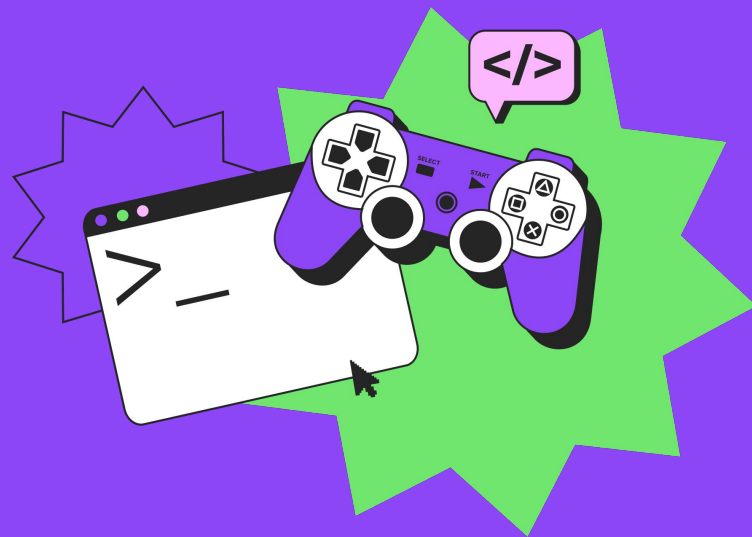


Инструменты разметки данных.

Семинар 9

Сбор и разметка данных





Сбор и разметка данных

1

Основы клиент-серверного взаимодействия. Парсинг API.

2

Парсинг HTML. BeautifulSoup.

3

СУБД MongoDB и ClickHouse в Python

4

Парсинг HTML. XPath.

5

Scrapy.

6

Scrapy. Парсинг фото и файлов.

7

Selenium в Python.




8

Работа с данными.

9

Инструменты разметки наборов данных.

Что будет на уроке сегодня

-  Использование инструментов разметки данных в зависимости от требований проекта.
-  Применение методов разметки данных на основе правил для автоматической генерации размеченных данных.
-  Реализация подходов к разметке данных с участием человека.





Викторина



Активное обучение - это метод, который:

1. включает в себя разметку всех точек данных в датасете
2. Случайно выбирает точки данных для разметки
3. Выбирает наиболее информативные точки данных для разметки
4. Помечает точки данных на основе заранее определенных правил



Активное обучение - это метод, который:

1. включает в себя разметку всех точек данных в датасете
2. Случайно выбирает точки данных для разметки
3. Выбирает наиболее информативные точки данных для разметки
4. Помечает точки данных на основе заранее определенных правил



Какой метод обычно используется для оценки неопределенности в активном обучении?

1. Кластеризация К-средних
2. Анализ главных компонент (PCA)
3. Разметка на основе правил
4. Оценка энтропии



Какой метод обычно используется для оценки неопределенности в активном обучении?

1. Кластеризация К-средних
2. Анализ главных компонент (PCA)
3. Разметка на основе правил
4. Оценка энтропии



Какой тип инструмента разметки данных позволяет итеративно улучшать прогнозы модели благодаря обратной связи с человеком?

1. Human-in-the-loop
2. Разметка на основе правил
3. Активное обучение
4. Semi-supervised learning



Какой тип инструмента разметки данных позволяет итеративно улучшать прогнозы модели благодаря обратной связи с человеком?

1. **Human-in-the-loop**
2. Разметка на основе правил
3. Активное обучение
4. Semi-supervised learning



Какой тип инструмента разметки данных предполагает разделение задачи разметки на микрозадачи и распределение их между группами работников?

1. Разметка на основе правил
2. Краудсорсинг
3. Активное обучение
4. Human-in-the-loop



Какой тип инструмента разметки данных предполагает разделение задачи разметки на микрозадачи и распределение их между группами работников?

1. Разметка на основе правил
2. Краудсорсинг
3. Активное обучение
4. Human-in-the-loop



В чем заключается основная проблема разметки данных с помощью краудсорсинга?

1. Обеспечение высокого качества разметки
2. Управление большими объемами данных
3. Балансировка рабочей нагрузки между маркировщиками
4. Обеспечение конфиденциальности и безопасности данных



В чем заключается основная проблема разметки данных с помощью краудсорсинга?

1. **Обеспечение высокого качества разметки**
2. Управление большими объемами данных
3. Балансировка рабочей нагрузки между маркировщиками
4. Обеспечение конфиденциальности и безопасности данных



Вопросы?

Вопросы?



Вопросы?





Практика



Задание 1

- Загрузите датасет в pandas DataFrame.
- Определите функцию разметки на основе правил, которая присваивает метки на основе определенных условий или закономерностей в данных.
- Примените функцию разметки для автоматической разметки подмножества данных.



20 минут



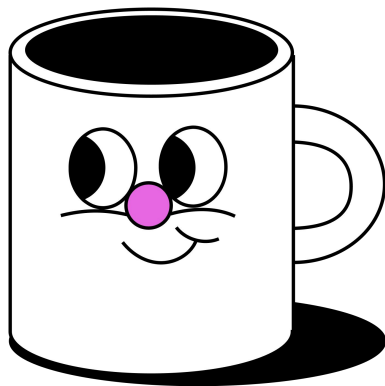
Задание 1

- Загрузите датасет tweets.csv в pandas DataFrame.
- Выберите подмножество данных для маркировки (10% от датасета).
- Разметьте выбранное подмножество вручную с помощью Label Studio.
- Сохраните размеченные вручную данные вместе с исходным датасетом.



30 минут

Перерыв



<<5:00->>



Задание 3

Разметка данных с использованием активного обучения

- Вам предоставлен датасет, содержащий рецензии на фильмы в файле 'movie.csv'. Ваша задача - построить модель анализа настроений, используя шаблон кода из лекции, и оценить ее производительность на тестовом наборе данных в файле.



50 минут



Домашнее задание

Задание 1.

Выберите датасет, который имеет отношение к вашей области интересов или исследований. Датасет должен содержать неструктурированные данные, требующие разметки для решения конкретной задачи, например, анализа настроений или распознавания именованных сущностей.

Задание 2.

Выполните разметку на основе правил (rule-based labeling) на подмножестве выбранного датасета. Разработайте и реализуйте набор правил или условий, которые позволят автоматически присваивать метки данным на основе определенных шаблонов или критериев.

Задача 3.

Выполните разметку вручную отдельного подмножества выбранного датасета с помощью выбранного вами инструмента разметки.



Домашнее задание

Задача 4.

Объедините данные, размеченные вручную, с данными, размеченными на основе правил.

Объедините два подмножества размеченных данных в один набор данных, сохранив при этом соответствующую структуру и целостность.

Задача 5.

Обучите модель машинного обучения, используя объединенный набор размеченных данных.

Разделите датасет на обучающий и тестовый наборы и используйте обучающий набор для обучения модели.

Задача 6.

Оценить эффективность обученной модели на тестовом датасете. Используйте подходящие метрики оценки. Интерпретируйте результаты и проанализируйте эффективность модели в решении задачи разметки.



Вопросы?

Вопросы?



Вопросы?





Спасибо за внимание!