ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ:
Директор института
Институт естественных и точных наук

Электронный документ, подписанный ПЭП, хранится в системе электронного документооборота Южно-Уральского государственного университета
ЮУрГУ
СВЕДЕНИЯ О ВЛАДЕЛЬЦЕ ПЭП
Кому выдан: Замышляева А. А.
Пользователь: zamyshliaevaaa
Дата подписания: 24.12.2021

А. А. Замышляева

РАБОЧАЯ ПРОГРАММА

дисциплины 1.Ф.М1.05 Интеллектуальный анализ текстов
для направления 01.04.02 Прикладная математика и информатика
уровень Магистратура
магистерская программа Технологии и методы искусственного интеллекта в фундаментальных и прикладных исследованиях
форма обучения очная
кафедра-разработчик Прикладная математика и программирование

Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 01.04.02 Прикладная математика и информатика, утверждённым приказом Минобрнауки от 10.01.2018 № 13

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

Электронный документ, подписанный ПЭП, хранится в системе электронного документооборота Южно-Уральского государственного университета
ЮУрГУ
СВЕДЕНИЯ О ВЛАДЕЛЬЦЕ ПЭП
Кому выдан: Замышляева А. А.
Пользователь: zamyshliaevaaa
Дата подписания: 24.12.2021

А. А. Замышляева

Разработчик программы,
старший преподаватель (-)

Электронный документ, подписанный ПЭП, хранится в системе электронного документооборота Южно-Уральского государственного университета
ЮУрГУ
СВЕДЕНИЯ О ВЛАДЕЛЬЦЕ ПЭП
Кому выдан: Ножкина Т. Г.
Пользователь: nozhkinatg
Дата подписания: 24.12.2021

Т. Г. Ножкина

СОГЛАСОВАНО

Руководитель образовательной программы
д.физ.-мат.н., проф.

Электронный документ, подписанный ПЭП, хранится в системе электронного документооборота Южно-Уральского государственного университета
ЮУрГУ
СВЕДЕНИЯ О ВЛАДЕЛЬЦЕ ПЭП
Кому выдан: Замышляева А. А.
Пользователь: zamyshliaevaaa
Дата подписания: 24.12.2021

А. А. Замышляева

Челябинск

1. Цели и задачи дисциплины

Цель дисциплины: изучить фундаментальные основы дисциплины обработки естественного языка и научиться применять машинное обучение и нейронные сети для решения задач обработки естественного языка. Задачи дисциплины: изучить математические основы представления текстовых данных, методы обработки текстов, методы классификации и кластеризации текстов, реализацию алгоритмов обработки и анализа текстов с помощью различных библиотек, методы обработки текстов с помощью глубоких нейронных сетей.

Краткое содержание дисциплины

В процессе обучения изучаются математические основы представления текстовых данных, методы обработки текстовой информации, методы анализа, классификации и кластеризации текстов. Рассматривается реализация алгоритмов обработки и анализа текстов с помощью различных современных специализированных библиотек для языка программирования Python, и методы обработки текстов с помощью глубоких нейронных сетей.

2. Компетенции обучающегося, формируемые в результате освоения дисциплины

Планируемые результаты освоения ОП ВО (компетенции)	Индикаторы достижения компетенции:	Планируемые результаты обучения по дисциплине
ПК-9 Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых субтехнологий искусственного интеллекта в прикладных областях	ПК-9.2. Руководит проектами в области сквозной цифровой субтехнологии «Обработка естественного языка»	Знает: принципы построения систем обработки естественного языка, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка» Умеет: руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой технологии «Обработка естественного языка»
ПК-12 Способен разрабатывать и применять алгоритмы анализа данных для решения прикладных задач	ПК-12.1. Разрабатывает и применяет алгоритмы анализа данных при решении профессиональных задач	Умеет: применять различные методы и алгоритмы предобработки текстов и разрабатывать алгоритмы анализа полученных данных Имеет практический опыт: классификации и тематического моделирования текстов на основе интеллектуального анализа

3. Место дисциплины в структуре ОП ВО

Перечень предшествующих дисциплин, видов работ учебного плана	Перечень последующих дисциплин, видов работ
Компьютерное зрение, Информационный поиск, анализ и предобработка данных	Не предусмотрены

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

Дисциплина	Требования
Информационный поиск, анализ и предобработка данных	Знает: основные принципы сбора, хранения и предобработки данных Умеет: выбирать методы и средства для анализа данных, оценивать возможности и ограничения используемых методов, осуществлять дискретизацию непрерывных данных с учётом решаемой задачи Имеет практический опыт: сбора первичной информации, организации и хранения данных для конкретного исследования, применения методов предобработки данных
Компьютерное зрение	Знает: основные виды нейронных сетей, применяющихся для анализа изображений, их эффективные конфигурации и методики обучения Умеет: применять алгоритмы компьютерного зрения для распознавания образов, очистки изображений и других прикладных задач Имеет практический опыт: применения методов, позволяющих производить детектирование, отслеживание и классификацию объектов на изображениях и в видеопотоке

4. Объём и виды учебной работы

Общая трудоемкость дисциплины составляет 3 з.е., 108 ч., 44,5 ч. контактной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		4
Общая трудоёмкость дисциплины	108	108
Аудиторные занятия:	36	36
Лекции (Л)	12	12
Практические занятия, семинары и (или) другие виды аудиторных занятий (ПЗ)	0	0
Лабораторные работы (ЛР)	24	24
Самостоятельная работа (СРС)	63,5	63,5

с применением дистанционных образовательных технологий	0	
Подготовка отчётов по лабораторным работам	20	20
Подготовка к экзамену	23,5	23.5
Самостоятельное изучение принципов работы трансформеров	20	20
Консультации и промежуточная аттестация	8,5	8,5
Вид контроля (зачет, диф.зачет, экзамен)	-	экзамен

5. Содержание дисциплины

№ раздела	Наименование разделов дисциплины	Объем аудиторных занятий по видам в часах			
		Всего	Л	ПЗ	ЛР
1	Задачи обработки естественного языка	6	2	0	4
2	Предобработка текстов	6	2	0	4
3	Статистические модели языка	12	4	0	8
4	Нейросетевые языковые модели	6	2	0	4
5	Классификация текстов. Перевод текстов	6	2	0	4
6	Основные принципы работы трансформеров	0	0	0	0

5.1. Лекции

№ лекции	№ раздела	Наименование или краткое содержание лекционного занятия	Кол-во часов
1	1	Естественный язык и текст. Особенности обработки естественных языков. Лингвистический анализ. Извлечение признаков. Прикладные задачи обработки текста и итоги.	2
2	2	Токенизация по предложениям. Токенизация по словам. Лемматизация и стемминг текста. Стоп-слова. Регулярные выражения. Мешок слов. Векторная модель текста и TF-IDF. Создаём нейросеть для работы с текстом. Теоретические задачи: Векторная модель текста. Классификация новостных текстов. Базовые нейросетевые методы работы с текстами	2
3	3	Базовые нейросетевые методы работы с текстами. Общий алгоритм работы с текстами с помощью нейросетей. Дистрибутивная семантика и векторные представления слов.. Рецепты еды и Word2Vec на PyTorch. Теоретические вопросы: Дистрибутивная семантика. Основные виды нейросетевых моделей для обработки текстов. Свёрточные нейросети для обработки текстов. POS-тэггинг свёрточными нейросетям. Свёрточные нейросети в обработке и анализе текстовых данных	2
4	3	Языковые модели и генерация текста. Рекуррентные нейросети. Моделирование языка. Генерация имён и лозунгов с помощью RNN. Агрегация, механизм внимания. Трансформер и self-attention. Моделирование языка с помощью Transformer	2
5	4	Преобразование последовательностей: 1-к-1 и N-к-M. Распознавание плоской структуры коротких текстов. Распознавание структуры рецептов. Аспектный сентимент-анализ как NER. Преобразование последовательностей (seq2seq). Transfer learning, адаптация моделей. Контекстуализированные представления и перенос знаний.	2
6	5	Глубокие нейронные сети в обработке естественного языка. Сверточные нейронные сети. Классификация и кластеризация текстов. Библиотека	2

		Pytorch.	

5.2. Практические занятия, семинары

Не предусмотрены

5.3. Лабораторные работы

№ занятия	№ раздела	Наименование или краткое содержание лабораторной работы	Кол-во часов
1-2	1	Введение в векторные представления текстовых данных	4
3-4	2	Реализация алгоритма word2vec, doc2vec. Классические алгоритмы	4
5-6	3	Основы нейронных сетей RNN, GRU. Имплементация сетей на фреймворке Pytorch	4
7-8	3	Основы нейронных сетей LSTM, bi-LSTM. Имплементация сетей на фреймворке Pytorch	4
9-10	4	Нейронный машинный перевод с последовательностью, вниманием и подсловами	4
11-12	5	Глубокие нейронные сети для анализа текстовых данных. Self-supervised обучение (самообучение) и fine-tuning (дообучение под конкретную задачу) с помощью моделей трансформеров	4

5.4. Самостоятельная работа студента

Выполнение СРС			
Подвид СРС	Список литературы (с указанием разделов, глав, страниц) / ссылка на ресурс	Семестр	Кол-во часов
Подготовка отчётов по лабораторным работам	ЭУМД. осн. лит. п. 1. стр. 6-100,, п. 4. стр. 6-207.	4	20
Подготовка к экзамену	ЭУМД. осн. лит. п. 1., стр. 6-100, п. 4. стр. 6-207.	4	23,5
Самостоятельное изучение принципов работы трансформеров	ЭУМД. осн. лит. п. 4, доп. лит. п. 5.	4	20

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	Учи-тыва-ется в ПА
1	4	Текущий контроль	КМ-1. Лабораторная	10	7	За каждый верно выполненный этап работы начисляется 1 балл.	экзамен

			работа 1			За этапы, выполненные не верно или не выполненные баллы не начисляются.	
2	4	Текущий контроль	КМ-2. Лабораторная работа 2	10	9	За каждый верно выполненный этап работы начисляется 1 балл. За этапы, выполненные не верно или не выполненные баллы не начисляются.	экзамен
3	4	Текущий контроль	КМ-3. Лабораторная работа 3	10	13	За каждый верно выполненный этап работы начисляется 1 балл. За этапы, выполненные не верно или не выполненные баллы не начисляются.	экзамен
4	4	Текущий контроль	КМ-4. Лабораторная работа 4	10	7	За каждый верно выполненный этап работы начисляется 1 балл. За этапы, выполненные не верно или не выполненные баллы не начисляются.	экзамен
5	4	Текущий контроль	КМ-5. Лабораторная работа 5	10	5	За каждый верно выполненный этап работы начисляется 1 балл. За этапы, выполненные не верно или не выполненные баллы не начисляются.	экзамен
6	4	Текущий контроль	КМ-6. Активная познавательная деятельность	40	36	На каждом из 18 занятий студент может получить 2 балла: Студент задает вопросы по изучаемому материалу - 1 балл; Студент правильно отвечает на вопросы по изучаемому материалу - 1 балл. В противном случае баллы не начисляются.	экзамен
7	4	Текущий контроль	КМ-7. Доклад	10	5	Подготовлен доклад - 1 балл; Подготовлена презентация - 1 балл; Оформление презентации соответствует ГОСТ - 1 балл; Тема раскрыта - 1 балл; Доклад вызвал интерес у аудитории - 1 балл.	экзамен
8	4	Проме-жуточная аттестация	КМ-8. Экзамен	1	6	Вопрос 1. 2 балла - студент дал полный верный ответ на вопрос; 1 балл - студент дал не полный ответ на вопрос; 0 баллов - студент не ответил на вопрос или ответ был не верный. Вопрос 2. 2 балла - студент дал полный верный ответ на вопрос; 1 балл - студент дал не полный ответ на вопрос; 0 баллов - студент не ответил на вопрос или ответ был не верный. Вопрос 3. 2 балла - студент дал полный верный ответ на вопрос; 1 балл - студент дал не полный ответ на вопрос; 0 баллов - студент не ответил на вопрос или ответ был не верный.	экзамен

6.2. Процедура проведения, критерии оценивания

Вид промежуточной аттестации	Процедура проведения	Критерии оценивания
экзамен	На экзамене происходит оценивание учебной деятельности обучающихся по дисциплине на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля. Студент может улучшить свой рейтинг, пройдя контрольное мероприятие промежуточной аттестации, которое не является обязательным. Контрольное мероприятие промежуточной аттестации проводится во время экзамена в виде устного опроса. Студенту выдаётся экзаменационный билет, содержащий 3 вопроса из разных тем курса. Студенту дается 60 минут на подготовку ответов. Затем студент озвучивает свои ответы.	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
ПК-9	Знает: принципы построения систем обработки естественного языка, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой технологии «Обработка естественного языка»			+	+	+	+	+	+
ПК-9	Умеет: руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой технологии «Обработка естественного языка»				+	+	+		+
ПК-12	Умеет: применять различные методы и алгоритмы предобработки текстов и разрабатывать алгоритмы анализа полученных данных	+	+	+	+	+	+		+
ПК-12	Имеет практический опыт: классификации и тематического моделирования текстов на основе интеллектуального анализа		+	+	+	+	+		

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. ГОСТ Оформления отчёта

из них: учебно-методическое обеспечение самостоятельной работы студента:

1. ГОСТ Оформления отчёта

Электронная учебно-методическая документация

№	Вид литературы	Наименование ресурса в электронной форме	Библиографическое описание
1	Основная литература	Электронно-библиотечная система издательства Лань	Беляева, Л. Н. Сетевые лингвистические технологии : монография / Л. Н. Беляева, О. Н. Камшилова, К. Р. Пиотровская. — Санкт-Петербург : РГПУ им. А. И. Герцена, 2019. — 111 с. — ISBN 978–5–8064–2701–5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/136751 (дата обращения: 22.09.2021). — Режим доступа: для авториз. пользователей.
2	Дополнительная литература	Электронно-библиотечная система издательства Лань	Риз, Р. Обработка естественного языка на Java : учебное пособие / Р. Риз ; перевод с английского А. В. Снастина. — Москва : ДМК Пресс, 2016. — 264 с. — ISBN 978-5-97060-331-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/93272 (дата обращения: 22.09.2021). — Режим доступа: для авториз. пользователей.
3	Дополнительная литература	Электронно-библиотечная система издательства Лань	Нишит, П. Искусственный интеллект для .NET: речь, язык и поиск. Конструирование умных приложений с использованием Microsoft Cognitive Services APIs / П. Нишит ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2018. — 298 с. — ISBN 978-5-97060-605-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/112929 (дата обращения: 23.09.2021). — Режим доступа: для авториз. пользователей.
4	Основная литература	Электронно-библиотечная система издательства Лань	Онтологии и тезаурусы: модели, инструменты, приложения : учебное пособие / Б. В. Добров, В. В. Иванов, Н. В. Лукашевич, В. Д. Соловьев. — 2-е изд. — Москва : ИНТУИТ, 2016. — 207 с. — ISBN 978-5-9963-0007-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/100277 (дата обращения: 23.09.2021). — Режим доступа: для авториз. пользователей.
5	Дополнительная 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. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140584 (дата обращения: 23.10.2021). — Режим доступа: для авториз. пользователей.

Перечень используемого программного обеспечения:

1. Microsoft-Office(бессрочно)

2. -Java SE SDK (комплект для разработки на Java SE)(бессрочно)
3. -Python(бессрочно)
4. -Microsoft Visual Studio (бессрочно)
5. Microsoft-Microsoft Imagine Premium (Windows Client, Windows Server, Visual Studio Professional, Visual Studio Premium, Windows Embedded, Visio, Project, OneNote, SQL Server, BizTalk Server, SharePoint Server)(04.08.2019)

Перечень используемых профессиональных баз данных и информационных справочных систем:

1. -База данных polpred (обзор СМИ)(бессрочно)

8. Материально-техническое обеспечение дисциплины

Вид занятий	№ ауд.	Основное оборудование, стенды, макеты, компьютерная техника, предустановленное программное обеспечение, используемое для различных видов занятий
Лабораторные занятия	327 (3б)	Компьютер.
Самостоятельная работа студента	332 (3б)	Компьютер
Лекции	336 (3б)	Компьютер, проектор.