# Машинное обучение и NLP - библиотеки, pipeline

NLP (Natural Language Processing, обработка естественного языка) — ээто область искусственного интеллекта, которая помогает компьютерам понмиать, интерпретировать и генерировать человеческий язык осмысленным способом.

## Основное определение

- **NLP** объединяет лингвистику, машинное обучение и информатику для того, чтобы научить компьютеры воспринимать и использовать естественный язык так, как это делают люди.data-light+1
- Основными задачами являются: автоматический перевод, распознавание и синтез речи, классификация текстов, генерация ответов на вопросы, фильтрация спама и автоисправление.skillfactory+2

## Как работает NLP

- Для решения задач NLP используют методы машинного обучения, в частности нейронные сети (RNN, CNN, Transformer), которые обучаются на больших наборах текстовых данных.trainingdata+2
- Система анализирует текст, определяет структуру, смысл, эмоции и контекст, после чего связывает это с целями пользователя или системой команд.lia+1

## Примеры применения

- Голосовые помощники (Siri, Алиса, Google Assistant).
- Автоматические переводчики.
- Чат-боты и системы поддержки клиентов.
- Поисковые системы и рекомендательные сервисы.mchost+1
- Машинный перевод(Google Translate, DeepL и другие системы позволяют мгновенно переводить тексты между сотнями языков)

## Отличия от классического ML:

В NLP мы работаем с неструктурированными текстовыми данными, которые требуют специальной предобработки и представления.

# Типовые задачи NLP

## Классификация текста

Определение категорий документов (спам/не спам, тематика, статьи, анализ настроений). Одна из самых распространенных задач в коммерческих применениях.

## Извлечение сущностей

Named Entity Recognition (NER) - поиск и классификация именованных сущностей в тексте: имена людей, организации, геолокации, даты.

## Машинный перевод

Автоматический перевод текста с одного языка на другой.

## Генерация текста

# Представление текста в ML

## Bag-of-Words

Самый простой подход: текст представляется как набор слов без учета порядка. Быстро, но теряет семантику и конткест.

## TF-IDF

Term Frequency-Inverse Document Frequency. Учитывает важность слова в документе относительно всей коллекции. Более информативен чем BoW.

## Word Embeddings

Векторные представления слов в многомерном пространстве. Word2Vec, GloVe захватывает семантические отношения между словами

# Основные библиотеки для NLP

## NLTK

Natural Language Toolkit - академическая библиотека с обширным набором инструментов. Отлично подходит для обучения и исследований, включается корпуса и готовые модели.

- Токенизация
- Частеречная заметка
- Обширная документация

## spaCy

Промышленная библиотека для высокопроизводительной обработки текста. Оптимизирована для продакшана-систем, поддерживает множество языков из коробки.

## Hugging Face Transformers

# Что такое ML Pipline?

Pipline в ML - это последовательность шагов обработки данных, которые автоматически применяются к входным данным для получения финального предсказания.

## 01 Предобработка данных

Очистка, нормализация, удаление шума.

## 02 Извлечение признаков

Преобразование данных в числовые представления(векторизация).

Важно подобрать метрики для оценки качества и эффективности.

## 03 Обучение модели

Применение алгоритма машинного обучения к подготовленным данным.

## 04 Предсказание

Получение результатов на новых, ранее не виденных данных.

# NLP Pipline для анализа тональности

## 01 Очистка данных

Удаление HTML-тегов, специальных символов, приведение к нижнему регистру, удаление стоп-слов

## 02 Токенизация

Разбиение текста в численные векторы с учетом важности слов в корпусе

## Векторизация (TF-IDF)

Преобразование текста в численные векторы с учетом важности слов в корпусе

## Классификация

Применение алгоритма (логистическая регрессия, SVM) для определения тональности.

Весь pipline можно реализовать в несколько строк кода с использованием scikit-learn.pipline - это обеспечивает консинтеность и воспроизводимость.

# Практический пример

Задача: анализ тональности отзывов

Создадим систему для определения положительных и отрицательных отзывов клиентов интернет - магазина.
Данные: CSV-файл с колонками "текст отзыва" и "оценка"(позитив/негатив)

Подходы:
Классический: sklearn Pipline с TF-DF + логистическая регрессия
Современный: Hugging Face pipline ("sentiment-analysis")

# Современные тренды

## Large Language Model

BERT, GPT, LLaMA революционизировали NLP. Трансформеры с attintion-механизмами показывают рекордные результаты в большинстве задач.

## Few-shot обучение

Современные модели могут решать новые задачи с минимальным количеством примеров или вовсе без дообучения (zero-shot).

## Промышленное применение

От чат-ботов до автоматизации документооборота - NLP трансформирует целые индустрии и бизнесс-процессы.