Бауманский учебный центр «Специалист»

Программа курса Нейронные сети. Уровень 2. Компьютерная обработка естественного языка

Код курса: **ПИМЛ3**

Тема	Ак. часов
Модуль 1. Введение в обработку естественного языка • Основные задачи обработки естественного языка. Классификация предложений: определение эмоциональной окраски, обнаружение спама, проверка грамматики определение связи между предложениями. Классификация слов в предложении: синтаксический анализ, выделение именованных сущностей. Генерация текста: завершение предложений, заполнение пропусков. Вопросно-ответные системы. Генерация предложений. • Токенизация и векторное представление текста. One-hot encoding, мешок слов, TF, TF-IDF. Методы векторного представления текста, учитывающие отношения между словами: Word2Vec, GloVe, FastText. • Нейронные сети с архитектурой «Трансформер». Кодировщики, декодировщики, слои внимания.Архитектуры и веса. • Библиотека transformers. Стандартные конвейеры (пайплайны) в библиотеке transformers. • **Лабораторная работа:** пайплайны в transformers: sentiment-analysis, zero-shot-classification, translation, text-generation, fill-mask, ner, question-answering, summarization.	**4**
Модуль 2. Использование библиотеки transformers • Токенизация в библиотеке transformers: на основе слов, символов, подслов. Кодирование и декодирование слов. • Трансферное обучение. Предварительно обученные модели и головы моделей. Маски внимания. • Обработка нескольких последовательностей. Обработка длинных последовательностей. • Загрузчики данных. Оценка качества моделей. Платформа Hugging Face и библиотека Datasets. • **Лабораторная работа:** Настройка токенизатора DistilBERT. • **Лабораторная работа:** Дообучение модели bert-base-uncased для классификации пар предложение на наборе данных MRPC. • **Лабораторная работа:** Семантический поиск с помощью FAISS.	**8**
Модуль 3. Токенизаторы в библиотеке transformers • Дообучение токенизаторов. Быстрые токенизаторы. Нормализация и предварительная токенизация. Токенизация Byte-Pair Encoding, WordPiece, Unigram. • **Лабораторная работа:** Быстрые токенизаторы в вопросно-ответном конвейере. • **Лабораторная работа:** Создание токенизатора.	**8**
Модуль 4. Практическое применение моделей обработки естественного языка • Практическая реализация моделей классификации токенов, классификации текстов, фильтрации спама, перевода, суммаризации. • Перплексия языковых моделей. • Обучение с помощью Trainer и Accelerate. • **Лабораторная работа:** Дообучение модели маскированного языкового моделирования. • **Лабораторная работа:** Полное обучение каузальной языковой модели с нуля. • **Лабораторная работа:** Собственный проект по тематике профессиональных интересов слушателей.	**12**

Аудиторная нагрузка в классе с преподавателем	**32 ак.ч. + 32 ак.ч. самостоятельно**

По окончании обучения на курсе проводится итоговая аттестация. Аттестация проводится в виде теста на последнем занятии или на основании оценок практических работ, выполняемых во время обучения на курсе.

Контакты центра

Для частных лиц
+7 (495) 232-32-16

Для организаций
+7 (495) 780-48-44

E-mail
info@specialist.ru