Искусственный интеллект: разметка данных и построение моделей в каскаде

Онищук О.Н.

О лекторе

Онищук Олег Николаевич

- Руководитель направления аналитики данных,
  Правовой департамент Центрального Аппарата ПАО «Сбербанк»

- Магистр в сфере международного частного права, МГЮА, 2018

Магистерская диссертация: «Блокчейн как технология охраны авторских прав в трансграничных отношениях»

План лекции

Часть 2. Разметка данных

- Что такое разметка
- Инструменты разметки
- Влияние плохой разметки на результат

Часть 3. Построение иерархического NER

- Как модели искусственного интеллекта могут работать в каскаде.
- Применение каскада ML моделей в LegalTech

Разметка данных

Разметка. Основное и важное

- Разметка (аннотирование) данных - это добавление меток или тегов к обучающему набору данных для придания контексту и значению данных.

- Все виды данных, включая текст, изображения, аудио и видео, размечаются перед вводом в модель искусственного интеллекта.

- Размеченные данные помогают моделям машинного обучения изучать и распознавать закономерности, делать прогнозы или генерировать аналитические данные на основе помеченных данных.

- Качество и точность аннотаций к данным имеют решающее значение для производительности и надежности моделей машинного обучения.

Разметка. Классификация по типам задач

Компьютерное зрение («Computer Vision» или «CV»)

В CV аннотирование данных включает в себя маркировку и разметку визуальных элементов на изображениях, фотографиях и видео для обучения моделей по таким задачам, как:

- распознавание объектов,
- распознавание лиц,
- отслеживание движения,
- автономное вождение,
- и другим.

Разметка. Классификация по типам задач

Обработка данных в естественном языке («Natural language processing» или «NLP»):

В NLP разметка данных фокусируется на текстовой информации и элементах, связанных с языком.

Аннотация данных NLP предназначена для обучения моделей искусственного интеллекта понимать человеческую речь, понимать естественный язык и выполнять такие задачи, как:

- классификация текста,
- анализ настроений,
- распознавание именованных объектов (NER)
- машинный перевод.

Разметка данных

I. Задачи компьютерного зрения (CV)

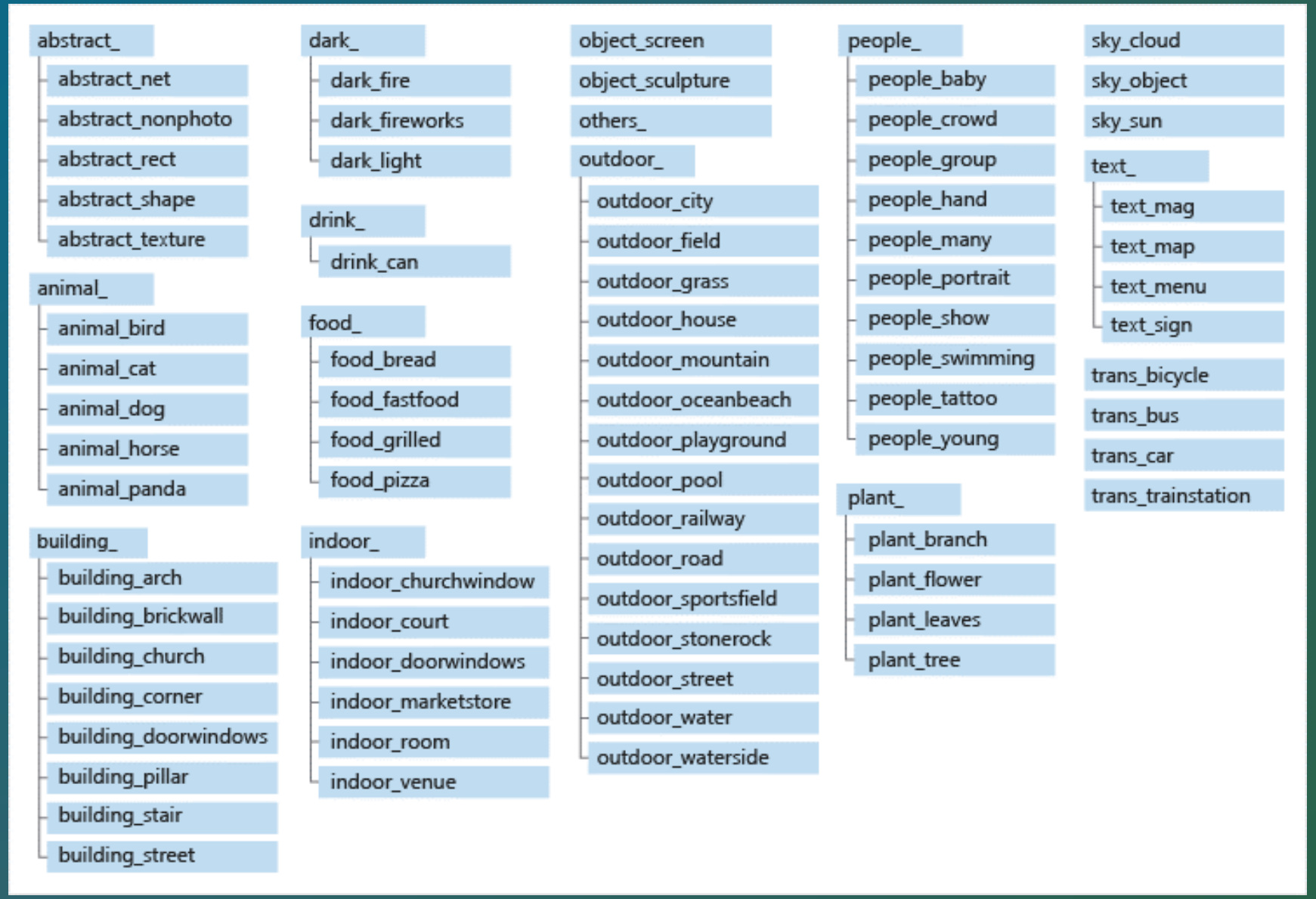
Разметка для CV

I. Классификация изображений (Image categorization)

- Решает главную задачу: «Что изображено на картинке?»
- Ответ: на изображении {объект}

В процессе создания аннотаций для данной задачи могут быть использованы различные методы, такие как аннотация в виде ограничивающего прямоугольника, семантическая сегментация и аннотация с ориентирами.

Пример таксономии категорий тэгов модели Image Analysis 3.2 (Microsoft Azure):



Результат отработки модели Image Analysis 3.2 (Microsoft Azure):

Предоставляем изображение CV модели:

Получаем на выходе:

JSON


```
{
    "categories": [
        {
            "name": "people_",
            "score": 0.81640625
        }
    ],
    "requestId": "bae7f76a-1cc7-4479-8d29-48a694974705",
    "metadata": {
        "height": 200,
        "width": 300,
        "format": "Jpeg"
    }
}
```

Разметка для CV

II. Распознавание / обнаружение объектов

Решает задачи: «Что, где и сколько изображено на картинке?»

Определение наличия, местоположения и количества объектов на изображении и нанесение на них меток называется распознаванием объектов.

Интересующие объекты на изображении могут быть помечены различными способами, включая ограничивающие рамки, полигоны, рамки с надписями и т.д.

Разметка для CV

III. Сегментация

Сегментация - это сложный метод разметки изображений, который включает в себя разделение изображения на разделы или сегменты и их разметку в соответствии с их визуальным содержанием.

Распространенные типы Сегментации в CV:

- Распознавание границ
- Семантическая сегментация
- Сегментация по экземплярам
- Паноптическая сегментация

Разметка для CV

Семантическая сегментация

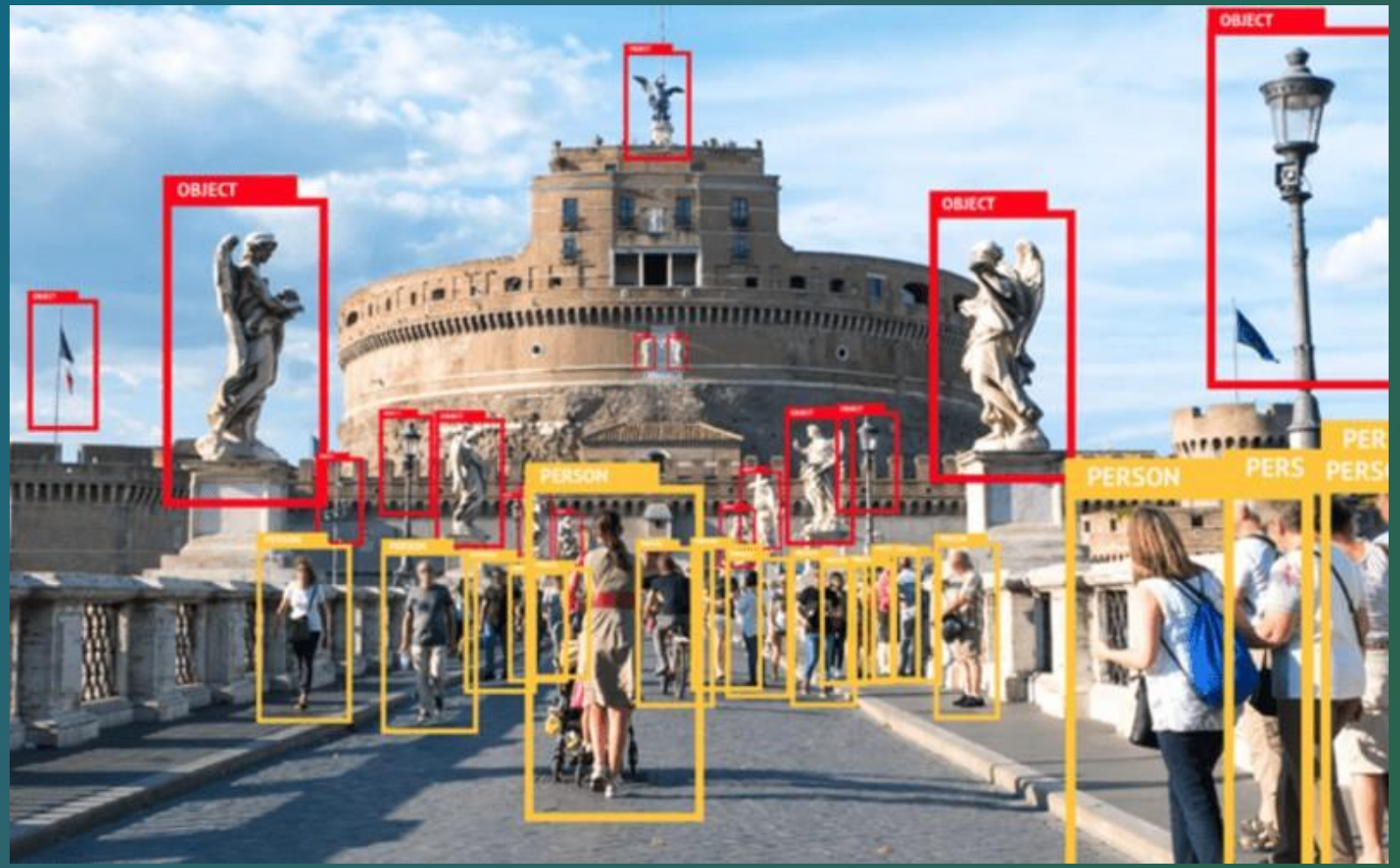
Задача: «Что это? Это тот же объект?»

Установление границ между связанными объектами на изображении и присвоение им одинакового идентификатора.

При данном методе компьютерного зрения каждому пикселю изображения присваивается метка или категория.

Разметка для CV

Семантическая сегментация



Разметка для CV

Сегментация по экземплярам

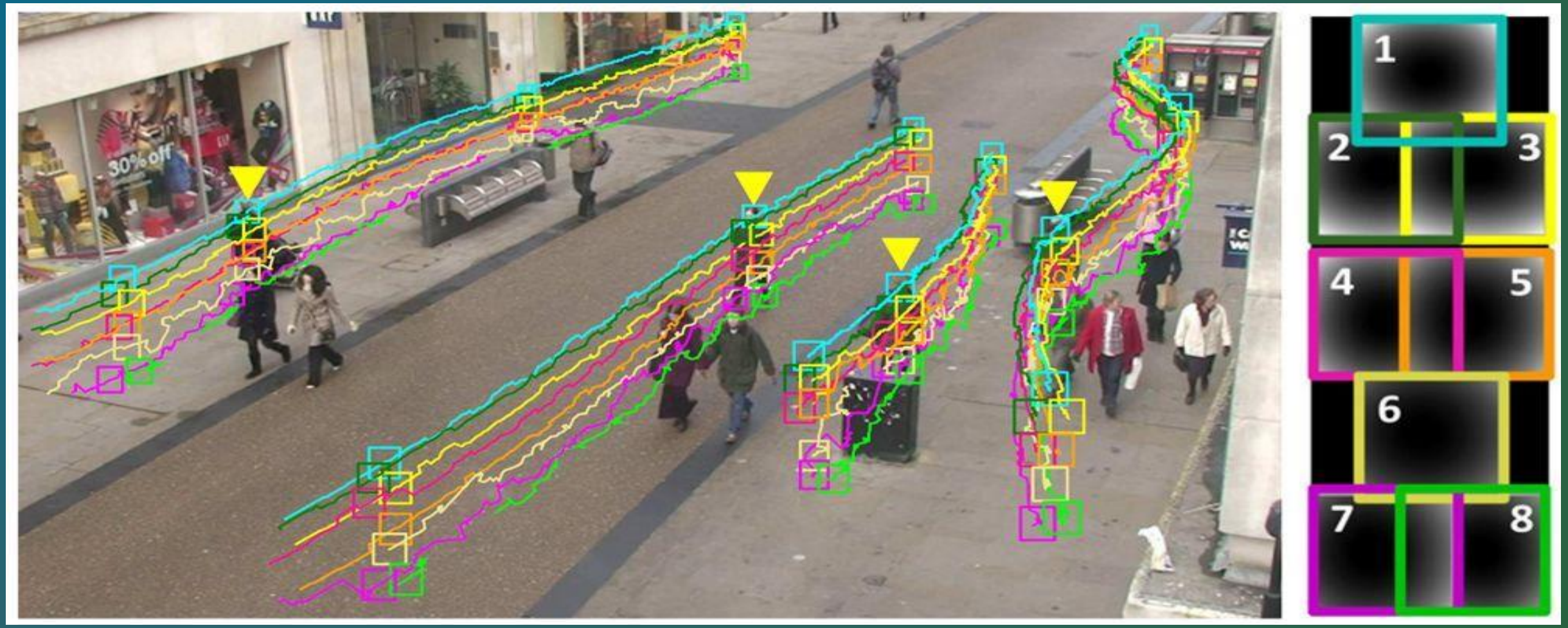
Задача: «Что это? Ага, это такой же объект. Чуть левее – что-то другое»

Сегментация по экземплярам - это более продвинутая версия семантической сегментации, которая позволяет различать различные экземпляры объектов на изображении.

В отличие от семантической сегментации, которая группирует все пиксели одного класса в одну категорию, сегментация по экземплярам присваивает уникальные метки каждому экземпляру объекта.

Разметка для CV

Сегментация по экземплярам



Разметка для CV

Распознавание границ

Задача: «Граница 1 ведет до… Граница 2 означает…»

Аннотации к изображениям для распознавания границ необходимы для обучения моделей машинного обучения распознавать узоры на изображениях без надписей.

Узоры могут быть простыми и сложными, иметь значения прямого действия или подразумеваемые, - например знак «STOP» или двойная сплошная линия разметки дорожного полотна.

Разметка для CV

Паноптическая сегментация

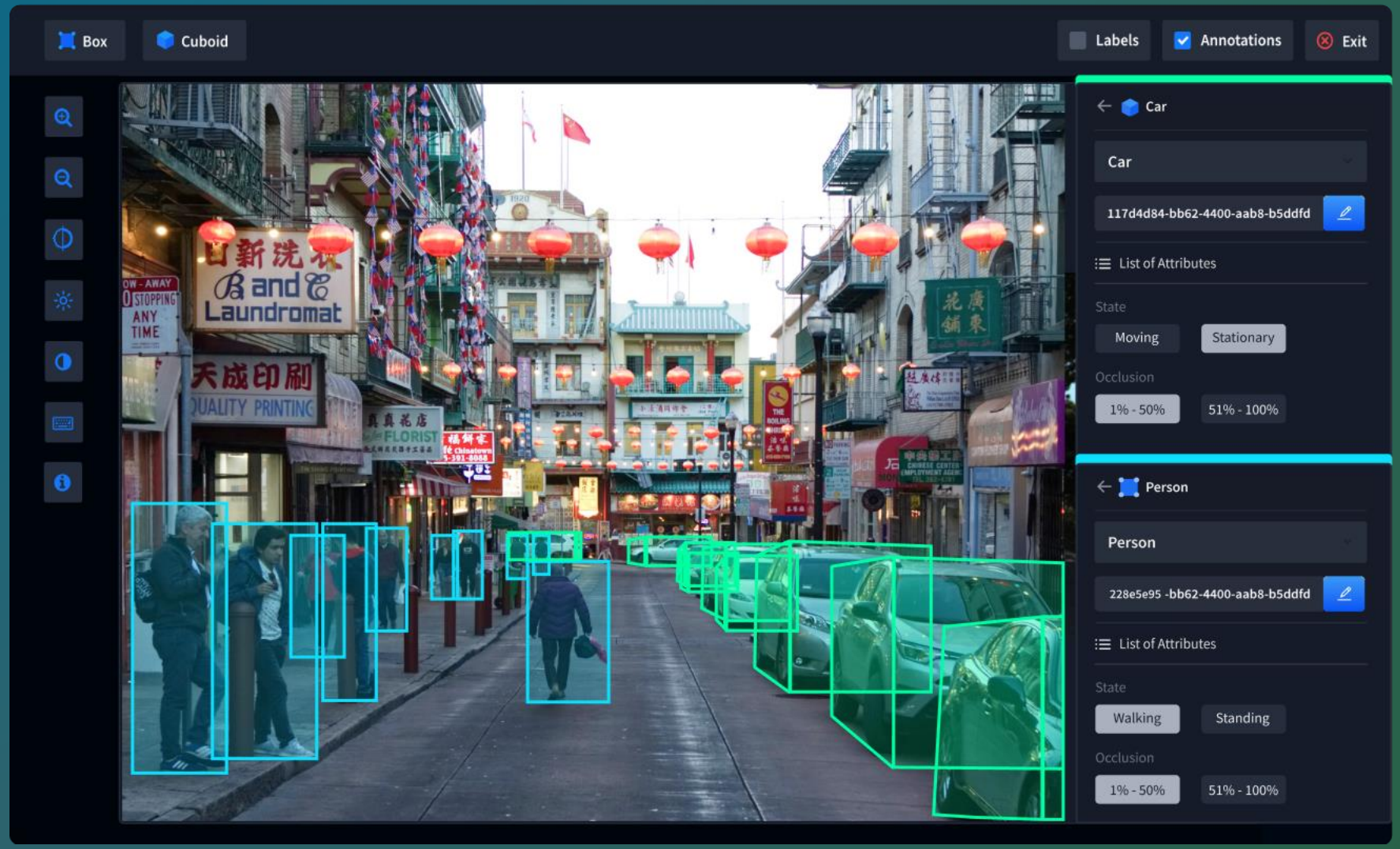
Задача: «На картинке обнаружено {O1, O2, N1, M1…} с точностью до пикселя»

Паноптическая сегментация - это задача компьютерного зрения, которая сочетает в себе семантическую сегментацию и сегментацию по экземплярам для получения полного представления об изображении.

Она направлена на разделение изображения на семантически значимые области и идентификацию каждого экземпляра объекта в них.

Разметка для CV

Разметка для паноптической сегментации



Разметка для CV

Примеры сегментации

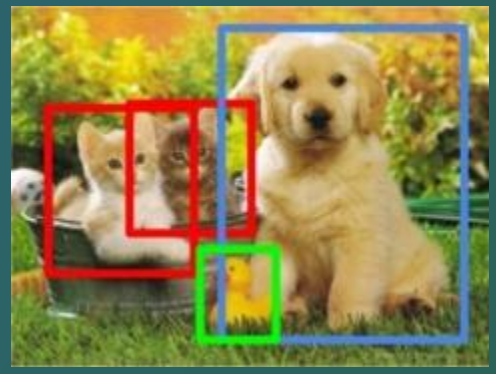
Классификация

Кошка

Классификация
+ локализация

Кошка

Обнаружение
объектов



Кошка + утка +
собака

Сегментация по
экземплярам

Кошка + утка +
собака

Единичный объект

Множество объектов

Разметка для CV

Популярные инструменты для разметки CV

- Encord Annotate
- Scale
- CVAT
- Labelbox
- Playment
- Appen
- Dataloop
- V7 Labs
- Hive
- Labelstudio
- Supervisely
- VoTT

* наименования представлены в качестве примера, носят информативный характер для образовательных целей и не являются рекламой

** все зарегистрированные товарные знаки принадлежат их правообладателям

Разметка данных

II. Задачи обработки естественного языка NLP

Типы разметки для NLP

В зависимости от задач:

- Именованные сущности (Entity annotation)
- Части речи (POS annotation)
- Синтаксические зависимости (Dependency annotation)
- Семантические роли (Semantic role annotation)
- Кореферентность (Coreference annotation)
- Связи (Named entity linking или «NEL»)

В зависимости от уровня:

- Токен (Token-level)
- Предложение (Sentence-level)
- Документ (Document-level)

Типы разметки для NLP

- Token-level annotation: разметка отдельных токенов (слов или символов) в тексте.
- Sentence-level annotation: аннотация отдельных предложений в тексте.
- Document-level annotation: аннотация целого документа или текста.

Типы разметки для NLP

Разметка именованных сущностей (Entity annotation): метод аннотации именованных сущностей, таких как имена, даты, места и организации.

Это позволяет NER и LLM извлекать соответствующую информацию из текстовых данных и использовать ее для принятия решений или классификации.

Типы разметки для NLP

Разметка части речи (Part-of-speech или POS annotation): метод аннотации частей речи, таких как глаголы, существительные, прилагательные и т.д.

Это позволяет системам NER и LLM понимать грамматическую структуру текста и извлекать соответствующую информацию.

Типы разметки для NLP

Разметка зависимостей (Dependency annotation): метод аннотации синтаксических зависимостей между словами в тексте.

Это позволяет системам NER и LLM понимать синтаксическую структуру текста и извлекать соответствующую информацию.

Типы разметки для NLP

Разметка семантики (Semantic role annotation): метод аннотации семантических ролей, таких как агент, пациент, тема документа, роль в документе и т.д.

Это позволяет системам NER и LLM понимать смысл текста и извлекать соответствующую информацию.

Типы разметки для NLP

Разметка кореферентности (Coreference annotation): метод аннотации кореферентности, то есть связей между словами или фразами, которые относятся к одному и тому же объекту или сущности.

Это позволяет системам NER и LLM понимать контекст текста и извлекать соответствующую информацию.

Если сократить – находим псевдонимы и объединяем с одним и тем же объектом, к которому другие сущности принадлежат (кореферентны).

Типы разметки для NLP

Связывание сущностей (Named entity linking): метод аннотации сущностей, который включает в себя поиск и привязку именованных сущностей в тексте к соответствующим записям в базе знаний или базе данных.

Разметка для NLP | NER

Популярные инструменты для разметки текста

- Labelbox
- Diffgram
- Prodigy
- Brat
- TagMe

* наименования представлены в качестве примера, носят информативный характер для образовательных целей и не являются рекламой

** все зарегистрированные товарные знаки принадлежат их правообладателям

Методы разметки (CV + NLP)

- Ориентир
- Ограничивающие рамки, иногда – «2D боксы»
- Многоугольная сегментация
- Полилинии
- Отслеживание
- Трехмерные ограничивающие рамки иногда – «Трехмерные кубоиды»
- Полигональная аннотация

Иерархический NER

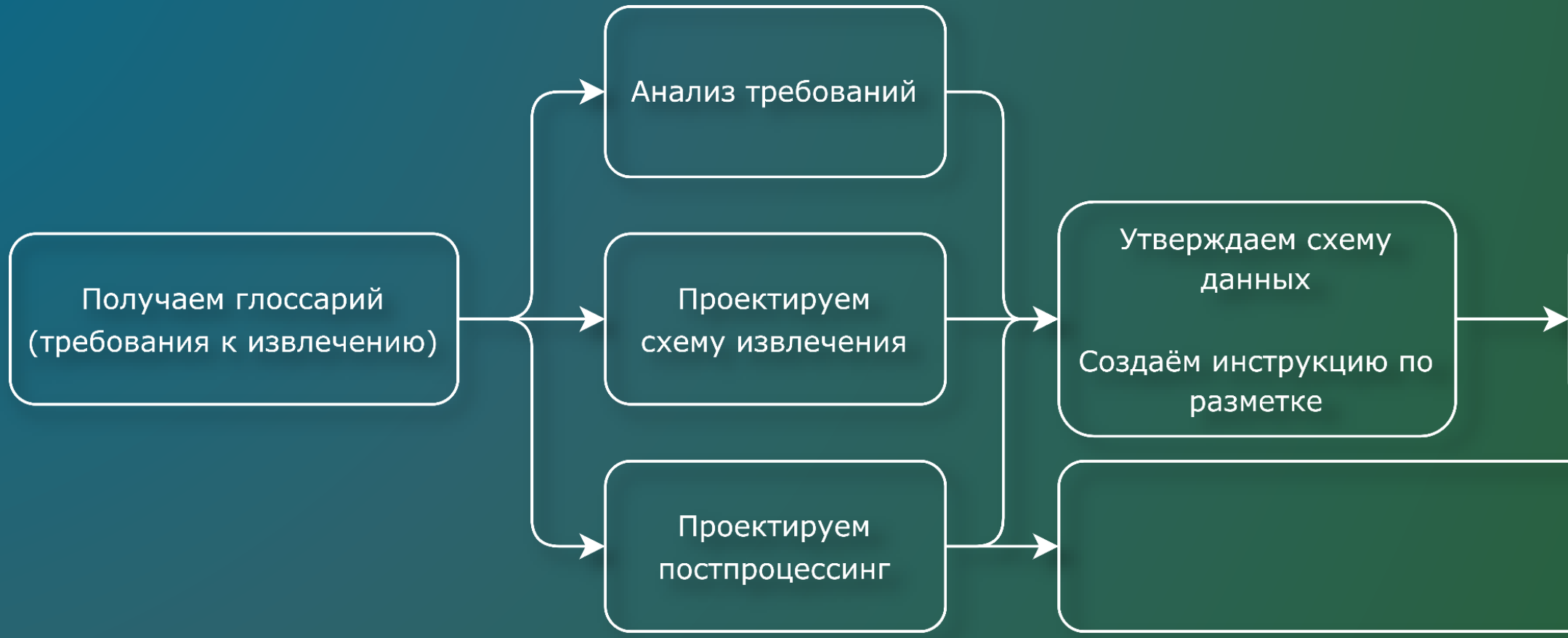
Часть 3: Построение иерархического NER

Построение иерархического NER

🚨 Требования! 🚨

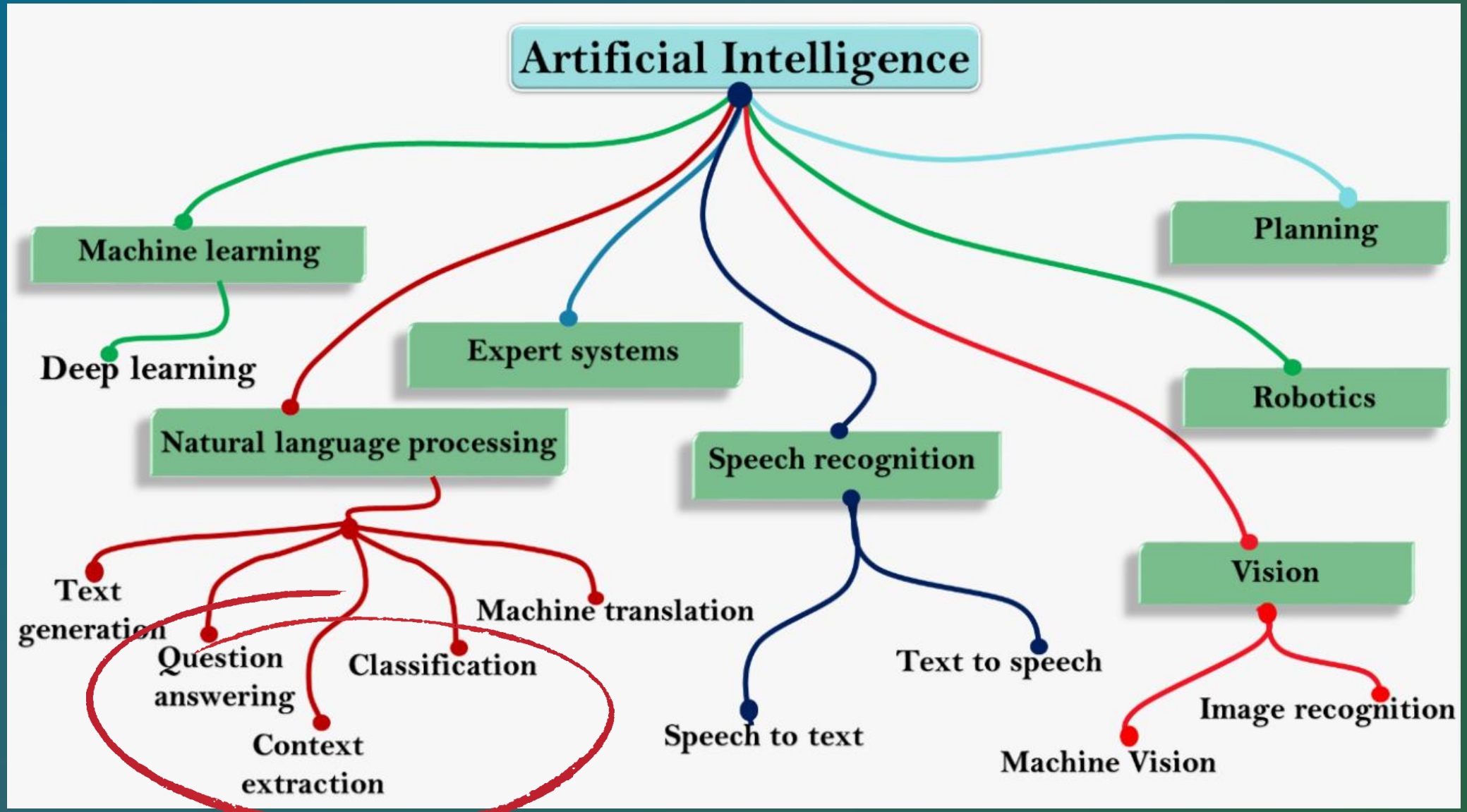
I. Понимать технические возможности вашего инструментария
II. Данные. Вычищенные, разнообразные, в большом количестве
III. Иметь чёткое и понятное ТЗ (и составлять такое же)
IV. Понимать, какие операции необходимы с документом (извлечение, нормализация, построение графов знаний и т.д.)
V. По возможности, - сегментируйте! (документ, операции с документом)
VI. Научный подход хорош, но это не повод отказываться от экспериментов. Пробуйте новые инструменты и подходы. Всегда.

Построение иерархического NER



Построение иерархического NER

Итоги


Artificial Intelligence
Machine learning
Deep learning
Expert systems
Natural language processing
Speech recognition
Planning
Robotics
Vision
Text generation
Question answering
Classification
Machine translation
Context extraction
Speech to text
Text to speech
Machine Vision
Image recognition

Ну и что-нибудь «на почитать»

- «Named Entity Recognition in Historic Legal Text: A Transformer and State Machine Ensemble Method» - ссылка
- «E-NER - An Annotated Named Entity Recognition Corpus of Legal Text» - ссылка
- «Named entity recognition in the Romanian legal domain» - ссылка
- «Legal Named Entity Recognition with Multi-Task Domain Adaptation» - ссылка
- «SemEval-2023 Task 6: LegalEval - Understanding Legal Texts» - ссылка

Спасибо за внимание!