не охватит другие голоса. Нужно учесть условия, в которых распознают речь: темп, произношение, микрофон и фоновый шум. Файлы могут открываться без ошибок и всё же не подходить для нужной задачи.

29 ноября 2017 года Mozilla объявила первый выпуск Common Voice — открытого набора добровольно предоставленной речи. В нём было около 400 тысяч записей, примерно 500 часов. Разработчикам нужны были не только часы звука, но и расшифровки и разные участники: иначе система могла хуже понимать часть пользователей.

Реальный случай: часы речи и разные голоса

Mozilla собирала записи добровольцев для развития распознавания речи. В сообщении о выпуске объяснялось, что ограниченность доступных данных мешала новым разработчикам, а разнообразие голосов было отдельной задачей.

Вывод: масштаб и разнообразие решают разные проблемы. Один показатель размера не сообщает, какие люди и условия представлены в наборе.

Для учителя

Числа относятся к выпуску 2017 года, а не к текущему объёму Common Voice. Не называйте многоязычным именно этот первый выпуск: в сообщении он описан как англоязычный.

06. Этап 4. Каталог, которому можно доверять 14 минут

Говорим ученикам

Источник и вид данных — разные вещи. Рисунок и подпись к нему могут быть двумя частями одной записи. А фотография страницы и переписанный с неё текст не становятся двумя независимыми наблюдениями. Сейчас проверим это на доступных нам учебных материалах.

Задание. Соберите маленькую коллекцию

В группе выберите тему предстоящей контрольной, выступления или общего обсуждения. Найдите 4–6 подходящих фрагментов в материалах, которые уже доступны на занятии. Это могут быть собственные записи, рисунки и разрешённые учителем материалы. Копировать или загружать их в сервис не нужно.

1. Составьте каталог на бумаге или в редакторе: номер фрагмента, что он содержит, вид данных, где его найти и кто его создал. Если части связаны, дайте им общий номер.
2. Выберите один вопрос по теме и найдите по каталогу материал для ответа. Отметьте, что удалось найти, а чего в коллекции нет.
3. Один участник временно закрывает описание у выбранной записи, оставляя сам материал. Остальные определяют, что ещё можно понять и какой смысл потерялся. Верните описание и уточните его.

4. Решите, можно ли передать эту коллекцию другой группе: что разрешено использовать, что нужно спросить у автора, какие личные сведения убрать. Назовите один источник, который закрыл бы найденный пробел.

Составьте каталог, по которому можно найти нужный материал. Отметьте, чего в нём не хватает. Критерии: материалы можно найти; источник прослеживается; связанные части не перепутаны; пробел обнаружен по конкретному вопросу; условия использования понятны.

Для учителя

Собственных записей двух участников достаточно для двух источников, но они могут повторять один и тот же материал. Спросите, что действительно добавил второй источник. Если в классе нет цифровых файлов, каталог описывает бумажные фрагменты: это полноценный вариант работы. Фотографирование, поиск в интернете и регистрация не требуются. Не нужно искусственно набирать все виды данных. Если материал не подходит к выбранному вопросу, его можно исключить с объяснением. Большие данные здесь не создаются: после работы спросите, что изменилось бы при непрерывном поступлении множества разных материалов.

07. Этап 5. Итоги урока 6 минут

Говорим ученикам

Данные имеют форму, происхождение и границы применимости. Большие данные добавляют трудности обработки, но не отменяют этих требований. Умение выбрать и описать подходящий набор часто важнее, чем быстро накопить больше файлов.

Главное за урок

- Вид датасета зависит от того, какую структуру нужно сохранить: поля, текст, изображение, звук или порядок наблюдений.
- Происхождение данных позволяет оценить условия сбора, качество и допустимое использование.
- Большие данные создают трудности объёма, скорости поступления и разнообразия.
- Размер набора не гарантирует, что в нём представлены нужные условия и пользователи.

Итоговые вопросы

Ответьте коротко. Затем обсудим объяснения двух-трёх участников.

1. Почему звуковые записи и их расшифровки нельзя перемешать независимо друг от друга?
2. Какие вопросы о происхождении нужно задать перед использованием чужого датасета?

3. Почему тысяча почти одинаковых примеров может быть менее полезна, чем меньший, но разнообразный набор?
4. Что может усложнять обработку больших данных кроме количества файлов?

Для учителя

1. Исчезнет соответствие звука и правильного текста; обучающие пары станут неверными.
2. Кто, как, когда и для какой цели собрал данные; какие условия и разрешения действуют.
3. Разнообразный набор может покрывать больше условий нужной задачи; количество копий не добавляет таких условий.
4. Скорость поступления, разные форматы, необходимость связывать и обрабатывать обновления.

Принимаются ответы по смыслу. Важна связь с общим знанием, а не воспроизведение деталей реального случая.

Для учителя: источники

1. Mozilla. Initial Release of Speech Recognition Model and Voice Dataset. Дата материала: 29.11.2017. Материал старше 2024 года. Подтверждающая фраза: «nearly 400,000 recordings, representing 500 hours of speech».