

Машинное обучение и NLP: библиотеки, pipeline

Фадеев Виталий Олегович

Зачем изучать NLP?

Поиск и информационные системы

Современные поисковые системы используют NLP для понимания запросов пользователей и ранжирования результатов по релевантности.

Чат-боты и виртуальные ассистенты

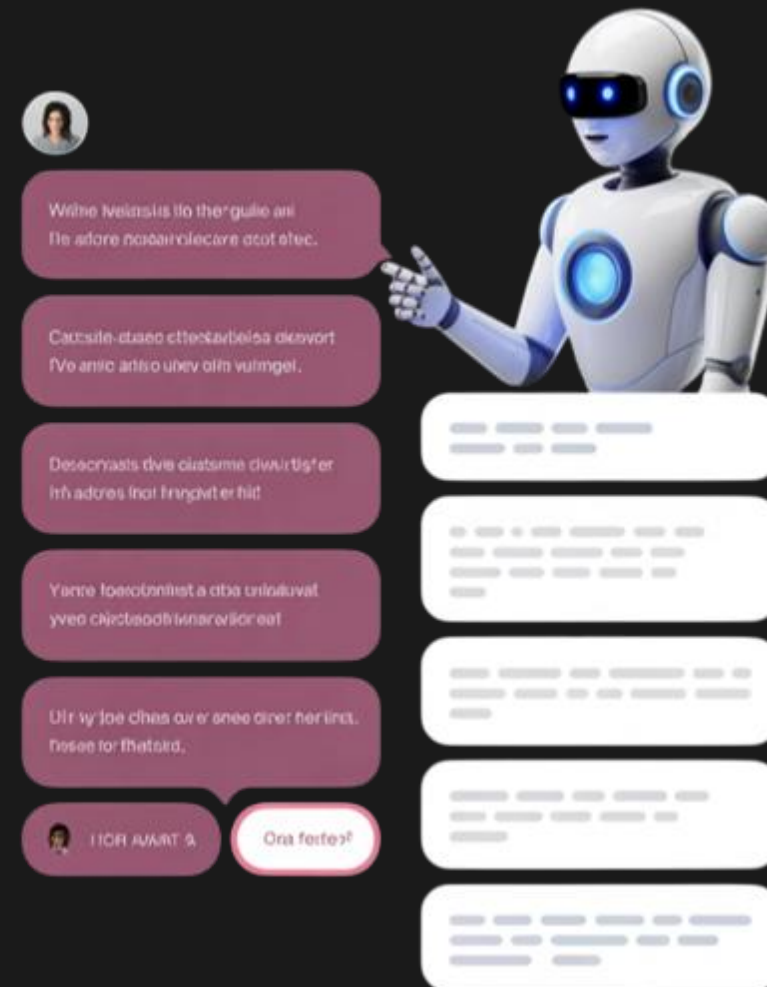
От Алисы до ChatGPT – NLP лежит в основе всех современных диалоговых систем и голосовых помощников.

Анализ тональности

Автоматический анализ настроений в социальных сетях, отзывах клиентов и мониторинг репутации брендов.

Машинный перевод

Google Translate, DeepL и другие системы позволяют мгновенно переводить тексты между сотнями языков.



План нашего занятия

01

Основы NLP и ML

Понимание ключевых концепций и отличий от классического машинного обучения

02

Инструменты и библиотеки

Понимание ключевых концепций и отличий от классического машинного обучения

03

Построение NLP pipeline

Практические подходы к созданию конвейеров обработки текстовых данных

04

Практический pipeline

Pipeline: создание системы анализа тональности отзывов

05

Современные тренды

Large Language Models и их влияние на развитие области

Что такое NLP?

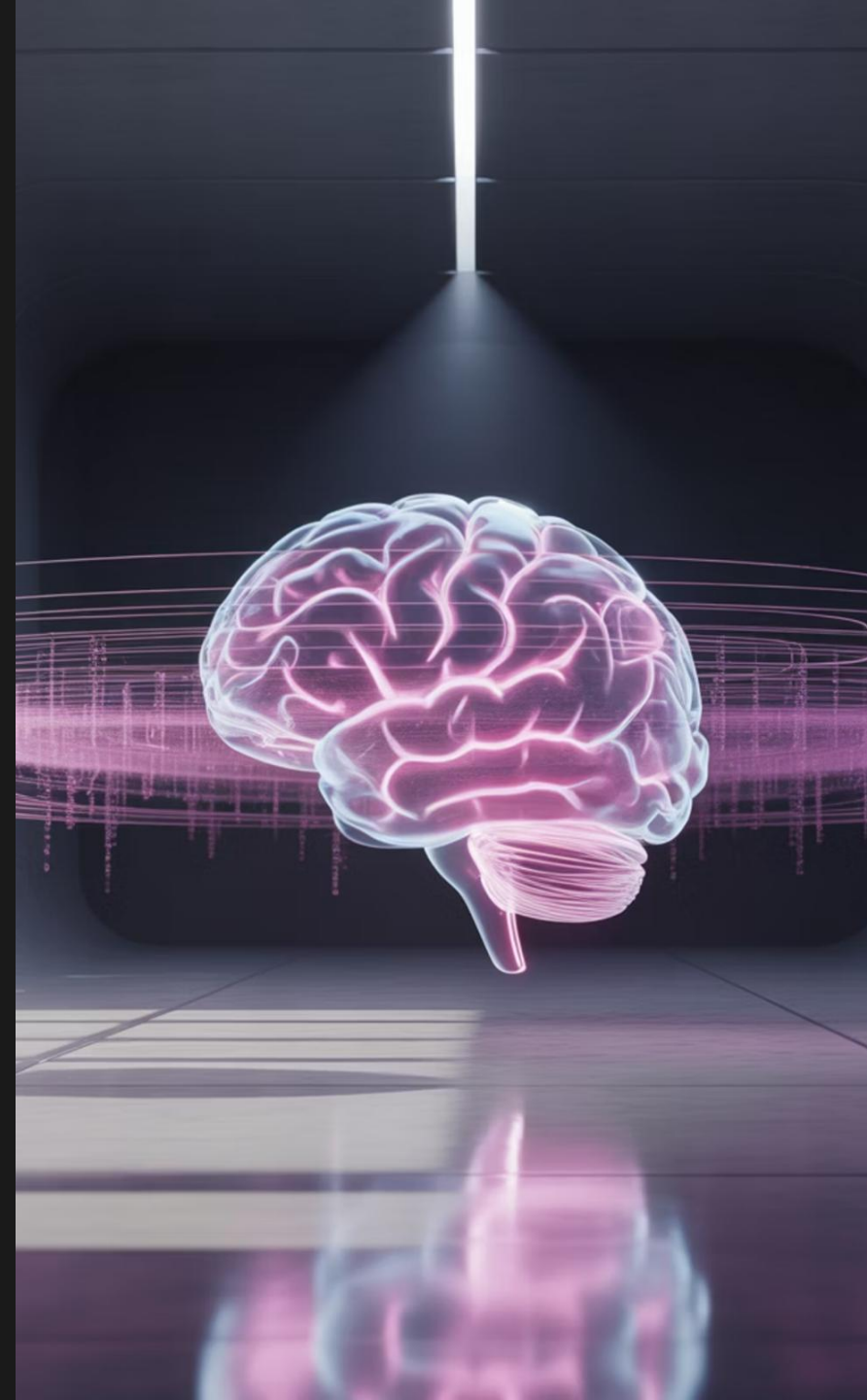
Natural Language Processing

NLP (обработка естественного языка) — это область искусственного интеллекта, которая помогает компьютерам понимать, интерпретировать и генерировать человеческий язык осмысленным способом.

Ключевое отличие от классического ML:

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

В отличие от традиционного машинного обучения, где данные уже структурированы, в NLP основная сложность заключается в преобразовании текста в числовые представления, которые может обработать алгоритм.



Машинное обучение и NLP: библиотеки, pipeline

Типовые задачи NLP

Классификация текста

Определение категории документа (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. Современные системы используют нейронные сети и attention-механизмы.

Генерация текста

Создание связного текста на основе контекста или промптов. От простых шаблонов до сложных языковых моделей как GPT.



Представление текста в ML



Language

1

Bag-of-Words

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

2

TF-IDF

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

3

Word Embeddings

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

Основные библиотеки для NLP



NLTK

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

- Токенизация и стемминг
- Частеречная разметка
- Обширная документация



spaCy

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

- Быстрая обработка больших объемов
- Предтренированные модели
- Простое API



Hugging Face Transformers

Современные трансформеры - доступ к SOTA моделям: BERT, GPT, RoBERTa. Простая интеграция предтренированных моделей и fine-tuning.

- Тысячи готовых моделей
- Поддержка PyTorch и TensorFlow
- Активное сообщество

Сравнение библиотек

Критерий	NLTK	spaCy	Transformers
Производительность	Низкая	Высокая	Средняя
Простота использования	Средняя	Высокая	Высокая
Готовые модели	Базовые	Отличные	SOTA
Размер библиотеки	Большой	Средний	Очень большой
Применение	Обучение, исследования	Продакшн	Исследования, продакшн

Выбор библиотеки зависит от конкретной задачи: для обучения лучше NLTK, для продакшн-систем - spaCy, для современных SOTA решений - Transformers.

Машинное обучение и NLP: библиотеки, pipeline

Что такое ML Pipeline?

Pipeline в машинном обучении - это **последовательность шагов обработки данных**, которые автоматически применяются к входным данным для получения финального предсказания.

01 Предобработка данных

Очистка, нормализация, удаление шума из исходных данных

02 Извлечение признаков

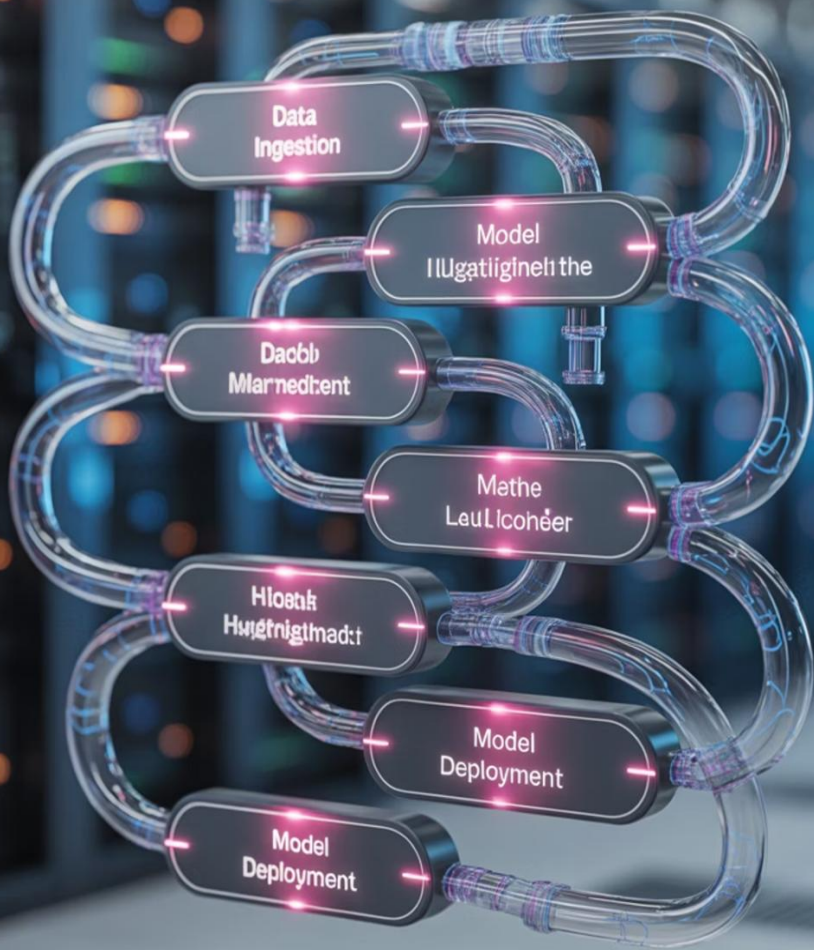
Преобразование данных в числовые представления (векторизация)

03 Обучение модели

Применение алгоритма машинного обучения к подготовленным данным

04 Предсказание

Получение результатов на новых, ранее не виденных данных



Машинное обучение и NLP: библиотеки, pipeline

NLP Pipeline для анализа тональности

01 Очистка текста

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

02 Токенизация

Разбиение текста на отдельные слова или токены, обработка пунктуации

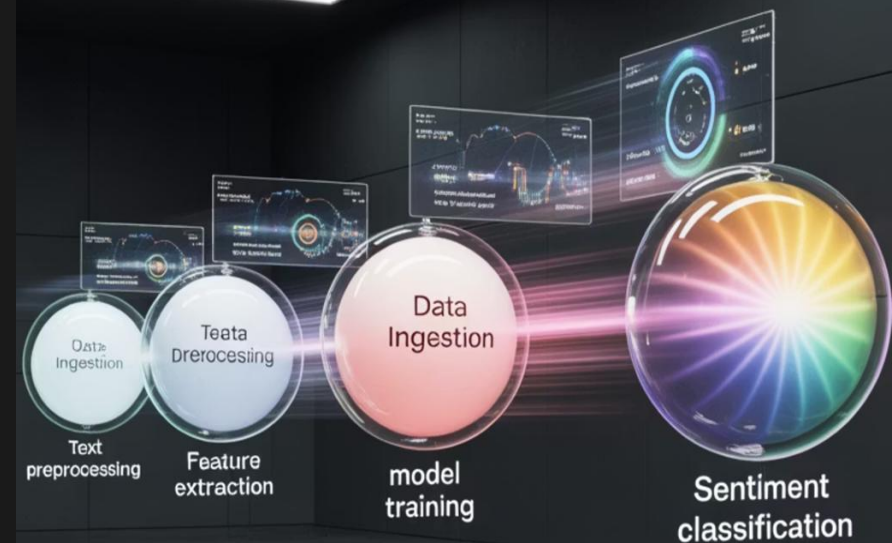
03 Векторизация (TF-IDF)

Преобразование текста в численные векторы с учетом важности слов в корпусе

04 Классификация

Применение алгоритма (логистическая регрессия, SVM) для определения тональности

Весь pipeline можно реализовать в несколько строк кода с использованием `scikit-learn.Pipeline` - это обеспечивает консистентность и воспроизводимость результатов.



Машинное обучение и NLP: библиотеки, pipeline

Практический пример

Задача: анализ тональности отзывов

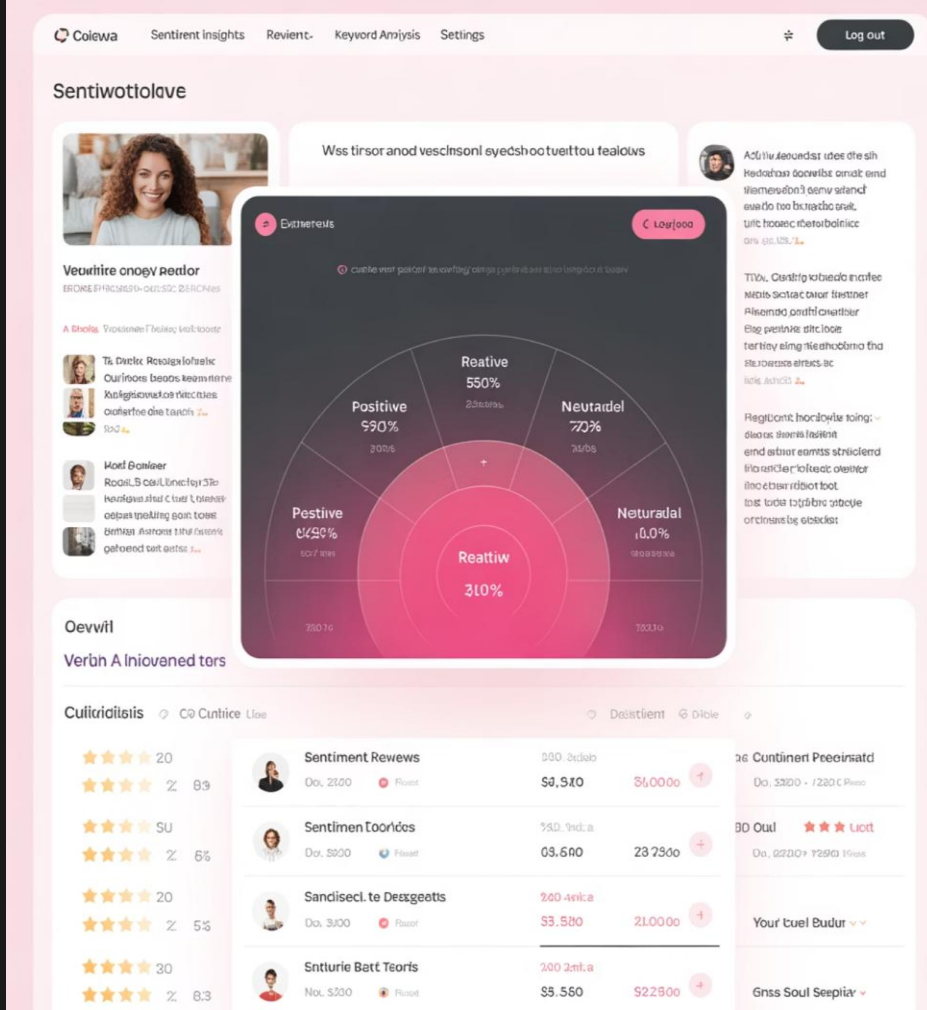
Создадим систему для определения положительных и отрицательных отзывов клиентов интернет-магазина.

Данные: CSV-файл с колонками "текст отзыва" и "оценка" (позитив/негатив)

Подходы:

- **Классический:** sklearn Pipeline с TF-IDF + логистическая регрессия
- **Современный:** Hugging Face pipeline("sentiment-analysis")

Сравнение: точность предсказаний, скорость работы, простота реализации



```
# Sklearn Pipelinepipe = Pipeline([ ('tfidf',  
TfidfVectorizer()), ('clf',  
LogisticRegression())])# Hugging  
Faceclassifier = pipeline( "sentiment-  
analysis", model="blanchefort/rubert-base-  
cased-sentiment")
```

Современные тренды

Large Language Models

BERT, GPT, LLaMA революционизировали NLP. Трансформеры с attention-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 бизнес-процессы.

Ключевые выводы

Библиотеки

NLTK для обучения

spaCy для продакшн

Transformers для SOTA решений

Pipeline:

структурированный подход от сырого текста к предсказанию

Баланс:

не всегда нужны сложные модели - иногда простой ML-pipeline эффективнее

Рекомендации для изучения: практикуйтесь на реальных данных, изучайте документацию библиотек, следите за новыми исследованиями в области NLP.