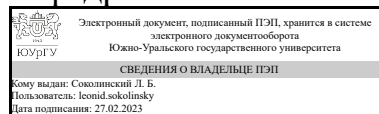


УТВЕРЖДАЮ:
Заведующий выпускающей
кафедрой



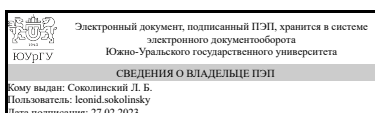
Л. Б. Соколинский

РАБОЧАЯ ПРОГРАММА

дисциплины 1.Ф.П0.03 Подготовка данных для машинного обучения
для направления 09.03.04 Программная инженерия
уровень Бакалавриат
профиль подготовки Инженерия информационных и интеллектуальных систем
форма обучения очная
кафедра-разработчик Системное программирование

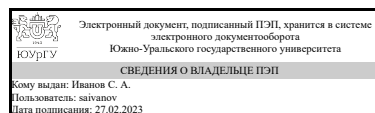
Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 09.03.04 Программная инженерия, утверждённым приказом Минобрнауки от 19.09.2017 № 920

Зав.кафедрой разработчика,
д.физ.-мат.н., проф.



Л. Б. Соколинский

Разработчик программы,
к.физ.-мат.н., доцент



С. А. Иванов

1. Цели и задачи дисциплины

Целью дисциплины является формирование базовых представлений, знаний и умений в области подготовки данных для машинного обучения. Основные задачи дисциплины: ознакомить студента с основными методами и подходами сбора и подготовки данных для машинного обучения, дать описание основных характеристик наборов данных, показать способы предварительной обработки данных.

Краткое содержание дисциплины

Изложение наиболее важных понятий, определений и методов работы с данными при подготовке датасетов для машинного обучения. В курс входят следующие разделы: математические основы, основы работы с изображениями и видео, основы работы с текстовыми данными, основы работы с аудио данными. Практические задания подразумевают работу по сбору и разметке данных в том числе при помощи современных средств разметки и с использованием методов машинного обучения для автоматической разметки.

2. Компетенции обучающегося, формируемые в результате освоения дисциплины

Планируемые результаты освоения ОП ВО (компетенции)	Планируемые результаты обучения по дисциплине
ПК-3 (ПК-1 модели) Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта	Знает: базовые подходы к сбору, разметке и предварительной подготовке данных для моделей машинного обучения Умеет: ПК-1.3. У-1. Умеет осуществлять сбор исходной информации с использованием платформ данных (облачных и внутрикорпоративных) Имеет практический опыт: создания собственных наборов данных для моделей машинного обучения при решении задач с учетом особенностей решаемой задачи
ПК-6 (ПК-7 модели) Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	Знает: ПК-7.1. 3-2. Знает уровни представления данных (ODS DDL, семантический слой, модель данных); ПК-7.1. 3-3 . Знает основные инструменты, библиотеки и технологии Data Science; ПК-7.2. 3-1. Знает методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки структурированных и неструктурированных данных; ПК-7.2. 3-2. Знает методы планирования вычислительного эксперимента, формирования обучающей и контрольной выборок; Умеет: ПК-7.2. У-1. Умеет выявлять и исключать из массива данных ошибочные данные и выбросы; ПК-7.1. У-1. Умеет отделять достоверные источники данных от сомнительных, осуществлять критических отбор данных, проверять их на целостность и

	<p>непротиворечивость; ПК-7.1. У-2. Умеет использовать инструменты и библиотеки для Data Science для поиска данных в открытых источниках, специализированных библиотеках и репозиториях; ПК-7.2. У-3. Умеет осуществлять разметку структурированных и неструктурированных данных; использовать инструменты библиотеки и технологии Data Science для подготовки и ПК-7.2. У-4 . Умеет использовать инструменты библиотеки и технологии Data Science для подготовки и разметки структурированных и неструктурированных данных для машинного обучения;</p> <p>Имеет практический опыт: использования инструментов и библиотек для Data Science для поиска данных в открытых источниках, специализированных библиотеках и репозиториях</p>
<p>ПК-9 (ПК-6 модели) Способен создавать и поддерживать системы искусственного интеллекта на основе нейросетевых моделей и методов</p>	<p>Знает: ПК-6.2. З-1. Знает принципы построения систем искусственного интеллекта на основе искусственного интеллекта на основе искусственных нейронных сетей, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта в том числе в условиях малого количества данных искусственных моделей;</p> <p>Умеет: применять известные алгоритмы предобработки данных для решения проблемы малой обучающей выборки</p> <p>Имеет практический опыт: применения подходов к предобработке малых наборов данных при построении систем искусственного интеллекта</p>
<p>ПК-11 (ПК-5 модели) Способен использовать инструментальные средства для решения задач машинного обучения</p>	<p>Знает: ПК-5.2. З-2. Знает принципы проведения машинного эксперимента, проблемы переобучения и недообучения модели, требования к обучающей, тестовой и валидационной выборкам для решения задач анализа данных и машинного обучения;</p> <p>Умеет: осуществлять оценку и отбор инструментальных средств для сбора и разметки наборов данных</p> <p>Имеет практический опыт: применения различных инструментальных средств для сбора и разметки наборов данных</p>

3. Место дисциплины в структуре ОП ВО

Перечень предшествующих дисциплин, видов работ учебного плана	Перечень последующих дисциплин, видов работ
<p>Технологии аналитической обработки информации, Структуры и алгоритмы обработки данных, Основы интеллектуального анализа данных, Глубокое обучение,</p>	<p>Введение в компьютерное зрение, Современные языки программирования систем искусственного интеллекта, Введение в обработку естественного языка</p>

Требования к «входным» знаниям, умениям, навыкам студента, необходимым при освоении данной дисциплины и приобретенным в результате освоения предшествующих дисциплин:

Дисциплина	Требования
Основы машинного обучения	<p>Знает: Знает классические методы и алгоритмы машинного обучения: предиктивные - обучение с учителем, дескриптивные - обучение без учителя, методы и инструментальные средства решения задач с использованием систем искусственного интеллекта в зависимости от особенностей проблемной области, критерии выбора методов и инструментальных средств решения интеллектуальных задач, подходы к выбору методов и инструментальных средств систем искусственного интеллекта, процесс, стадии и методологии разработки решений на основе искусственного интеллекта, возможности современных инструментальных средств и систем программирования для решения задач анализа данных и машинного обучения</p> <p>Умеет: проводить сравнительный анализ и осуществлять выбор, настройку при необходимости разработку методов и алгоритмов для решения задач машинного обучения, осуществлять оценку критериев выбора методов и инструментальных средств решения задач с помощью систем искусственного интеллекта и выбор методов и инструментальных средств в зависимости от особенностей проблемной и предметной областей</p> <p>Имеет практический опыт: применения методов машинного обучения для решения задач, использования инструментальных средств решения задач искусственного интеллекта</p>
Технологии аналитической обработки информации	<p>Знает: методы и критерии оценки качества моделей машинного обучения, общедоступные репозитории и специализированные библиотеки, содержащие наборы больших данных, постановку базовых задач интеллектуального анализа данных (поиск шаблонов, классификация, кластеризация) и базовые методы их решения</p> <p>Умеет: сопоставить задачам предметной области классы задач машинного обучения, настраивать и оптимизировать конфигурацию программного и аппаратного обеспечения с целью интеграции больших данных, планировать и выполнять машинные эксперименты, оценивать точность и качество построенных моделей</p> <p>Имеет практический опыт: анализа требований и идентификации классов задач для реализации приложений машинного обучения, разработки программных компонент</p>

	<p>для извлечения и подготовки больших данных для аналитической обработки информации, разработки приложений для аналитической обработки информации с помощью современных инструментальных средств</p>
<p>Основы интеллектуального анализа данных</p>	<p>Знает: общедоступные репозитории и специализированные библиотеки, содержащие наборы больших данных, постановку и методы решения основных задач интеллектуального анализа данных (поиск шаблонов, классификация, кластеризация), методы и критерии оценки качества моделей машинного обучения Умеет: настраивать и оптимизировать конфигурацию программного и аппаратного обеспечения с целью интеграции больших данных, планировать и выполнять машинные эксперименты, оценивать точность и качество построенных моделей, сопоставить задачам предметной области классы задач машинного обучения Имеет практический опыт: разработки программных компонент для извлечения и подготовки больших данных для интеллектуального анализа, разработки моделей машинного обучения для решения основных задач интеллектуального анализа данных (поиск шаблонов, классификация, кластеризация) и проведения вычислительных экспериментов по оценке точности и качества построенных моделей, анализа требований и определения необходимых классов задач для реализации приложений машинного обучения; определения метрик и критериев качества оценки моделей машинного обучения</p>
<p>Структуры и алгоритмы обработки данных</p>	<p>Знает: виды представления данных, методы поиска и парсинга данных., базовые структуры данных и основные алгоритмы их обработки Умеет: выбирать оптимальные алгоритмы для решения задач предметной области и осуществлять их программную реализацию Имеет практический опыт: применения наиболее распространенных алгоритмов для решения задач с использованием сложных структур данных</p>
<p>Глубокое обучение</p>	<p>Знает: базовые архитектуры и модели искусственных нейронных сетей, базовые подходы к поиску и подготовке данных для моделей искусственных нейронных сетей, основы работы и построения моделей искусственных нейронных сетей Умеет: осуществлять оценку и отбор моделей искусственных нейронных сетей для решения задач, выделять входные и выходные переменные с целью использования предиктивных моделей, применять современные инструментальные средства и системы программирования для разработки моделей</p>

	машинного обучения Имеет практический опыт: разработки систем искусственного интеллекта на основе моделей искусственных нейронных сетей, подготовки и разметки данных для моделей искусственных нейронных сетей, разработки моделей нейронных сетей для решения задач
Введение в искусственный интеллект	Знает: основные определения искусственного интеллекта и систем искусственного интеллекта, историю развития науки об искусственном интеллекте, эволюцию и главные тренды систем искусственного интеллекта; классы решаемых задач с помощью систем искусственного интеллекта; основные параметры идентификации задач искусственного интеллекта: назначение, сфера применения, виды используемых знаний, временные аспекты решения задач, международные и национальные стандарты в области искусственного интеллекта и смежных областях, основные принципы, тенденции развития и перспективы исследований и разработок в области искусственного интеллекта Умеет: определять принадлежность проблемной и предметной областей к классу решаемых задач с помощью систем искусственного интеллекта и основные параметры идентификации задач систем искусственного интеллекта, применять международные и национальные стандарты в области искусственного интеллекта и смежных областях для решения задач в профессиональной деятельности, применять технологии искусственного интеллекта при реализации проектов на различных стадиях жизненного цикла Имеет практический опыт:

4. Объём и виды учебной работы

Общая трудоёмкость дисциплины составляет 3 з.е., 108 ч., 54,25 ч. контактной работы

Вид учебной работы	Всего часов	Распределение по семестрам
		в часах
		Номер семестра
		6
Общая трудоёмкость дисциплины	108	108
<i>Аудиторные занятия:</i>	48	48
Лекции (Л)	16	16
Практические занятия, семинары и (или) другие виды аудиторных занятий (ПЗ)	32	32
Лабораторные работы (ЛР)	0	0
<i>Самостоятельная работа (СРС)</i>	53,75	53,75

Подготовка собственного набора данных изображений на платформа Robooflow.	20	20
Подготовка собственного набора текстовых данных.	33,75	33.75
Консультации и промежуточная аттестация	6,25	6,25
Вид контроля (зачет, диф.зачет, экзамен)	-	зачет

5. Содержание дисциплины

№ раздела	Наименование разделов дисциплины	Объем аудиторных занятий по видам в часах			
		Всего	Л	ПЗ	ЛР
1	Математические основы	6	2	4	0
2	Основы работы с изображениями и видео	14	4	10	0
3	Основы работы с текстовыми данными	20	8	12	0
4	Основы работы с аудио данными	8	2	6	0

5.1. Лекции

№ лекции	№ раздела	Наименование или краткое содержание лекционного занятия	Кол-во часов
1	1	Математические и статистические характеристики наборов данных. Обзор современных систем разметки данных.	2
2-3	2	Представление изображений и видео в компьютере и связь таких представлений с машинным обучением. Операции над данными в рамках предварительной подготовки для машинного обучения. Разметка изображений.	4
4-6	3	Представление текста в компьютере и связь таких представлений с машинным обучением. Операции над текстовыми данными в рамках предварительной подготовки для машинного обучения. Особенности сбора и обработки текстовых данных для машинного обучения.	6
7	3	Особенности аугментации текстовых данных. Разметка текстов.	2
8	4	Представление аудио в компьютере и связь таких представлений с машинным обучением. Операции над аудио данными в рамках предварительной подготовки для машинного обучения. Разметка аудио.	2

5.2. Практические занятия, семинары

№ занятия	№ раздела	Наименование или краткое содержание практического занятия, семинара	Кол-во часов
1-2	1	Статистические характеристики наборов данных. Работа с табличными данными, текстами и изображениями.	4
3-5	2	Представление изображений в компьютере. Рассмотрение современных библиотек работы с изображениями Pillow, OpenCV, Albumentation. Аугментация.	6
6-7	2	Представление видео данных. Кадрирование и аугментация данных.	4
8-10	3	Представление текстов в компьютере. Мешок слов, one-hot-encoding, векторное представление, embeddings. Методы работы с текстами sklearn и tensorflow.	6
11-13	3	Парсинг текстов. Подготовка текстовых данных для решения различных задач: классификации и распознавания именованных сущностей.	6
14-16	4	Представление аудио в компьютере. Рассмотрение современной библиотеки	6

		работы с аудио Librosa. Подготовка аудио данных с помощью методов tensorflow. Аугментация.	
--	--	--	--

5.3. Лабораторные работы

Не предусмотрены

5.4. Самостоятельная работа студента

Выполнение СРС			
Подвид СРС	Список литературы (с указанием разделов, глав, страниц) / ссылка на ресурс	Семестр	Кол-во часов
Подготовка собственного набора данных изображений на платформа Robooflow.	Методические указания раздел "Подготовка собственного набора данных изображений на платформе Robooflow." https://app.roboflow.com	6	20
Подготовка собственного набора текстовых данных.	Методические указания раздел "Подготовка собственного набора текстовых данных." https://appen.com	6	33,75

6. Фонд оценочных средств для проведения текущего контроля успеваемости, промежуточной аттестации

Контроль качества освоения образовательной программы осуществляется в соответствии с Положением о балльно-рейтинговой системе оценивания результатов учебной деятельности обучающихся.

6.1. Контрольные мероприятия (КМ)

№ КМ	Се-местр	Вид контроля	Название контрольного мероприятия	Вес	Макс. балл	Порядок начисления баллов	Учитывается в ПА
1	6	Текущий контроль	Тест 1	5	5	Компьютерный тест состоит из 5 вопросов, позволяющих оценить сформированность компетенций. На ответы отводится 10 мин. Стоимость одного вопроса - 1 балл. 5 баллов: задание полностью выполнено без ошибок 1-4 баллов: задание выполнено частично или выполнено с ошибками 0 баллов: задание не выполнено	зачет
2	6	Текущий контроль	Практическая 1	5	5	5 заданий, каждое задание 1 балл. 0 баллов: задание не выполнено	зачет
3	6	Текущий контроль	Практическая 2	5	5	5 заданий, каждое задание 1 балл. 0 баллов: задание не выполнено	зачет
4	6	Текущий контроль	Практическая 3	5	5	5 заданий, каждое задание 1 балл. 0 баллов: задание не выполнено	зачет
5	6	Текущий контроль	Практическая 4	5	5	5 заданий, каждое задание 1 балл. 0 баллов: задание не выполнено	зачет
6	6	Текущий контроль	Практическая 5	5	5	5 заданий, каждое задание 1 балл. 0 баллов: задание не выполнено	зачет

7	6	Текущий контроль	Практическая 6	5	5	5 заданий, каждое задание 1 балл. 0 баллов: задание не выполнено	зачет
8	6	Текущий контроль	Тест 2	5	5	Компьютерный тест состоит из 5 вопросов, позволяющих оценить сформированность компетенций. На ответы отводится 10 мин. Стоимость одного вопроса - 1 балл. 5 баллов: задание полностью выполнено без ошибок 1-4 баллов: задание выполнено частично или выполнено с ошибками 0 баллов: задание не выполнено	зачет
9	6	Промежуточная аттестация	зачет	-	15	Компьютерный тест состоит из 15 вопросов, позволяющих оценить сформированность компетенций. На ответы отводится 30 минут. Стоимость одного вопроса - 1 балл. 15 баллов: задание полностью выполнено без ошибок 1-14 баллов: задание выполнено частично или выполнено с ошибками 0 баллов: задание не выполнено	зачет

6.2. Процедура проведения, критерии оценивания

Вид промежуточной аттестации	Процедура проведения	Критерии оценивания
зачет	<p>При оценивании результатов учебной деятельности обучающегося по дисциплине используется балльно-рейтинговая система оценивания результатов учебной деятельности обучающихся (Положение о БРС утверждено приказом ректора от 24.05.2019 г. № 179, в редакции приказа ректора от 10.03.2022 г. № 25-13/09). Оценка за дисциплину формируется на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля. Зачтено: Величина рейтинга обучающегося по дисциплине 60...100 %. Незачтено: Величина рейтинга обучающегося по дисциплине 0...59 %. Если студент не согласен с оценкой, полученной по результатам текущего контроля, студент проходит мероприятие промежуточной аттестации в виде тестирования. Тестирование проводится в системе edu.susu.ru. Тест содержит 15 вопросов. На выполнение теста дается 30 минут. В этом случае оценка за дисциплину рассчитывается на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля и промежуточной аттестации. Фиксация результатов учебной деятельности по дисциплине проводится в день зачета при личном присутствии студента.</p>	В соответствии с пп. 2.5, 2.6 Положения

6.3. Паспорт фонда оценочных средств

Компетенции	Результаты обучения	№ КМ								
		1	2	3	4	5	6	7	8	9
ПК-3	Знает: базовые подходы к сбору, разметке и предварительной подготовке данных для моделей машинного обучения	+		+	+	+	+	+	+	+
ПК-3	Умеет: ПК-1.3. У-1. Умеет осуществлять сбор исходной информации с			+	+	+	+	+	+	+

	использованием платформ данных (облачных и внутрикорпоративных)									
ПК-3	Имеет практический опыт: создания собственных наборов данных для моделей машинного обучения при решении задач с учетом особенностей решаемой задачи									
ПК-6	Знает: ПК-7.1. 3-2. Знает уровни представления данных (ODS DDL, семантический слой, модель данных); ПК-7.1. 3-3 . Знает основные инструменты, библиотеки и технологии Data Science; ПК-7.2. 3-1. Знает методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки структурированных и неструктурированных данных; ПК-7.2. 3-2. Знает методы планирования вычислительного эксперимента, формирования обучающей и контрольной выборок;									
ПК-6	Умеет: ПК-7.2. У-1. Умеет выявлять и исключать из массива данных ошибочные данные и выбросы; ПК-7.1. У-1. Умеет отделять достоверные источники данных от сомнительных, осуществлять критический отбор данных, проверять их на целостность и непротиворечивость; ПК-7.1. У-2. Умеет использовать инструменты и библиотеки для Data Science для поиска данных в открытых источниках, специализированных библиотеках и репозиториях; ПК-7.2. У-3. Умеет осуществлять разметку структурированных и неструктурированных данных; использовать инструменты библиотеки и технологии Data Science для подготовки и ПК-7.2. У-4 . Умеет использовать инструменты библиотеки и технологии Data Science для подготовки и разметки структурированных и неструктурированных данных для машинного обучения;									
ПК-6	Имеет практический опыт: использования инструментов и библиотек для Data Science для поиска данных в открытых источниках, специализированных библиотеках и репозиториях									
ПК-9	Знает: ПК-6.2. 3-1. Знает принципы построения систем искусственного интеллекта на основе искусственного интеллекта на основе искусственных нейронных сетей, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта в том числе в условиях малого количества данных искусственных моделей;									
ПК-9	Умеет: применять известные алгоритмы предобработки данных для решения проблемы малой обучающей выборки									
ПК-9	Имеет практический опыт: применения подходов к предобработке малых наборов данных при построении систем искусственного интеллекта									
ПК-11	Знает: ПК-5.2. 3-2. Знает принципы проведения машинного эксперимента, проблемы переобучения и недообучения модели, требования к обучающей, тестовой и валидационной выборкам для решения задач анализа данных и машинного обучения;									
ПК-11	Умеет: осуществлять оценку и отбор инструментальных средств для сбора и разметки наборов данных									
ПК-11	Имеет практический опыт: применения различных инструментальных средств для сбора и разметки наборов данных									

Типовые контрольные задания по каждому мероприятию находятся в приложениях.

7. Учебно-методическое и информационное обеспечение дисциплины

Печатная учебно-методическая документация

а) основная литература:

Не предусмотрена

б) *дополнительная литература:*

Не предусмотрена

в) *отечественные и зарубежные журналы по дисциплине, имеющиеся в библиотеке:*

Не предусмотрены

г) *методические указания для студентов по освоению дисциплины:*

1. Методическое пособие

из них: учебно-методическое обеспечение самостоятельной работы студента:

1. Методическое пособие

Электронная учебно-методическая документация

№	Вид литературы	Наименование ресурса в электронной форме	Библиографическое описание
1	Основная литература	Электронно-библиотечная система издательства Лань	Бизли, Д. Python. Книга рецептов / Д. Бизли, Б. К. Джонс ; перевод с английского Б. В. Уварова. — Москва : ДМК Пресс, 2019. — 646 с. — ISBN 978-5-97060-751-0. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/131723
2	Дополнительная литература	Электронно-библиотечная система издательства Лань	Паттерсон, Д. Глубокое обучение с точки зрения практика / Д. Паттерсон, А. Гибсон. — Москва : ДМК Пресс, 2018. — 418 с. — ISBN 978-5-97060-481-6. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/116122
3	Основная литература	Электронно-библиотечная система издательства Лань	Ганегедара, Т. Обработка естественного языка с TensorFlow : руководство / Т. Ганегедара ; перевод с английского В. С. Яценкова. — Москва : ДМК Пресс, 2020. — 382 с. — ISBN 978-5-97060-756-5. https://e.lanbook.com/book/140584
4	Дополнительная литература	Электронно-библиотечная система издательства Лань	Годин, А. М. Статистика : учебник / А. М. Годин. — 13-е изд. — Москва : Дашков и К, 2021. — 412 с. https://e.lanbook.com/book/229796
5	Основная литература	Электронно-библиотечная система издательства Лань	Антонио, Д. Библиотека Keras – инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow / Д. Антонио, П. Суджит ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2018. — 294 с. — ISBN 978-5-97060-573-8. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/111438

Перечень используемого программного обеспечения:

Нет

Перечень используемых профессиональных баз данных и информационных справочных систем:

Нет

8. Материально-техническое обеспечение дисциплины

Вид занятий	№ ауд.	Основное оборудование, стенды, макеты, компьютерная техника, предустановленное программное обеспечение, используемое для различных видов занятий
Лекции	110 (3г)	Проектор, компьютерный класс
Практические занятия и семинары	110 (3г)	Проектор, компьютерный класс, Платформа обработки данных на основе искусственного интеллекта https://aiplatform.ru/ ML Space: российская платформа для ML-разработки полного цикла https://sbercloud.ru/ru/aicloud/mlspace
Зачет, диф.зачет	110 (3г)	Компьютерный класс