Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Программная инженерия. Разработка ПО (Python для продвинутых специалистов. Машинное обучение)

Модуль: Предобработка данных и машинное обучение

Лекция 15: Основы обработки естественного языка (NLP)

Дата: 07.07.2025

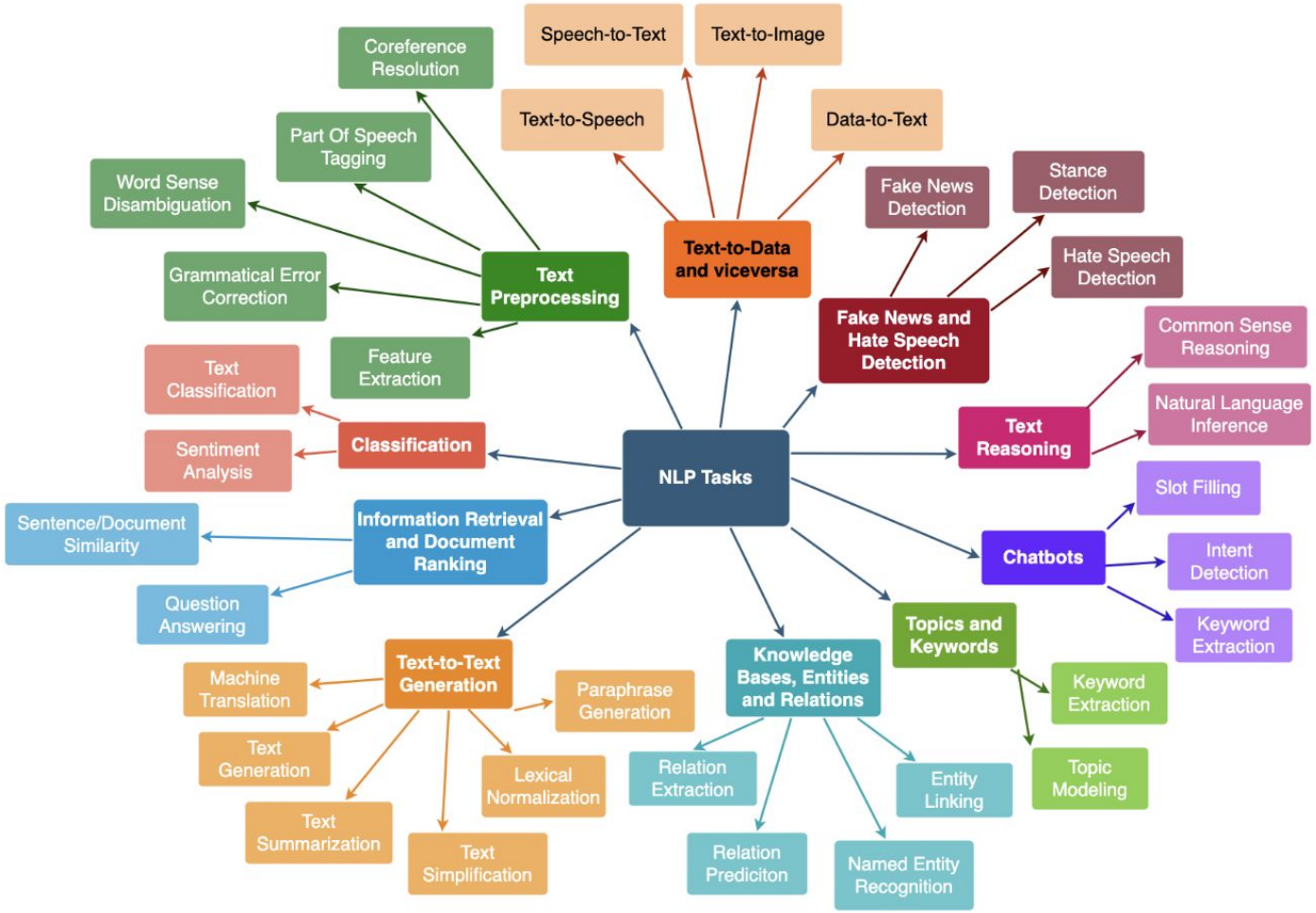
Что такое Natural Language Processing (NLP)

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

NLP — одно из направлений искусственного интеллекта, которое работает

- с анализом,
- пониманием и
- генерацией живых языков,

для того, чтобы взаимодействовов с компьютерами и устно, и письменно, используя естественные языки вместо компьютерных.



источник картинки - https://www.nlplanet.org/

Основные задачи работы с текстом NLP

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- **Токенизация (Tokenization):**

Разделение текста на токены (слова, предложения или другие более мелкие единицы), чтобы сделать текст более структурированным для последующей обработки.

- **Предобработка текста (Text Preprocessing):**

Очистка и преобразование текста перед анализом. Это может включать удаление стоп-слов, пунктуации, приведение к нижнему регистру, лемматизацию и стемминг.

- **Парсинг (Parsing):**

Анализ синтаксической структуры предложений для определения связей между словами и создания деревьев разбора.

Основные задачи работы с текстом NLP

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- **Частеречная разметка (Part-of-Speech Tagging):**
Присвоение частям речи (существительное, глагол, прилагательное и т.д.) для
каждого слова в предложении.

- **Извлечение информации (Information Extraction):**
Извлечение структурированных сущностей из текста, таких как имена, даты,
местоположения и т.д.

- **Извлечение фактов и связей (Relation Extraction):**
Определение связей и отношений между сущностями в тексте.

Основные задачи работы с текстом NLP

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- **Разрешение семантической неоднозначности (Word Sense Disambiguation):**
Определение правильного значения слова в контексте, когда у слова есть несколько возможных смыслов.

- **Обработка естественного языка с помощью глубокого обучения (Deep Learning NLP):**
Применение глубоких нейронных сетей для различных задач NLP, таких как машинный перевод, анализ тональности, вопросно-ответные системы и т.д.

- **Анализ эмоциональной окраски (Sentiment Analysis):**
Определение эмоциональной окраски текста (негативной, позитивной, нейтральной).

- **Генерация текста (Text Generation):**
Создание текстовых данных с помощью алгоритмов, которые могут автоматически составлять предложения и тексты.

Этапы работы с текстом NLP

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

➔Этап 1. Обработка текста (очистка и синтаксис)
- Токенизация (разделение исходного текста на токены).
- Поиск частей речи
- Лемматизация (приведение слов к нормальной словарной форме)
- Удаление «стоп слов»
- Тематическое моделирование
- Поиск устойчивых словосочетаний (n-gramm)

➔Этап 2. Векторизация текста (статистический анализ, глубокое обучение, семантический анализ)
- Кодирование данных помощью методологии TF-IDF
- Поиск близких по смыслу слов с помощью векторной модели word2vec, bert, обработка rnn сетями

➔Этап 3. Выбор модели классификации и обучение
(выбор алгоритма для работы)

Этапы работы с текстом NLP

ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Практика

Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Спасибо за внимание