

Полная программа

# Natural Language Processing (NLP)

Длительность курса: 104 часа

## Модуль 1. Python для работы с текстами

### Тема 1

#### Введение в NLP

Цель занятия

рассмотреть область Natural Language Processing; рассмотреть задачи NLP; обсудить современные достижения в области NLP.

Краткое содержание

область NLP; задачи NLP; современные достижения NLP.

### Тема 2

#### Recap python/ data analysis/ визуализации

Цель занятия

вспомнить основы работы с Python для анализа данных и визуализации; поработать с библиотеками pandas, numpy, matplotlib, seaborn.

Краткое содержание

основные библиотеки Python для анализа данных и визуализации.

### Тема 3

#### Работа со строками + регулярные выражения

Цель занятия

рассмотреть методы работы со строками; освоить работу с регулярными выражениями.

Краткое содержание

работа со строками в Python; регулярные выражения; практика с регулярными выражениями.

### Тема 4

#### Парсинг данных

Цель занятия

рассмотреть открытые источники данных; использовать API парсинг; создать свои датасеты.

Краткое содержание

парсинг.

#### Домашние задания

##### Откуда берутся датасеты

Цель

В этом ДЗ вы напишите свой парсер, который будет бегать по страничкам и автоматически что-то собирать.

## Модуль 2. Введение в DL

### Тема 1

#### Введение в нейросети

Цель занятия

познакомиться с моделью искусственного нейрона; булевы операции над нейроном; разобрать как устроена нейронная сеть.

Краткое содержание

искусственный нейрон; полносвязная нейронная сеть.

### Тема 2

#### Градиентный спуск и backpropagation

Цель занятия

вспомнить, что такое градиентный спуск; повторить правило цепочки производных; рассмотреть алгоритм backpropagation; понять, как нейронная сеть обучается.

Краткое содержание

градиентный спуск; chain rule; граф вычислений; backpropagation.

### Тема 3

#### PyTorch. Часть 1

Цель занятия

познакомиться с фреймворком для глубокого обучения PyTorch и его архитектурой; разобраться, как устроены тензоры, какие операции над ними поддерживаются; изучить, как работать с данными в PyTorch, как использовать объекты dataset и dataloader; создать свою нейронную сеть, и обучить ее решать задачу классификации рукописных цифр на примере датасета MNIST.

Краткое содержание

pytorch; torchvision; тензоры; dataloader; dataset; MNIST.

#### Домашние задания

##### Практика по PyTorch

Цель

В этом ДЗ вы попрактикуетесь с PyTorch.

### Тема 4

#### PyTorch. Часть 2

Цель занятия

углубиться в PyTorch, и рассмотреть отличия версии 2x от 1.x; кратко пойтись по его модулям; проверить на примерах компиляцию моделей torch.compile(), и убедиться, действительно ли PyTorch 2.x работает быстрее 1.x; на примере MNIST / CIFAR рассмотреть другие архитектуры сетей (CNN); попробовать также Transfer Learning: MNIST -> CIFAR -> MNIST; обучить нейросеть задаче регрессии.

Краткое содержание

pytorch 2x; torch.compile(); Transfer Learning; CIFAR / MNIST.

### Тема 5

#### Рекуррентные сети. Часть 1

Цель занятия

разобрать, что такое рекуррентная нейронная сеть; познакомиться с видами рекуррентных сетей.

Краткое содержание

RNN; LSTM; GRU.

### Тема 6

#### Рекуррентные сети. Часть 2

Цель занятия

рассмотреть классификацию и генерацию текста рекуррентной сетью.

Краткое содержание

построение генеративной модели, классификационной модели, shallow-parsing модели средствами RNN.

## Модуль 3. Классические методы NLP

### Тема 1

#### Предобработка данных и понятие векторных представлений слов

Цель занятия

обсудить pipeline предобработки текстовых данных (удаление стоп-слов, токенизация, лемматизация); рассмотреть библиотеки для работы с текстом, такие как nltk и rpython2.

Краткое содержание

токенизация; стоп-слова; нормализация, векторные представления, BOW, tf-idf.

### Тема 2

#### Векторные представления слов и работа с предобученными эмбедингами

Цель занятия

рассмотреть алгоритмы word2vec, fasttext; научиться работать с векторными представлениями слов; использовать основные векторные представления для решения NLP-задач.

Краткое содержание

word2vec; fasttext; предобученные эмбединги.

### Тема 3

#### Задача NER

Цель занятия

рассмотреть задачу NER и области ее применения.

Краткое содержание

NER; подходы для решения.

### Тема 4

#### Языковые модели (n-граммные языковые модели)

Цель занятия

познакомиться с понятием языковых моделей.

Краткое содержание

понятие языковой модели в NLP; статические языковые модели; краткий экскурс по нейросетевому подходу.

### Тема 5

#### Тематическое моделирование

Цель занятия

обсудить тематическое моделирование; рассмотреть общую схему решения задач NLP.

Краткое содержание

тематическое моделирование, PLSA, LDA; практика на Python.

#### Домашние задания

##### Что в векторе твоем?

Цель

В этом ДЗ вы освоите работу с предобученными векторными представлениями.

## Модуль 4. Нейросетевые языковые модели

### Тема 1

#### Нейросетевые языковые модели и стратегии генерации текста

Цель занятия

рассмотреть концепцию нейросетевых языковых моделей и процесс их обучения с помощью cross-entropy loss; рассмотреть основные методы генерации текста: beam search, sampling, top-k sampling.

Краткое содержание

нейросетевые языковые модели; обучение нейросетевых моделей через cross-entropy; методы генерации текста: beam search, sampling, top-k sampling.

### Тема 2

#### Машинный перевод и seq2seq

Цель занятия

познакомиться с задачей seq2seq и энкодер-декодер архитектурой на примере машинного перевода.

Краткое содержание

задача seq2seq; энкодер-decoder подход для решения задачи машинного перевода; RNN для решения задачи машинного перевода.

### Тема 3

#### Архитектура Transformer и концепция attention mechanism

Цель занятия

рассмотреть архитектуру Transformer; познакомиться с идеей attention mechanism.

Краткое содержание

идея attention mechanism; архитектура Transformer.

### Тема 4

#### Transfer learning; BERT model

Цель занятия

познакомиться с идеей transfer learning и идеей pretraining+finetuning подхода для решения NLP задач; изучить модель BERT.

Краткое содержание

transfer learning; pretraining+finetuning подход; BERT; практика по дообучению предобученного BERTa.

### Тема 5

#### Практическое занятие: работа с предобученными языковыми моделями на практическом примере

Цель занятия

научиться на практике дообучать языковые модели для различных задач.

Краткое содержание

практика по дообучению трансформерных языковых моделей.

#### Домашние задания

##### Почувствуй мощь трансформеров в бою

Цель

Научиться работать с трансформерными моделями и применять их для различных NLP задач.

### Тема 6

#### Генеративные языковые модели GPT3 и методы few, zero-shot learning

Цель занятия

рассмотреть модель GPT3 и концепцию few-/zero-shot learning; концепция p-tuning.

Краткое содержание

GPT3; few-/zero-shot learning; практика на finetune GPT3 для генеративных задач и применение few-/zero-shot методов с (Ru)GPT3.

### Тема 7

#### Towards ChatGPT

Цель занятия

разобрать методы, лежащие в основе ChatGPT; разобрать подход обучения с подкреплением на основе отзывов (RLHF); разработать идею тонинга инструкций и модель InstructGPT.

Краткое содержание

подход обучения с подкреплением на основе отзывов (RLHF); тоннинг инструкций; модель InstructGPT.

## Модуль 5. Практические методы применения LLM и фундаментальных моделей

### Тема 1

#### Теория промптинга LLM

Цель занятия

познакомиться с подходами к промптингу больших языковых моделей; освоить методы промптинга инструктивных моделей.

Краткое содержание

методы промптинга Zero-shot, Few-shot, CoT, Self consistency, Knowledge generated prompting, Agency prompting, ReAct.

### Тема 2

#### Sentence-transformers

Цель занятия

познакомиться с библиотекой sentence-transformers.

Краткое содержание

библиотека sentence-transformers; архитектуры bi-encoder, cross-encoder; обучение архитектур bi-encoder, cross-encoder под собственные данные.

### Тема 3

#### Langchain

Цель занятия

познакомиться с библиотекой LangChain.

Краткое содержание

библиотека LangChain; сегментация текста; векторный поиск и векторные БД; реализация пайплайнов на langchain.

### Тема 4

#### RAG - генерация на основе базы знаний

Цель занятия

познакомиться с подходом RAG.

Краткое содержание

обзор пайплайна RAG; промптинг для RAG, задача поиска и ранжирования; подготовка данных для RAG.

## Модуль 6. Дополнительные главы NLP

### Тема 1

#### Оценка языковых моделей; классические NLP-бенчмарки

Цель занятия

обсудить основные англо-, русско- и мультиязычные бенчмарки и датасеты для оценки языковых моделей.

Краткое содержание

англоязычные датасеты и бенчмарки; русскоязычные датасеты и бенчмарки; мультиязычные датасеты и бенчмарки; практика по оценке трансформеров на бенчмарках.

### Тема 2

#### Вопросно-ответные системы (задача question-answering)

Цель занятия

познакомиться с задачей Question-Answering.

Краткое содержание

задача Question-Answering; подходы к решению задач QA.

### Тема 3

#### Распределенное обучение

Цель занятия

освоить методы распределенного обучения языковых моделей.

Краткое содержание

методы распределенного обучения.

### Тема 4

#### Создание телеграм-бота

Цель занятия

научиться создавать собственных телеграм-ботов.

Краткое содержание

практика, как создавать телеграм-бота.

#### Домашние задания

##### И пусть твой бот заговорит!

Цель

В этом ДЗ вы создадите телеграм-бота.

## Модуль 7. Проектный модуль

### Тема 1

#### Выбор темы и организация проектной работы

Цель занятия

выбрать и обсудить тему проектной работы; спланировать работу над проектом; ознакомиться с регламентом работы над проектом.

Краткое содержание

правила работы над проектом и специфика проведения итоговой защиты; требования к результату проекта и итоговой документации.

#### Домашние задания

##### Проектная работа

Цель

Выбрать и утвердить в чате по ДЗ темы проекта, разработать и презентовать проект.

### Тема 2

#### Презащита №1

Цель занятия

обсудить прогресс по проектной работе; решить сложности, возникшие при выполнении проектной работы.

Краткое содержание

обсуждение со студентами их продвижений по проектной работе.

### Тема 3

#### Презащита №2

Цель занятия

обсудить прогресс по проектной работе; решить сложности, возникшие при выполнении проектной работы.

Краткое содержание

обсуждение со студентами их продвижений по проектной работе.

### Тема 4

#### Защита проектных работ

Цель занятия

защитить проект; получить рекомендации экспертов.

Краткое содержание

презентация проектов перед комиссией; вопросы и комментарии по проектам.