Аннотация к рабочей программе дисциплины
Инжиниринг данных

Направление подготовки: 09.04.01 Информатика и вычислительная техника
Направленность (профиль*):* Инженерия искусственного интеллекта
Квалификация выпускника: магистр
Целью освоения дисциплины является: формирование у студентов теоретических и практических знаний о методах сбора, извлечения и обработки данных, методах построения систем хранения данных, обучении технологии функционирования платформ распределенной обработки больших наборов данных
Объем дисциплины: 3 зачетные единицы– 108 часов
Семестр: 1
Краткое содержание основных разделов дисциплины:

№ п/п раздела	Основные разделы дисциплины	Краткое содержание разделов дисциплины
1	**Работа с данными в Python**	Библиотеки для работы с данными в различных форматах в Python: файлы CSV, JSON, HTML. Работа с базами данных в Python. Работа с изображениями, видео и звуковыми файлами. Общедоступные платформы для хранения данных. Подход Data-Centric AI. Форматы хранения больших данных и работа с ними: Parquet, Avro. Графы знаний. Библиотеки для работы с данными в различных форматах в Python: файлы CSV, JSON, HTML. Работа с базами данных в Python. Работа с изображениями, видео и звуковыми файлами.
2	**Подготовка данных для систем машинного обучения**	Сбор данных и формирование набора данных для систем машинного обучения. Загрузка данных из интернет и социальных сетей. Методы очистки и подготовки данных. Очистка и подготовка данных на Python. Разметка данных. Сбор данных и формирование набора данных для систем машинного обучения. Загрузка данных из интернет и социальных сетей.
3	**Параллельная и распределенная обработка данных**	Архитектура центров обработки данных, кластеры для параллельных и распределенных вычислений. Экосистема для распределенного хранения и обработки больших объемов данных: Apache Hadoop, HDFS. Распределенная обработка данных в Apache Spark. Архитектура Apache Spark: Resilient Distributed Dataset (RDD), действия трансформации. Работа с данными с использованием Spark DataFrame. Источники данных для Spark DataFrame. Обработка данных в Spark DataFrame. Использование SQL в Spark DataFrame. Методы очистки и подготовки данных. Очистка и подготовка данных на Python. Разметка данных.

Форма промежуточной аттестации: зачет