МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Воронежский государственный технический университет»

УТВЕРЖДАЮ

И.о. декана факультета информационных технологий и компьютерной безопасности

_____/ А.В. Бредихин /
19.03.2024

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Технологии обработки естественного языка»

Направление подготовки 09.03.02 Информационные системы и технологии

Профиль Технологии искусственного интеллекта

Квалификация выпускника бакалавр

Нормативный период обучения 4 года

Форма обучения очная

Год начала подготовки 2024

Автор программы _____ В.Н. Кострова

И.о. заведующего кафедрой систем автоматизированного проектирования и информационных систем _____ П.Ю. Гусев

Руководитель ОПОП _____ Д.В. Иванов

Воронеж 2024

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

1.1. Цели дисциплины

ознакомление студентов с методами и технологиями в области искусственного интеллекта на основе принципов построения систем обработки естественного языка

1.2. Задачи освоения дисциплины

овладение основными приложениями, подходами, источниками данных и инструментами обработки естественного языка, современными методами анализа естественного языка, основанными на нейронных сетях и машинном обучении

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Технологии обработки естественного языка» относится к дисциплинам части, формируемой участниками образовательных отношений блока Б1.

3. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Процесс изучения дисциплины «Технологии обработки естественного языка» направлен на формирование следующих компетенций:

ПК-5 - Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта

ПК-7 - Способен использовать знание основных методов искусственного интеллекта в последующей профессиональной деятельности в качестве научных сотрудников, преподавателей образовательных организаций высшего образования, инженеров, технологов

Компетенция	Результаты обучения, характеризующие сформированность компетенции
ПК-5	знать современные методы анализа естественного языка и принципы построения систем обработки естественного языка
	уметь применять методы и технологии искусственного интеллекта для анализа и обработки естественного языка
	владеть различными алгоритмами машинного обучения и архитектурами искусственных нейронных сетей в рамках задач обработки естественного языка
ПК-7	знать принципы, методы и средства анализа и структурирования профессиональной информации для решения задач обработки естественного языка
	уметь применять основные средства искусственного интеллекта с целью построения систем обработки естественного языка в последующей профессиональной деятельности

	владеть методами и технологиями искусственного интеллекта для анализа и обработки естественного языка при решении профессиональных задач, в том числе в междисциплинарном контексте

4. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины «Технологии обработки естественного языка» составляет 3 з.е.

Распределение трудоемкости дисциплины по видам занятий

очная форма обучения

Виды учебной работы	Всего часов	Семестры
		7
Аудиторные занятия (всего)	54	54
В том числе:		
Лекции	18	18
Лабораторные работы (ЛР)	36	36
Самостоятельная работа	54	54
Часы на контроль	-	-
Виды промежуточной аттестации - зачет	+	+
Общая трудоемкость: академические часы зач.ед.	 108 3	 108 3

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1 Содержание разделов дисциплины и распределение трудоемкости по видам занятий

очная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Лаб. зан.	СРС	Всего, час
1	Теоретические аспекты обработки естественного языка	Введение в обработку естественного языка (NLP). Основные задачи NLP. Представления текстовых данных. Предобработка текста, лемматизация, стемминг.	2	4	7	13
2	Машинное обучение и нейронные сети для решения задач анализа и обработки естественного языка	Методы машинного обучения для классификации текстовых документов на основе частотных мер (TF-IDF). Деревья решений, наивный байесовский классификатор, логистическая регрессия в задаче классификации текстов.	3	6	8	17
3	Машинное обучение и нейронные сети для	Языковые модели. Нейросетевые модели языка. Мера семантической близости. Классификация текстов	2	4	7	13

	решения задач анализа и обработки естественного языка	на основе нейросетевых моделей языка.				
4	Машинное обучение и нейронные сети для решения задач анализа и обработки естественного языка	Кластеризация текстовых документов. Тематическое моделирование Методы LSA, pLSA. Аддитивная регуляризация тематических моделей.	2	6	8	16
5	Машинное обучение и нейронные сети для решения задач анализа и обработки естественного языка	Классификация текстов с помощью глубоких нейронных сетей: CNN, LSTM.	3	6	8	17
6	Машинное обучение и нейронные сети для решения задач анализа и обработки естественного языка	Задачи обработки последовательностей: машинной перевод, автоматическое реферирование (summarization), вопросно-ответные системы. Механизм внимания (attention). Архитектуры encoder-decoder-attention.	3	6	8	17
7	Машинное обучение и нейронные сети для решения задач анализа и обработки естественного языка	Transfer learning в задачах анализа текстов. Self-Attention. Архитектуры трансформеров: BERT, GPT в задачах классификации текстов, предсказания пропущенных слов, генерации текстов. Fine-tuning трансформеры.	3	4	8	15
		Итого	**18**	**36**	54	108

5.2 Перечень лабораторных работ

Предварительная обработка текста для анализа.

Классификация текста с использованием классических методов машинного обучения.

Классификация текста с использованием глубоких нейронных сетей.

Языковая модель. Обучение языковой модели.

Автоматическая генерация текста.

Поиск именованных объектов в тексте.

Механизм внимания в нейронных сетях. Сети с трансформаторной архитектурой.

Передача обучения в задачах обработки текстов.

6. ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ ПРОЕКТОВ (РАБОТ) И КОНТРОЛЬНЫХ РАБОТ

В соответствии с учебным планом освоение дисциплины не предусматривает выполнение курсового проекта (работы) или контрольной работы.

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

7.1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

7.1.1 Этап текущего контроля

Результаты текущего контроля знаний и межсессионной аттестации оцениваются по следующей системе:

«аттестован»;

«не аттестован».

Компе-тенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Аттестован	Не аттестован
ПК-5	знать современные методы анализа естественного языка и принципы построения систем обработки естественного языка	Выполнение тестового задания	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	уметь применять методы и технологии искусственного интеллекта для анализа и обработки естественного языка	Выполнение лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	владеть различными алгоритмами машинного обучения и архитектурами искусственных нейронных сетей в рамках задач обработки естественного языка	Выполнение лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ПК-7	знать принципы, методы и средства анализа и структурирования профессиональной информации для решения задач обработки естественного языка	Выполнение тестового задания	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	уметь применять основные средства искусственного интеллекта с целью	Выполнение лабораторных работ	Выполнение работ в срок, предусмотренны	Невыполнение работ в срок, предусмотренн

	построения систем обработки естественного языка в последующей профессиональной деятельности		й в рабочих программах	ый в рабочих программах
	владеть методами и технологиями искусственного интеллекта для анализа и обработки естественного языка при решении профессиональных задач, в том числе в междисциплинарном контексте	Выполнение лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

7.1.2 Этап промежуточного контроля знаний

Результаты промежуточного контроля знаний оцениваются в 7 семестре для очной формы обучения оцениваются по следующей системе:

«зачтено»;

«не зачтено».

Компе-тенция	**Результаты обучения, характеризующие сформированность компетенции**	**Критерии оценивания**	**Отлично**	**Хорошо**	**Удовл.**	**Неудовл.**
ПК-5	знать современные методы анализа естественного языка и принципы построения систем обработки естественного языка	Тест	Выполнение теста на 70-100%	Выполнение менее 70%	Выполнение теста на 70- 80%	В тесте менее 70% правильных ответов
	уметь применять методы и технологии искусственного интеллекта для анализа и обработки естественного языка	Решение стандартных задач	Выполнение на 70-100%	Выполнение менее 70%	Продемонстр ирован верный ход решения в большинстве задач	Задачи не решены
	владеть различными алгоритмами машинного	Решение прикладных задач	Выполнение на 70-100%	Выполнение менее 70%	Продемонстр ирован верный ход решения в	Задачи не решены

	обучения и архитектурами искусственных нейронных сетей в рамках задач обработки естественного языка				большинстве задач	
ПК-7	знать принципы, методы и средства анализа и структурирования профессиональной информации для решения задач обработки естественного языка	Тест	Выполнение теста на 70-100%	Выполнение менее 70%	Выполнение теста на 70- 80%	В тесте менее 70% правильных ответов
	уметь применять основные средства искусственного интеллекта с целью построения систем обработки естественного языка в последующей профессиональной деятельности	Решение стандартных задач	Выполнение на 70-100%	Выполнение менее 70%	Продемонстр ирован верный ход решения в большинстве задач	Задачи не решены
	владеть методами и технологиями искусственного интеллекта для анализа и обработки естественного языка при решении профессиональных задач, в том числе в междисциплинарном контексте	Решение прикладных задач	Выполнение на 70-100%	Выполнение менее 70%	Продемонстр ирован верный ход решения в большинстве задач	Задачи не решены

7.2 Примерный перечень оценочных средств (типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности)

7.2.1 Примерный перечень заданий для подготовки к тестированию

1. Какой из следующих методов можно использовать для нормализации ключевых слов в NLP, процесс преобразования ключевого слова в его

базовую форму:

а) Лемматизация

б) Soundex

с) Косинус сходства

д) N-грамм

2. Какой из методов синтаксического анализа текста можно использовать для обнаружения именной фразы, глагольной фразы, обнаружения субъекта и обнаружения объекта в NLP:

а) Тегирование части речи

б) Пропустить извлечение граммов и N-граммов

в) Непрерывный мешок слов

г) Анализ зависимостей и анализ групп

3. Что из нижеперечисленного относится к методам нормализации ключевых слов в NLP?

а) Стемминг

б) Часть речи

в) Распознавание именованных объектов

г) Лемматизация

4. Что из нижеперечисленного относится к вариантам использования NLP?

а) Обнаружение объектов на изображении

б) Распознавание лиц

в) Речь биометрическая

г) Обобщение текста

5. TF-IDF помогает вам установить?

а) наиболее часто встречающееся слово в документе

б) самое важное слово в документе

6. Что из нижеперечисленного не относится к методам предварительной обработки в NLP?

а) Стемминг и лемматизация

б) Преобразование в нижний регистр

в) Удаление знаков препинания

г) Удаление стоп-слов

д) Анализ настроений

7. В NLP слова, представленные в виде векторов, называются нейронными вложениями слов

а) Истинный

б) Ложь

8. В NLP поддерживается контекстное моделирование, с помощью которого одно из следующих вложений слов

а) Word2Vec

б) GloVe

в) BERT

г) Все вышеперечисленное

7.2.2 Примерный перечень заданий для решения стандартных задач

1. Необходимо обучить языковую модель русскому языку и использовать ее для создания текста. Для выполнения задачи необходимо выполнить следующее:

- подготовить набор данных с текстами на русском языке. Возможно использовать готовые наборы данных или создать свои собственные;
- обучить языковую модель на подготовленном наборе данных;
- используя обученную языковую модель, сгенерировать пять примеров текстов на русском языке;
- разместить набор данных, код и обученную модель в открытом доступе на GitHub.

2. Необходимо обучить предварительно обученную сеть с архитектурой Transformer для классификации текстов на русском языке. Для выполнения задачи необходимо выполнить следующее:

- подготовить набор данных с текстами на русском языке для классификации. Возможно использовать готовые наборы данных или создать свои собственные;
- выбрать предварительно обученную нейронную сеть с трансформаторной архитектурой, подходящую для задачи классификации текстов на русском языке;
- выполнить дополнительное обучение выбранной нейронной сети на подготовленном наборе данных;
- выполнить тестирование классификации текста с использованием обученной нейронной сети и оценить качество сети;
- поместить набор данных, код и завершенную модель в открытый доступ на GitHub.

7.2.3 Примерный перечень заданий для решения прикладных задач

1. Необходимо создать конвейеры для следующих задач обработки естественного языка:

- Классификация текста.
- Определение эмоциональной окраски текста.
- Автоматическая генерация текста.
- Поиск именованных объектов в тексте.

2. Необходимо разработать последовательность действий для решения задачи анализа текста с использованием машинного обучения. Конвейер должен включать:

- Способ подготовки текста к обработке.
- Подход к маркировке текста.
- Подход к векторизации текста.
- Используемая модель машинного обучения.
- Метод обучения модели.

- Метод оценки качества модели.
- Использование обученной модели для решения задачи анализа текста.
- Другие шаги, которые могут потребоваться при решении проблемы.

7.2.4 Примерный перечень вопросов для подготовки к зачету

1. Теоретические аспекты обработки естественного языка.
2. Особенности обработки текста на английском языке.
3. Особенности обработки текста на русском языке.
4. Предварительная обработка текста. Очистка текста. Удаление стопслов/наиболее и наименее частых слов.
5. Токенизация, вывод, лемматизация текста.
6. Классические методы машинного обучения для решения задач классификации текста.
7. Классические методы машинного обучения для решения задачи определения тональности текста.
8. Архитектуры нейронных сетей для обработки текста: LSTM.
9. Архитектуры нейронных сетей для обработки текста: GRU.
10. Архитектуры нейронных сетей для обработки текста: одномерные сверточные сети.
11. Классификация текста с использованием нейронных сетей.
12. Определение тональности текста с помощью нейронных сетей.
13. Языковая модель.
14. Обучение языковой модели.
15. Основные подходы к генерации текста.
16. Механизм внимания в нейронных сетях.
17. Применение механизма внимания для обработки текста.
18. Архитектура трансформаторной нейронной сети.
19. Предварительно обученные нейронные сети для обработки текста BERT.
20. Предварительно обученные нейронные сети для обработки текста GPT.
21. Передача обучения задачам обработки текстов.
22. Классификация текста с использованием сетей с трансформаторной архитектурой.
23. Генерация текста с использованием сетей с трансформаторной архитектурой.
24. Поиск именованных объектов в тексте с использованием сетей с архитектурой трансформатора.

7.2.5 Примерный перечень заданий для подготовки к экзамену

Не предусмотрено учебным планом

7.2.6. Методика выставления оценки при проведении промежуточной аттестации

Зачёт проводится по тест-билетам, каждый из которых содержит 10 вопросов. Каждый правильный ответ на вопрос в тесте оценивается 1 баллом. Максимальное количество набранных баллов – 10. Незачёт ставится в случае, если студент набрал менее 5 баллов.

7.2.7 Паспорт оценочных материалов

№ п/п	Контролируемые разделы (темы) дисциплины	Код контролируемой компетенции	Наименование оценочного средства
1	Введение в обработку естественного языка (NLP). Основные задачи NLP. Представления текстовых данных. Предобработка текста, лемматизация, стемминг.	ПК-5, ПК-7	Тест, защита лабораторных работ
2	Методы машинного обучения для классификации текстовых документов на основе частотных мер (TF-IDF). Деревья решений, наивный байесовский классификатор, логистическая регрессия в задаче классификации текстов.	ПК-5, ПК-7	Тест, защита лабораторных работ
3	Языковые модели. Нейросетевые модели языка. Мера семантической близости. Классификация текстов на основе нейросетевых моделей языка.	ПК-5, ПК-7	Тест, защита лабораторных работ
4	Кластеризация текстовых документов. Тематическое моделирование Методы LSA, pLSA. Аддитивная регуляризация тематических моделей.	ПК-5, ПК-7	Тест, защита лабораторных работ
5	Классификация текстов с помощью глубоких нейронных сетей: CNN, LSTM.	ПК-5, ПК-7	Тест, защита лабораторных работ
6	Задачи обработки последовательностей: машинной перевод, автоматическое реферирование (summarization), вопросно-ответные системы. Механизм внимания (attention). Архитектуры encoder-decoder-attention.	ПК-5, ПК-7	Тест, защита лабораторных работ
7	Transfer learning в задачах анализа текстов. Self-Attention. Архитектуры трансформеров: BERT, GPT в задачах классификации текстов, предсказания	ПК-5, ПК-7	Тест, защита лабораторных работ

	пропущенных слов, генерации текстов. Fine-tuning трансформеры.		

7.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Тестирование осуществляется с использованием выданных тест-заданий на бумажном носителе. Время тестирования 30 мин. Затем осуществляется проверка теста экзаменатором и выставляется оценка согласно методики выставления оценки при проведении промежуточной аттестации.

Тестирование осуществляется с использованием выданных тест-заданий на бумажном носителе. Время тестирования 30 мин. Затем осуществляется проверка теста экзаменатором и выставляется оценка согласно методики выставления оценки при проведении промежуточной аттестации.

8 УЧЕБНО МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ)

8.1 Перечень учебной литературы, необходимой для освоения дисциплины

1. Маккини, У. Python и анализ данных / У. Маккини ; перевод А. А. Слинкин. — 2-е изд. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/125361.html

2. Гольдберг, Й. Нейросетевые методы в обработке естественного языка / Й. Гольдберг ; перевод А. А. Слинкин. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/124564.html

3. Йылдырым, С. Осваиваем архитектуру Transformer. Разработка современных моделей с помощью передовых методов обработки естественного языка / С. Йылдырым, М. Асгари-Ченаглу. — Москва : ДМК Пресс, 2022. — 319 с. — ISBN 978-5-93700-106-1. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/125183.html

4. Ганегедара, Т. Обработка естественного языка с TensorFlow / Т. Ганегедара ; перевод В. С. Яценков. — Москва : ДМК Пресс, 2020. — 382 с. — ISBN 978-5-97060-756-5. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/130336.html

8.2 Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень лицензионного программного обеспечения, ресурсов

информационно-телекоммуникационной сети «Интернет», современных профессиональных баз данных и информационных справочных систем:
www.elbib.ru
www.eLibrary.ru
www.intuit.ru

9 МАТЕРИАЛЬНО-ТЕХНИЧЕСКАЯ БАЗА, НЕОБХОДИМАЯ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА

Специализированная лекционная аудитория, оснащенная оборудованием для лекционных демонстраций и проекционной аппаратурой

Компьютерный класс, оснащенный программным обеспечением для проведения лабораторного практикума

10. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

По дисциплине «Технологии обработки естественного языка» читаются лекции, проводятся лабораторные работы.

Основой изучения дисциплины являются лекции, на которых излагаются наиболее существенные и трудные вопросы, а также вопросы, не нашедшие отражения в учебной литературе.

Лабораторные работы выполняются на лабораторном оборудовании в соответствии с методиками, приведенными в указаниях к выполнению работ.

Вид учебных занятий	Деятельность студента
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначение вопросов, терминов, материала, которые вызывают трудности, поиск ответов в рекомендуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на лекции или на практическом занятии.
Лабораторная работа	Лабораторные работы позволяют научиться применять теоретические знания, полученные на лекции при решении конкретных задач. Чтобы наиболее рационально и полно использовать все возможности лабораторных для подготовки к ним необходимо: следует разобрать лекцию по соответствующей теме, ознакомится с соответствующим разделом учебника, проработать дополнительную литературу и источники, решить задачи и выполнить другие письменные задания.
Самостоятельная работа	Самостоятельная работа студентов способствует глубокому усвоения учебного материала и развитию навыков самообразования. Самостоятельная работа предполагает следующие составляющие: - работа с текстами: учебниками, справочниками, дополнительной литературой, а также проработка конспектов лекций;

	- выполнение домашних заданий и расчетов; - работа над темами для самостоятельного изучения; - участие в работе студенческих научных конференций, олимпиад; - подготовка к промежуточной аттестации.
Подготовка к промежуточной аттестации	Готовиться к промежуточной аттестации следует систематически, в течение всего семестра. Интенсивная подготовка должна начаться не позднее, чем за месяц-полтора до промежуточной аттестации.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Перечень вносимых изменений	Дата внесения изменений	Подпись заведующего кафедрой, ответственной за реализацию ОПОП