Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

Кафедра математики и цифровых технологий

РАБОЧАЯ ПРОГРАММА

ДИСЦИПЛИНЫ

«Б1.Д.В.Э.1.1 Компьютерная лингвистика»

Уровень высшего образования

МАГИСТРАТУРА

Направление подготовки
09.04.02 Информационные системы и технологии
(код и наименование направления подготовки)

Искусственный интеллект в промышленности
(наименование направленности (профиля) образовательной программы)

Квалификация
Магистр

Форма обучения
Очная

Год набора 2025

Рабочая программа дисциплины «*Б1.Д.В.Э.1.1 Компьютерная лингвистика*» рассмотрена и утверждена на заседании кафедры

Кафедра математики и цифровых технологий
наименование кафедры

протокол № 6 от "19" февраля 2025 г.

Заведующий кафедрой
Кафедра математики и цифровых технологий _____ А.Е. Шухман
наименование кафедры *подпись* *расшифровка подписи*

Исполнители:
Зав. кафедрой _____ А.Е. Шухман
должность *подпись* *расшифровка подписи*

СОГЛАСОВАНО:
Председатель методической комиссии по направлению подготовки
09.04.02 Информационные системы и технологии _____ И.П. Болодурина
код наименование *личная подпись* *расшифровка подписи*

Научный руководитель магистерской программы _____ И.П. Болодурина
личная подпись *расшифровка подписи*

Заведующий отделом формирования фонда и научной обработки документов
_____ С.Н. Биктимирова
личная подпись *расшифровка подписи*

Уполномоченный по качеству института
_____ С.Н. Морозова
личная подпись *расшифровка подписи*

№ регистрации _____

© Шухман А.Е., 2025
© ОГУ, 2025

1 Цели и задачи освоения дисциплины

Познакомить студентов с актуальным направлением искусственного интеллекта – современными методами обработки естественных языков.

Задачи:

– Знакомство с основными методами предобработки текстов, применением векторных моделей.

– Изучение нейросетевых методов работы с текстом.

– Приобретение опыта решения базовых задач обработки текстов, в том числе на русском языке.

2 Место дисциплины в структуре образовательной программы

Дисциплина относится к дисциплинам (модулям) по выбору вариативной части блока Д «Дисциплины (модули)»

Пререквизиты дисциплины: *Б1.Д.Б.6 Теоретическая информатика, Б1.Д.Б.10 Моделирование сложных систем*

Постреквизиты дисциплины: *Отсутствуют*

3 Требования к результатам обучения по дисциплине

Процесс изучения дисциплины направлен на формирование следующих результатов обучения

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
ПК*-3 Способен использовать методы научных исследований и математического моделирования в области проектирования и управления системами искусственного интеллекта, в том числе универсального искусственного интеллекта	ПК*-3-В-1 Применяет логические методы и приемы научного исследования, методологические принципы современной науки, направления, концепции, источники знания и приемы работы с ними, основные особенности научного метода познания, программно-целевые методы решения научных проблем в профессиональной деятельности ПК*-3-В-2 Осуществляет методологическое обоснование научного исследования, создание и применение библиотек искусственного интеллекта	**Знать**: возможности библиотек обработки текстов **Уметь**: разрабатывать прикладные программы обработки текстов **Владеть**: инструментами обработки текстов, в том числе на русском языке
ПК*-11 Способен управлять процессами разработки и сопровождения требований к системам, способен управлять качеством систем	ПК*-11-В-1 Управляет процессами разработки и сопровождения требований к системам	**Знать:** основные методы автоматизированной обработки естественных языков **Уметь**: решать базовые задачи обработки естественных языков **Владеть**: методами анализа алгоритмов обработки естественных языков

4 Структура и содержание дисциплины

4.1 Структура дисциплины

Общая трудоемкость дисциплины составляет 3 зачетные единицы (108 академических часов).

Вид работы	Трудоемкость, академических часов	
	3 семестр	всего
Общая трудоёмкость	**108**	**108**
Контактная работа:	**34,25**	**34,25**
Лекции (Л)	18	18
Лабораторные работы (ЛР)	16	16
Промежуточная аттестация (зачет, экзамен)	0,25	0,25
Самостоятельная работа: *- выполнение индивидуального творческого задания (ИТЗ);* *- самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий;* *- изучение разделов курса в системе электронного обучения;* *- подготовка к лабораторным занятиям;* *- подготовка к коллоквиумам;* *- подготовка к рубежному контролю).*	**73,75**	**73,75**
Вид итогового контроля (зачет, экзамен, дифференцированный зачет)	**зачет**	

Разделы дисциплины, изучаемые в 3 семестре

№ раздела	Наименование разделов	Количество часов				
		всего	аудиторная работа			внеауд. работа
			Л	ПЗ	ЛР	
1	Основы компьютерной лингвистики	6	2			4
2	Предобработка текстов	18	4		4	10
3	Векторные модели текстов	28	4		4	20
4	Базовые нейросетевые методы работы с текстами	28	4		4	20
5	Языковые модели и генерация текста	28	4		4	20
	Итого:	108	18		16	74
	Всего:	108	18		16	74

4.2 Содержание разделов дисциплины

1. Основы компьютерной лингвистики

Основные задачи обработки естественных языков. Виды лингвистического анализа: лексический анализ (токенизация), синтаксический анализ, морфологический анализ, семантический анализ. Основные методы анализа текстов. Лингвистические ресурсы. Оценка качества анализа текстов.

2. Предобработка текстов

Стемминг и лематизация. Регулярные выражения. Основные правила. Применение регулярных выражений.

3. Векторные модели текстов.

Признаки текстов. Векторные модели, TF_IDF. Выделение ключевых слов. Выделение n-грамм и коллокаций. Классификация текстов на основе векторных моделей.

4. Нейросетевые методы работы с текстами

Общий алгоритм работы с текстами с помощью нейросетей. Дистрибутивная семантика и векторные представления слов. Свёрточные нейросети для обработки текстов. Рекуррентные нейросети.

5. Языковые модели и генерация текста.

Моделирование языка. Агрегация, механизм внимания. Трансформеры. Генерация последовательностей. Контекстуализированные представления и перенос обучения. Большие языковые модели. Дообучение БЯМ. Методы ускорения БЯМ. Промпты. Дополненные промпты.

4.3 Лабораторные работы

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов
1	2	Токенизация, стемминг, лематизация	2
2	2	Регулярные выражения	2
3	3	Векторные модели	2
4	3	Классификация текстов	2
5	4	Нейросетевая классификация текстов	2
6	4	Применение трансформеров. Перенос обучения	2
7	5	Промптинг больших языковых моделей	2
8	5	Технология дополненного промптинга	2
		Итого:	16

5 Учебно-методическое обеспечение дисциплины

5.1 Основная литература

1. Корягин, С. В. Методы анализа естественно-языковых текстов : учебно-методическое пособие / С. В. Корягин, А. М. Русаков. — Москва : РТУ МИРЭА, 2023. — 86 с. — ISBN 978-5-7339-1737-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/331649

2. Обработка естественного языка с использованием языка программирования Python : учебное пособие : в 2 частях / составитель А. Б. Мантусов. — Элиста : КГУ, 2022 — Часть 1 — 2022. — 56 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/360923

5.2 Дополнительная литература

1. Щипицина, Л. Ю. Информационные технологии в лингвистике [Текст] : учебное пособие / Л. Ю. Щипицина.- 3-е изд., стер. - Москва : Флинта : Наука, 2017. - 128 с. : ил. - Библиогр.: с. 105-110. - Прил.: с. 111-124. - ISBN 978-5-9765-1431-7.

2. Прикладная и компьютерная лингвистика [Текст] : коллективная монография / под ред.: И.С. Николаева, О. В. Митрениной, Т. М. Ландо.- 2-е изд. - Москва : ЛЕНАНД, 2017. - 315 с. : ил. - Библиогр. в конце гл. - Указ. терминов: с. 309-315. - ISBN 978-5-9710-4633-2

5.3 Интернет-ресурсы

1. http://machinelearning.ru - Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.
2. https://nlpub.mipt.ru/ — каталог ресурсов для обработки естественного языка.
3. https://stepik.org/course/54098/syllabus «Степик»; МООК: «Нейронные сети и обработка текстов»
4. https://rusvectores.org/ru/ семантические модели для русского языка

5.4 Программное обеспечение, профессиональные базы данных и информационные справочные системы

1. Операционная система РЕД ОС для рабочих станций, имеется лицензия, входит в реестр отечественного ПО.
2. LibreOffice – свободно распространяемый офисный пакет программ, включающий в себя текстовый и табличный редакторы, редактор презентаций и другие офисные приложения.
3. Система управления учебным процессом Moodle, свободно распространяемая.
4. Программная система для организации видео-конференц-связи «DION» (Конфигурация «DION EDU») На основании договора № 13/223-4.2.1.35/40-03 от 14.02.2025 г. Срок действия лицензий с 14.02.2025 г по 14.02.2026.
5. Программа для просмотра сайтов Яндекс.Браузер, свободно распространяемая, входит в реестр отечественного ПО.
6. Система программирования Python, свободно распространяемая по лицензии GPL

6 Материально-техническое обеспечение дисциплины

Учебные аудитории для проведения занятий лекционного типа, семинарского типа, для проведения групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации.

Аудитории оснащены комплектами ученической мебели, техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Для проведения лабораторных занятий используется компьютерный класс, оснащенный компьютерной техникой, подключенной к сети «Интернет».

Помещение для самостоятельной работы обучающихся оснащены компьютерной техникой, подключенной к сети "Интернет", и обеспечением доступа в электронную информационно-образовательную среду ОГУ.