Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Директор Института магистратуры
_____Иванова Е.А.
«__22__» __февраля__ 2022___г.

Рабочая программа дисциплины
Обработка естественного языка

Направление 01.04.02 Прикладная математика и информатика
магистерская программа 01.04.02.04 "Искусственный интеллект: математические модели и прикладные решения"

Для набора 2022 года

Квалификация
Магистр

Составитель программы:
Алексейчик Тамара Васильевна, к.э.н., доц, кафедры фундаментальной и прикладной математики

Программа одобрена на заседании кафедры высшей фундаментальной и прикладной математики «22» февраля 2022 г., протокол №6

ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цели освоения дисциплины: выработка у студентов компетенций, связанных с их теоретической и практической подготовкой к использованию средств и методов анализа и математического моделирования структур естественного языка.

Задачи:

овладеть теоретическими знаниями и практическими навыками использования средств и методов анализа и математического моделирования структур естественного языка;

разъяснить ограничения и особенности применения различных методов анализа и математического моделирования структур естественного языка;

практическое применение современных программных средств и специализированных библиотек для обработки текстов.

МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

2.1. Учебная дисциплина «Обработка естественного языка» (2 курс магистратуры, 3 семестр) относится к части блока дисциплин, формируемой участниками образовательных отношений, и является дисциплиной по выбору. Дисциплина преподается на русском языке.

2.2. Для изучения данной учебной дисциплины необходимы знания, умения и навыки, формируемые дисциплинами, изучаемыми на 1 курсе магистратуры «Избранные вопросы теории вероятностей и математической статистики», «Питон для анализа данных», «Основы нейронных сетей».

2.3. Знания и навыки, полученные в ходе изучения данной дисциплины (модуля), могут использоваться для решения профессиональных задач в научно-исследовательской, научно-производственной и проектной деятельности, в частности, при выполнении выпускной квалификационной работы.

ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс изучения дисциплины направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО и ОП ВО по данному направлению подготовки (специальности):

Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы:

Шифр и формулировка компетенций (результаты освоения ОП)	**Индикаторы компетенций**	**Элементы компетенций, формируемые дисциплиной**
Общекультурные и профессиональные компетенции (ОК и ПК)		
ПК-6. Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых субтехнологий искусственного интеллекта в прикладных областях	ПК-6.2. Руководит проектами в области сквозной цифровой субтехнологии «Обработка естественного языка»	ПК-6.2. З-1. 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» ПК-6.2. У-1. Умеет руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка»

СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 5 зачетных единиц, 180 часов,
из них 34 часа лекционных занятий, 34 часа лабораторных занятий, 112 часов на самостоятельную работу в течение семестра, включая 36 часов на подготовку к экзамену.

Форма отчетности: экзамен

4.1 Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

№ п/п	Раздел дисциплины/темы	Семестр	Виды учебной работы, включая самостоятельную работу обучающихся и трудоемкость (в часах)				Формы текущего контроля успеваемости Форма промежуточной аттестации *(по семестрам)*
			Контактная работа преподавателя с обучающимися			Самостоятельная работа	
			Лекции	Семинарские (практические занятия)	Лабораторные занятия		
1	**Раздел 1. Построение векторной модели текста и классификация длинных текстов.**	**3**	**12**	**0**	**12**	**25**	**Результаты выполнения лабораторных работ. Ответ на экзамене**
1.1	Введение. Естественные языки, особенности обработки естественного языка. Обзор основных задач обработки текстов на естественном языке.	3	2	0	0	6	
1.2	Методы построения векторной модели теста. Программные продукты и библиотеки, используемые для построения векторной модели текста.	3	6	0	8	12	
1.3	Применение сверточных нейронных сетей для решения задач обработки текстов.	3	4		4	7	

№ п/п	Раздел дисциплины/темы	Семестр	Виды учебной работы, включая самостоятельную работу обучающихся и трудоемкость (в часах)				Формы текущего контроля успеваемости Форма промежуточной аттестации *(по семестрам)*
			Контактная работа преподавателя с обучающимися			Самостоятельная работа	
2	**Раздел 2. Статистические языковые модели.**	**3**	**12**	**0**	**12**	**28**	**Результаты выполнения лабораторных работ. Ответ на экзамене.**
2.1	Моделирование языка.	3	8	0	8	16	
2.2	Применение рекуррентных нейронных сетей для решения задачи генерации текстов.	3	4	0	4	12	
3	**Раздел 3. Распознавание структуры коротких текстов и преобразование последовательностей.**	**3**	**10**	**0**	**10**	**23**	**Результаты выполнения лабораторных работ. Ответ на экзамене.**
3.1	Методы решения задачи выделения фрагментов текста и их соотнесения с заданными классам.	3	5	0	5	13	
3.2	Методы преобразования последовательностей.	3	5	0	5	10	
Итого часов			**34**	**0**	**34**	**76**	
Подготовка к экзамену		3				**36**	
Всего часов - 180		3	**34**	**0**	**34**	**112**	

4.2 План внеаудиторной самостоятельной работы обучающихся по дисциплине

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Затраты времени (час.)		

1	Раздел 1. Построение векторной модели текста и классификация длинных текстов.	Изучение учебной и научной литературы, материалов Интернет, программирование	6 недель	25	Результаты выполнения лабораторных работ. Ответ на экзамене	Материалы лекций, рекомендованная научная литература, материалы Интернет
2	Раздел 2. Статистические языковые модели.	Изучение учебной и научной литературы, материалов Интернет, программирование	6 недель	28	Результаты выполнения лабораторных работ. Ответ на экзамене	
3	Раздел 3. Распознавание структуры коротких текстов и преобразование последовательностей.	Изучение учебной и научной литературы, материалов Интернет, программирование	5 недель	23	Результаты выполнения лабораторных работ. Ответ на экзамене	
Подготовка к экзамену				**36**		
Общая трудоемкость самостоятельной работы по дисциплине (час)				**112**		
Бюджет времени самостоятельной работы, предусмотренный учебным планом для данной дисциплины (час)				**112**		

4.3 Содержание учебного материала

Раздел 1. Построение векторной модели текста и классификация длинных текстов.

Тема 1.1. Введение. Естественные языки, особенности обработки естественного языка. Обзор основных задач обработки текстов на естественном языке.

Естественные и формальные языки. Особенности естественного языка: нечетко зафиксированные правила, правила, неоднозначность. Основные группы задач работы с текстом на естественном языке: лингвистический анализ, извлечение признаков из текстов, прикладные задачи обработки текстов, генерация текста.

Тема 1.2. Методы построения векторной модели теста. Программные продукты и библиотеки, используемые для построения векторной модели текста.

Способы построения векторной модели текста: мешок слов, n-граммы, T-IDF. Дистрибутивно-семантические модели. Программная реализация модели Word2Vec. Модель FastText.

Тема 1.3. Применение сверточных нейронных сетей для решения задач обработки текстов.

Архитектура глубоких нейронных сетей, применяемых для решения задач обработки и генерации текстов.

Раздел 2. Статистические языковые модели.

Тема 2.1. Моделирование языка.

Основные проблемы моделирования языка. Моделирование естественного языка в задачах машинного перевода, исправления текста и др.

Тема 2.2. Применение рекуррентных нейронных сетей для решения задачи генерации текстов.

Рекуррентные нейронные сети и сети с долговременной памятью. Задача генерации текста. Модель Transformer.

Раздел 3. Распознавание структуры коротких текстов и преобразование последовательностей.

Тема 3.1. Методы решения задачи выделения фрагментов текста и их соотнесения с заданными классам.

Методы распознавания плоской структуры текстов и примеры их практического применения.

Тема 3.2. Методы преобразования последовательностей.

Методыseq2seq, Skip-Gram, ContinuousBagofWords, GloVe: сравнение и особенности применения.

Студент самостоятельно (при консультации преподавателя) изучает рекомендованную литературу и разбирает алгоритмы обработки текстов, описанные в литературе. Студент самостоятельно собирает, обрабатывает и анализирует научную информацию, необходимую для выполнения лабораторных работ, проводит самостоятельные исследования, используя современные информационные технологии, включая создание собственных компьютерных программ и использование готового программного обеспечения. Студент самостоятельно готовит отчет о выполнении каждой лабораторной работы.

ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

При проведении лекций и лабораторных занятий используются следующие образовательные технологии:

— классические лекции,

— мультимедийные лекции,
— лекции-презентации,
— семинары-дискуссии.

Учебный процесс базируется на концепции компетентностного обучения, ориентированного на формирование конкретного перечня профессиональных компетенций, актуализацию получаемых теоретических знаний. Развертывание компетентностной модели обучения предполагает широкое применение инновационных способов организации учебного процесса, в т.ч. применение метода проектного обучения, технологий управляемого самостоятельного обучения в том числе балльно-рейтинговой системы, а также внедрение системы онлайн-поддержки внеаудиторной работы студентов.

Дисциплина может быть реализована частично или полностью с использованием ЭИОС Университета (ЭО и ДОТ). Аудиторные занятия и другие формы контактной работы обучающихся с преподавателем могут проводиться с использованием платформ Microsoft Teams, ZOOM, Skype, MOODLE и др., в том числе, в режиме онлайн-лекций.

ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Полный комплект контрольно-оценочных материалов (Фонд оценочных средств) оформляется в виде приложения к рабочей программе дисциплины.

УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

1.1. Основная литература.

{Печатный ресурс} Тушан, Ганегедара Обработка естественного языка с TensorFlow / пер. с анг. В. С. Яценкова. – М.:ДМК Пресс, 2020. – 382 с.: ил., ISBN 978-5-97060-756-5

Гольдберг Й. Нейросетевые методы в обработке естественного языка / Й. Гольдберг; перевод с английского. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1, — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131704 (дата обращения: 10.10.2021). — Режим доступа: для авториз. пользователей.

1.2. Дополнительная литература.

- [**Электронный ресурс:** https://ebookcentral.proquest.com/lib/hselibrary-ebooks/reader.action?docID=4035461&query=computational+linguistics;] The Handbook of Computational Linguistics and Natural Language Processing / ed. by Alexander Clark, Chris Fox, Shalom Lappin; DB ebrary. – Chichester: John Wiley & Sons, 2013. – 203 p.

7.4. Периодические издания *(при необходимости)*

Периодические издания рекомендуются научным руководителем по выполнению ВКР.

7.5. Перечень ресурсов сети Интернет, необходимых для освоения дисциплины

Университетская библиотека online: http://biblioclub.ru/index.php?page=main_ub_red
Электронно-библиотечная система (ЭБС) ЮРАЙТ https://urait.ru/

7.6. Программное обеспечение информационно-коммуникационных технологий

Операционная система Microsoft Windows, пакет Microsoft Office

МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

8.1. Учебно-лабораторное оборудование

При проведении дисциплины учащиеся должны быть обеспечены:

1. Лекционной аудиторией с мультимедийным презентационным оборудованием для демонстрации презентаций и иллюстративного материала.

2. Аудиторией для лабораторных занятий с аппаратными и программными средствами в соответствии с реализуемой учебной тематикой.

8.2. Программные средства

Microsoft Windows, Microsoft Office, Windows CAL's - Договор 232.02.02.03-16/60 от 10.08.2018 г., с 10.08.2018 г. по 10.08.2019 г.; Договор №232.02.02.03-16/46 от 30.08.2019 г., с 31.07.2019 г. по 30.07.2020 г.; Государственный контракт № SC-P/5679-01/07 от 04.12.2007 г., с 21.12.2007 г. (срок использования ПО неограничен)

Среда программирования Microsoft Visual Studio (любая версия)

Среда разработки Jupyter-ноутбук (любая версия)

Методические указания для обучающихся по освоению дисциплины

Методические указания приведены в учебных пособиях, перечисленных в разделе VII.

УЧЕБНАЯ КАРТА ДИСЦИПЛИНЫ

Б1.В.ДВ.04.02. Обработка естественного языка

(Natural Language Processing)

Трудоемкость: 5 зач.ед.

Форма промежуточной аттестации: экзамен

Курс 2, семестр 3

Код и наименование направления подготовки (специальности): 01.04.02 «Прикладная математика и информатика» (магистратура)

Магистерская программа: «Искусственный интеллект: математические модели и прикладные решения»

№	**Виды контрольных мероприятий**	**Текущий контроль**	**Рубежный контроль** *(при наличии)*
	Модуль 1. Построение векторной модели текста и классификация длинных текстов.	**20**	
1.	Лабораторные занятия	20	
	Модуль 2. Статистические языковые модели.	**20**	
1.	Лабораторные занятия	20	
	Модуль 3 Распознавание структуры коротких текстов и преобразование последовательностей.	**20**	
1.	Лабораторные занятия	20	
	Всего	**60**	
	Промежуточная аттестация ***в форме экзамена***	**40** баллов	**Экзамен проводится письменно по билетам. Критерий оценивания сформулирован в ФОС**
	ИТОГО	**100 баллов**	

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»
Факультет компьютерных технологий и защиты информации
Кафедра фундаментальной и прикладной математики

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ

Б1.В.ДВ.04.02. ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

Код и наименование направления подготовки/специальности:
01.04.02 «Прикладная математика и информатика»

Уровень образования:
Магистратура

Магистерская программа:
«Искусственный интеллект: математические модели и прикладные решения»

Форма обучения:
Очная

Ростов-на-Дону, 2021

ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ, ФОРМИРУЕМЫХ ДИСЦИПЛИНОЙ «Обработка естественного языка»

Код компетенции	Формулировка компетенции
1	2
ПК	**ПРОФЕССИОНАЛЬНЫЕ КОМПЕТЕНЦИИ**
ПК-6	Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых субтехнологий искусственного интеллекта в прикладных областях

ПАСПОРТ ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ
«Обработка естественного языка»

№ п/п	*Контролируемые разделы дисциплины**	*Код контролируемой компетенции*	Наименование *оценочного средства***
1.	**Построение векторной модели текста и классификация длинных текстов.**	ПК-6	Лабораторные работы. Экзаменационные билеты
2.	**Статистические языковые модели.**	ПК-6	Лабораторные работы. Экзаменационные билеты
3.	**Распознавание структуры коротких текстов и преобразование последовательностей.**	ПК-6	Лабораторные работы. Экзаменационные билеты

** Наименование раздела указывается в соответствии с рабочей программой дисциплины.*
***Наименование оценочного средства указывается в соответствии с учебной картой дисциплины.*

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»
Факультет компьютерных технологий и защиты информации
Кафедра фундаментальной и прикладной математики

Вопросы к экзамену
по дисциплине **«Обработка естественного языка»**

1. Естественные языки, особенности обработки естественного языка.
2. Основные задачи обработки текстов на естественном языке: лингвистический анализ, извлечение признаков из текстов, прикладные задачи обработки текстов, генерация текста.
3. Метод s построения векторной модели теста: мешок слов, n-граммы.
4. Метод T-IDF.
5. Модель Word2Vec.
6. Метод Skip-Gram.
7. Метод Continuous Bag of Words.
8. Метод GloVe.
9. Модель FastText
10. Дистрибутивно-семантические модели.
11. Программные продукты и библиотеки, используемые для построения векторной модели текста.
12. Примеры применения сверточных нейронных сетей для решения задач обработки текстов.
13. Моделирование языка.
14. Применение рекуррентных нейронных сетей для решения задачи генерации текстов.
15. Методы решения задачи выделения фрагментов текста и их соотнесения с заданными классам.
16. Методы преобразования последовательностей.

Форма экзаменационного билета

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»
Факультет компьютерных технологий и защиты информации
Кафедра фундаментальной и прикладной математики

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №

По дисциплине **“**Обработка естественного языка”

II. Направление/специальность 01.04.02 «Прикладная математика и информатика»

1 Вопрос: Метод T-IDF.

2 Вопрос: Применение рекуррентных нейронных сетей для решения задачи генерации текстов

.

Составитель _____
(подпись)

Заведующий кафедрой _____
(подпись)

Критерии оценки

Баллы начисляются по результатам ответов на вопросы в соответствие с приведенной таблицей.

Оценочные таблицы экзамена по дисциплине «Обработка естественного языка»

1. Ответ на экзамене был полным и ясным?

	Абсолютно
	В основном да
	Лишь частично
	Практически нет
	Нет

2. Соответствует ли структура экзаменационного ответа особенностям излагаемого материала?

	Абсолютно
	В основном да
	Лишь частично
	Практически нет
	Нет

3. Соответствует ли содержание экзаменационного ответа уровню студента выпускного курса магистратуры?

	Абсолютно
	В основном да
	Лишь частично
	Практически нет
	Нет

4. Соответствует ли содержание ответов на экзамен современным знаниям в данной области и соответствует ли оно целям магистерской программы?

	Абсолютно
	В основном да
	Лишь частично
	Практически нет
	Нет

5. Имеется ли ясная связь между текстом, формулами, графиками, рисунками и примерами? Подходит ли объем материала, который представлен в экзаменационном ответе, для изучения на уровне студента магистратуры?

	Абсолютно
	В основном да
	Лишь частично
	Практически нет
	Нет

Таким образом, максимальное количество баллов, которое может получить студент за ответ на экзамене, – 40.

№	Абсолютно	В основном да	Лишь частично	Практически нет	Нет
1	8	6	4	2	0
2	8	6	4	2	0
3	8	6	4	2	0
4	8	6	4	2	0
5	8	6	4	2	0

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»
Факультет компьютерных технологий и защиты информации
Кафедра фундаментальной и прикладной математики

Информатики и вычислительного эксперимента

Перечень лабораторных работ

Работа № 1. Решить задачу тематической классификации корпуса текстов. Для решения задачи использовать разреженную векторную модель текста и обучить логистическую регрессионную модель. Оценить качество работы модели. Оформить отчет.

Работа № 2. Дан текст, в котором есть пропущенные слова. Решить задачу генерации слов для заполнения пропусков, для этого выполнить необходимую подготовительную работу и обучить рекуррентную нейронную сеть. Оценить качество работы нейронной сети. Оформить отчет.

Работа № 3. Решить задачу анализа текстов для предложенного массива отзывов об автомобилях. Обучить нейронную сеть, которая сможет автоматически выделить основные моменты упоминаемые в отзыве — например, цвет, пробег, комфорт, потребление горючего, и так далее. Оценить качество работы нейронной сети. Оформить отчет.

Критерии оценки

20 баллов ставится в том случае, если работа выполнена полностью, отчет оформлен в соответствии с предъявленными требованиями, студент глубоко изучил учебный материал по соответствующему разделу программы курса, правильно, уверенно, последовательно и исчерпывающе комментирует выполненную работу, теоретические знания иллюстрирует примерами из практики.

15 баллов ставится тогда, когда работа выполнена полностью, оформление отчета не полностью соответствует предъявленным требованиям, студент относительно хорошо знает материал, разбирается в литературе, но имеет некоторые погрешности в ответах.

10 балов ставится при условии, что в работе имеются ошибки, но общая схема выполнения работы правильна, студент на заданные вопросы отвечает, но недостаточно четко и полно.

5 баллов ставится, если работа не доведена до конца, существенные шаги не сделаны, студент отвечает только на часть заданных вопросов.

0 баллов ставится в том случае, когда работа не доведена до конца, существенные шаги не сделаны, студент не смог правильно ответить на поставленные вопросы.

Таким образом, максимальное количество баллов, которое может получить студент за лабораторную работу – 20.