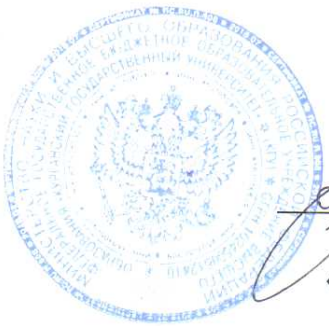
Министерство науки и высшего образования Российской Федерации

федеральное государственное бюджетное образовательное учреждение
высшего образования
«Курганский государственный университет»
(КГУ)

Кафедра «Безопасность информационных и автоматизированных систем»



УТВЕРЖДАЮ:
Первый проректор

_____ / Т.Р. Змызгова/

«31» августа 2022 г.

Рабочая программа учебной дисциплины

ВВЕДЕНИЕ В ОБРАБОТКУ ЕСТЕСТВЕННОГО ЯЗЫКА

образовательной программы высшего образования –
программы специалитета

10.05.03 – Информационная безопасность автоматизированных систем

Специализация:
Специализация №5 «Безопасность открытых информационных систем»

Форма обучения: **очная**

Курган 2022

Рабочая программа дисциплины «Введение в обработку естественного языка» составлена в соответствии с учебным планом по программе специалитета «Информационная безопасность автоматизированных систем» (Безопасность открытых информационных систем), утвержденным:
- для очной формы обучения «30» августа 2022 года.

Рабочая программа дисциплины одобрена на заседании кафедры «Безопасность информационных и автоматизированных систем» «29» августа 2022 года, протокол № 1.

Заведующий кафедрой «Безопасность информационных и автоматизированных систем» _____ Д.И. Дик

Согласовано:

Заведующий кафедрой «Безопасность информационных и автоматизированных систем» _____ Д.И. Дик

Специалист по учебно-методической работе учебно-методического отдела _____ Г.В. Казанкова

Начальник управления образовательной деятельности _____ И.В. Григоренко

1. ОБЪЕМ ДИСЦИПЛИНЫ

Всего: 4 зачетных единицы трудоемкости (144 академических часа)

Очная форма обучения

Вид учебной работы	На всю дисциплину	Семестр
		11
Аудиторные занятия (контактная работа с преподавателем), всего часов в том числе:	**64**	**64**
Лекции	32	32
Лабораторные работы	32	32
Самостоятельная работа, всего часов в том числе:	**80**	**80**
Подготовка к зачету	18	18
Другие виды самостоятельной работы (самостоятельное изучение тем (разделов) дисциплины)	62	62
Вид промежуточной аттестации	**Зачет**	**Зачет**
Общая трудоемкость дисциплины и трудоемкость по семестрам, часов	**144**	**144**

2 МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина «Введение в обработку естественного языка» является дисциплиной по выбору Блока 1 и относится к части, формируемой участниками образовательных отношений.

Изучение дисциплины базируется на результатах обучения, сформированных при изучении следующих дисциплин:

- Алгебра и геометрия;
- Основы программирования;
- Технологии и методы программирования.

Дисциплина «Введение в обработку естественного языка» является одной из заключительных дисциплин подготовки специалистов, изучается в последнем семестре, поэтому знания, умения и навыки, полученные в ходе изучения дисциплины необходимы для прохождения производственной практики и успешного написания выпускной квалификационной работы.

3 ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Целью изучения дисциплины является знакомство обучающихся с современными методами обработки естественного языка, основанными на глубоких нейронных сетях и машинном обучении, и приобретение комплекса практических навыков в области обработки естественного языка.

Задачами дисциплины являются:

- ознакомление с теоретическими принципами построения систем обработки естественного языка;
- формирование умений по решению типовых задач в области обработки естественного языка;
- приобретение обучающимися навыков использования технологий обработки естественного языка.

Компетенции, формируемые в результате освоения дисциплины:

- способен обрабатывать и анализировать научно-техническую информации и результаты исследований (ПК-1);
- способен разрабатывать и анализировать проектные решения по обеспечению безопасности автоматизированных систем (ПК-5);
- способен оценивать эффективность систем защиты информации, функционирующих в открытых информационных системах (ПК-8);

В результате изучения дисциплины обучающийся должен:

знать:

- принципы построения систем обработки естественного языка (ПК-1);
- принципы разработки программных средств для решения профессиональных задач (ПК-8);

уметь:

– применять средства обработки естественного языка с использованием технологий искусственного интеллекта (для ПК-5);

владеть:

– инструментами искусственного интеллекта для обработки естественного языка (для ПК-5).

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Учебно-тематический план

Рубеж	Номер темы	Наименование темы	Количество часов контактной работы с преподавателем	
			Лекции	Лабораторные работы
Рубеж 1	1	Теоретические аспекты обработки естественного языка.	4	-
	2	Предварительная обработка текста.	4	4
	3	Векторизация текста.	4	4
	4	Машинное обучение для обработки текстов.	4	6
		1-ый рубежный контроль (Тестирование)		2
Рубеж 2	5	Нейронные сети в решении задач текстовой обработки.	4	4
	6	Языковая модель.	4	-
	7	Поиск именованных сущностей.	4	4
	8	Механизм внимания. Трансформер.	4	6
		2-ый рубежный контроль (Тестирование)		2
		Всего:	**32**	**32**

4.2 Содержание лекционных занятий

Тема №1. Теоретические аспекты обработки естественного языка.

Синтаксический, морфологический, семантический и графематический анализ, омонимия, задачи лингвистического анализа.

Тема №2. Предварительная обработка текста.

Очистка текста, токенизация, стемминг, лемматизация, удаление стоп-слов, фильтрация наиболее частотных и наименее частотных слов.

Тема №3. Векторизация текста.

Построение словаря, мешок слов, TF-IDF, word2vec, fasttext, LDA, LSI, GloVe.

Тема №4. Машинное обучение для обработки текстов.

Решение задач классификации и определения тональности методами классического машинного обучения на основе векторных моделей.

Тема №5. Нейронные сети в решении задач текстовой обработки.

Архитектуры нейронных сетей для обработки текстов: рекуррентные (LSTM, GRU), одномерные сверточные. Применение нейронных сетей для обработки текстов.

Тема №6. Языковая модель.

Языковая модель и дистрибутивная семантика. Обучение векторной модели. Задача генерации текста. Различные подходы к генерации текста.

Тема №7. Поиск именованных сущностей..

Задача поиска именованных сущностей в тексте. Применение нейронных сетей для поиска именованных сущностей.

Тема №8. Механизм внимания. Трансформер.

Механизм внимания в нейронных сетях. Применение механизма внимания для обработки текста. Нейронные сети с архитектурой Transformer. Нейронные сети BERT, GPT. Перенос обучения.

4.3 Лабораторные работы

Номер темы	Наименование темы	Наименование лабораторной работы	Норматив времени, час.
2	Предварительная обработка текста	Предварительная обработка текста для анализа..	4
3	Векторизация текста.	Векторизация текста.	4
4	Машинное обучение для обработки текстов.	Классификация текста с использованием классических методов машинного обучения.	6
	1-ый рубежный контроль	***Тестирование***	***2***
5	Нейронные сети в решении задач текстовой обработки.	Классификация текста с использованием глубоких нейронных сетей.	4
7	Поиск именованных сущностей.	Поиск именованных объектов в тексте	4
8	Механизм внимания. Трансформер.	Создание чат-бота.	6
	2-ой рубежный контроль	***Тестирование***	***2***
		Итого:	**32**

5 МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Лекционный курс базируется на пассивном методе обучения, реализующем традиционную объяснительно-иллюстративную образовательную технологию, в рамках которой магистры выступают в роли слушателей, воспринимающих учебный материал и участвующих в дискуссиях и экспресс-опросах.

При прослушивании лекций рекомендуется в конспекте отмечать все важные моменты, на которых заостряет внимание преподаватель, в частности те, которые направлены на качественное выполнение соответствующей лабораторной работе.

Залогом качественного выполнения лабораторных работ является самостоятельная подготовка к ним накануне путем повторения материалов лекций. Преподавателем запланировано применение на лабораторных работах разбор конкретных ситуаций.

Для текущего контроля успеваемости по очной форме обучения преподавателем используется балльно-рейтинговая система контроля и оценки академической активности. Поэтому настоятельно рекомендуется тщательно прорабатывать материал дисциплины при самостоятельной работе, участвовать во всех формах обсуждения и взаимодействия, как на лекциях, так лабораторных работах в целях лучшего освоения материала и получения высокой оценки по результатам освоения дисциплины.

Выполнение самостоятельной работы подразумевает самостоятельное изучение разделов дисциплины, подготовку к лабораторным работам, к рубежным контролям (для очной формы обучения) и подготовку к зачету.

Рекомендуемая трудоемкость самостоятельной работы представлена в таблице:

Рекомендуемый режим самостоятельной работы

Наименование вида самостоятельной работы	Рекомендуемая трудоемкость, акад. час.
Самостоятельное изучение тем раздела:	**46**
Теоретические аспекты обработки естественного языка.	5
Предварительная обработка текста.	5
Векторизация текста.	5
Машинное обучение для обработки текстов.	7
Нейронные сети в решении задач текстовой обработки.	7
Языковая модель.	5
Поиск именованных сущностей.	5
Механизм внимания. Трансформер.	7
Подготовка к лабораторным работам (по 2 часа на каждую работу)	12
Подготовка к рубежным контролям (по 2 часа на каждый рубежный контроль)	4
Подготовка к зачету	18
Всего:	**80**

6 ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

6.1 Перечень оценочных средств

1. Балльно-рейтинговая система контроля и оценки академической активности обучающихся в КГУ (для очной формы обучения)
2. Отчеты обучающихся в по лабораторным работам.
3. Банк тестовых заданий к рубежным контролям № 1, № 2 (для очной формы обучения).
4. Вопросы к зачету.

6.2 Система балльно-рейтинговой оценки работы обучающихся по дисциплине (для очной формы обучения)

№	Наименование	Содержание					
1	Распределение баллов за семестры по видам учебной работы, сроки сдачи учебной работы *(доводятся до сведения обучающихся на первом учебном занятии)*	Распределение баллов					
		Вид учебной работы:	Посещение лекций	Выполнение лабораторных работ	Рубежный контроль №1	Рубежный контроль №2	Зачет
		Балльная оценка:	$1_б$ х 16 =$16_б$	$5_б$ х 6 = $30_б$	12	12	30
2	Критерий пересчета баллов в традиционную оценку по итогам работы в семестре и зачете	60 и менее баллов – неудовлетворительно; не зачтено; 61…73 – удовлетворительно; зачтено; 74… 90 – хорошо; 91…100 – отлично					
3	Критерии допуска к промежуточной аттестации, возможности получения автоматического зачета по дисциплине, возможность получения бонусных баллов	Для допуска к промежуточной аттестации (зачету) обучающийся должен набрать по итогам текущего и рубежного контроля не менее 50 баллов и должен выполнить все лабораторные работы. Для получения зачета «автоматически» обучающемуся необходимо набрать 61 балл По согласованию с преподавателем обучающемуся могут быть добавлены дополнительные (бонусные) баллы за активность на лабораторных работах, активное участие в научной и методической работе, оригинальность принятых решений в ходе выполнения лабораторных работ, за участие в значимых учебных и внеучебных мероприятиях кафедры.					

4	Формы и виды учебной работы для неуспевающих (восстановившихся на курсе обучения) обучающихся для получения недостающих баллов в конце семестра	В случае если к промежуточной аттестации (зачету) набрана сумма менее 50 баллов, обучающемуся необходимо набрать недостающее количество баллов за счет выполнения дополнительных заданий, до конца последней (зачетной) недели семестра. При этом необходимо проработать материал всех пропущенных лабораторных работ. Формы дополнительных заданий (назначаются преподавателем): - выполнение и защита пропущенной лабораторной работы (при невозможности дополнительного проведения лабораторной работы преподаватель устанавливает форму дополнительного задания по тематике пропущенной работы самостоятельно) – до 4 баллов. Ликвидация академических задолженностей, возникших из-за разности в учебных планах при переводе или восстановлении, проводится путем выполнения дополнительных заданий, форма и объем которых определяется преподавателем.

6.3 Процедура оценивания результатов освоения дисциплины

Мероприятия текущего контроля проводятся на аудиторных занятиях в соответствии с расписанием.

Основной вид текущего контроля результатов освоения дисциплины - защита отчетов по выполненным лабораторным работам.

В процессе защиты отчетов оценивается уровень понимания обучающимися методики проведения работы, полнота и качество выполнения заданий, а также обоснованность выводов, сделанных обучающимся по результатам выполнения заданий.

Рубежные контроли проводятся в форме письменного тестирования.

Перед проведением каждого рубежного контроля преподаватель прорабатывает с обучающимися основной материал соответствующих разделов дисциплины. Варианты тестовых заданий состоят для 1 и 2 рубежного контроля из 12 вопросов. На каждое тестирование при рубежном контроле обучающемуся отводится 2 академических часа.

Баллы обучающемуся выставляются в зависимости от числа правильно выбранных ответов. Итоговая оценка по тесту формируется путем суммирования набранных баллов и отнесения их к общему количеству вопросов в задании.

Зачет проводится в форме устного ответа на 2 вопроса. Билет состоит из 2 вопросов. Перечень вопросов преподаватель выдает заранее. Время, отводимое обучающемуся на подготовку вопросов, составляет 1 академический час. Каждый вопрос оценивается в 15 баллов.

Результаты текущего контроля успеваемости и зачета заносятся преподавателем в зачетную ведомость, которая сдается в организационный отдел института в день зачета, а также выставляются в зачетную книжку обучающегося.

6.4 Примеры оценочных средств для рубежных контролей и зачета

Примерные тестовые задания для рубежного контроля №1

1) Токенизация – это…
а) выделение заголовков, абзацев и т.п.;
б) разбиение текста на слова, удаление знаков препинания;
б) удаление высокочастотных служебных слов (предлогов, союзов и т.п.);
в) приведение слов к нормальной (например, словарной) форме;
г) определение частоты встречаемости слов в документе.

2) Лемматизация – это…
а) выделение заголовков, абзацев и т.п.;
б) разбиение текста на слова, удаление знаков препинания;
б) удаление высокочастотных служебных слов (предлогов, союзов и т.п.);
в) приведение слов к нормальной (например, словарной) форме;
г) определение частоты встречаемости слов в документе.

3) Представление текста в виде мешка слов можно описать как
а) текст представляется как набор содержащихся в нем слов, без учета грамматики и даже порядка слов, но с сохранением множественности.
б) формируется векторное представление, основанное на контекстной близости слов;
в) минимизирует разницу между произведением векторов слов и логарифмом вероятности их совместного появления;
г) каждое слово представляется композицией нескольких последовательностей символов определённой длины.

Примерные тестовые задания для рубежного контроля №2

1) Что из ниже перечисленного не относится к метрикам оценки качества перевода:
а) BLEU;
б) ROUGE;
в) METEOR;
г) COCO.

2) Состояние ячейки LSTM сети контролируется (укажите лишнее):
а) входным гейтом;
б) выходным гейтом;
в) забывающим гейтом;
г) вспоминающим гейтом;
д) промежуточным гейтом.

3) Выберите методы улучшения прогноза LSTM сети:

а) жадная выборка;
б) ленивая выборка;
в) лучевой поиск;
г) направленный поиск;
д) использование векторов слов вместо унитарного кодирования;
е) использование двунаправленных LSTM;
ж) нет верного ответа.

Примерный перечень вопросов к зачету

1. Теоретические аспекты обработки естественного языка.
2. Предварительная обработка текста. Очистка текста. Удаление стоп-слов/наиболее и наименее часто встречающихся слов.
3. Токенизация, стемминг, лемматизация текста.
4. Методы векторизации текста: создание словаря, набора слов.
5. Методы векторизации текста: TF-IDF.
6. Методы векторизации текста: word2vec.
7. Методы векторизации текста: fasttext
8. Методы векторизации текста: Перчатка.
9. Классические методы машинного обучения для решения задач классификации текстов.
10. Классические методы машинного обучения для решения задачи определения тональности текста.
11. Архитектуры нейронных сетей для обработки текста: LSTM.
12. Архитектуры нейронных сетей для обработки текста: GRU.
13. Архитектуры нейронных сетей для обработки текста: одномерные сверточные сети.
14. Классификация текста с использованием нейронных сетей.
15. Определение тональности текста с помощью нейронных сетей.
16. Языковая модель.
17. Обучение языковой модели.
18. Основные подходы к генерации текста.
19. Задача поиска именованных сущностей в тексте.
20. Использование нейронных сетей для поиска именованных объектов.
21. Механизм внимания в нейронных сетях.
22. Применение механизма внимания для обработки текста.
23. Архитектура нейронной сети трансформер.
24. Нейронные сети для обработки текста BERT.
25. Нейронные сети для обработки текста GPT.
26. Классификация текста с использованием сетей с архитектурой Transformer.

6.5. Фонд оценочных средств

Полный банк заданий для текущего, рубежных контролей и промежуточной аттестации по дисциплине, показатели, критерии, шкалы оценивания компетенций, методические материалы, определяющие процедуры оценивания образовательных результатов, приведены в учебно-методическом комплексе дисциплины.

7. ОСНОВНАЯ И ДОПОЛНИТЕЛЬНАЯ УЧЕБНАЯ ЛИТЕРАТУРА

7.1. Основная учебная литература

1. Риз, Р. Обработка естественного языка на Java [Электронный ресурс] / Р. Риз; пер. с англ. А. В. Снастина. - Москва : ДМК Пресс, 2016. - 264 с. - Доступ ЭБС «Консультант студента».

2. Йылдырым, С. Осваиваем архитектуру Transformer. Разработка современных моделей с помощью передовых методов обработки естественного языка [Электронный ресурс] / С. Йылдырым, М. Асгари-Ченаглу; пер. с анг. В. С. Яценкова. - Москва : ДМК Пресс, 2022. - 320 с. - Доступ ЭБС «Консультант студента».

7.2 Дополнительная учебная литература

1. Jurafsky, D. Speech and Language Processing [Electronic resourse] / Jurafsky D., J.H. Martin. – 3rd ed. draft. – Access mode: https://web.stanford.edu/~jurafsky/slp3/, free.

2. Браславский П.И. Введение в обработку естественного языка / П.И. Браславский. – Режим доступа: https://stepik.org/course/1233/, свободный.

3. Суворов, Р. Нейронные сети и обработка текста / Р. Суворов, А. Янина, А. Сильвестров, Н. Капырин. – Режим доступа: https://stepik.org/course/54098, свободный.

7.3 Методическая литература

1. Методические указания по выполнению лабораторных работ по дисциплине «Введение в обработку естественного языка».

8. РЕСУРСЫ СЕТИ «ИНТЕРНЕТ», НЕОБХОДИМЫЕ ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1. Сайт дистанционного обучения в НОУ (Национальный Открытый Университет) «ИНТУИТ» содержит бесплатные курсы, программы повышения квалификации и профессиональной переподготовки, интересные доклады и другую полезную информацию http://www.intuit.ru.
2. Федеральный портал «Российское образование» http://www.edu.ru/

3. Портал «Информационно-коммуникационные технологии в образовании» http://www.ict.edu.ru.
4. Федеральный портал ЭБС «Лань» - https://e.lanbook.com/;
5. ЭБС «Znanium» - https://znanium.com/;
6. ЭБС «Консультант студента» - https://www.studentlibrary.ru;
7. Электронная библиотека КГУ - http://dspace.kgsu.ru/xmlui/

9. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ И ИНФОРМАЦИОННЫЕ СПРАВОЧНЫЕ СИСТЕМЫ

Аналитическая платформа «KNIME» (лицензия на свободное ПО: GNU General Public License v3).

Библиотека глубокого обучения Tensorflow (лицензия на свободное ПО: Apache License Version 2.0).

Библиотека глубокого обучения Keras (лицензия на свободное ПО: Apache License Version 2.0).

Язык программирования Python (лицензия на свободное ПО: Python Software Foundation License).

При чтении лекций используются слайдовые презентации.

Минимальные требования к программному обеспечению компьютера, используемого при показе слайдовых презентаций: офисный пакет LibreOffice (лицензия Mozilla Public License Version 2.0).

10. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Материально-техническое обеспечение дисциплины включает в себя учебные лаборатории и классы, оснащенные современными компьютерами (все – в стандартной комплектации для лабораторных работ и самостоятельной работы), объединенными локальными вычислительными сетями с выходом в Интернет, мультимедийное оборудование (переносной персональный компьютер, мультимедийный проектор, мультимедийный экран).

11. ДЛЯ ОБУЧАЮЩИХСЯ С ИСПОЛЬЗОВАНИЕМ ДИСТАНЦИОННЫХ ОБРАЗОВАТЕЛЬНЫХ ТЕХНОЛОГИЙ

При использовании электронного обучения и дистанционных образовательных технологий (далее ЭО и ДОТ) занятия полностью или частично проводятся в режиме онлайн. Объем дисциплины и распределение нагрузки по видам работ соответствует п. 4.1. Распределение баллов соответствует п. 6.2 либо может быть изменено в соответствии с решением кафедры, в случае перехода на ЭО и ДОТ в процессе обучения. Решение кафедры об используемых технологиях и системе оценивания достижений обучающихся принимается с учетом мнения ведущего преподавателя и доводится до обучающихся.

Аннотация
рабочей программы учебной дисциплины
«Введение в обработку естественного языка»
образовательной программы высшего образования –
программы специалитета
10.05.03 – Информационная безопасность автоматизированных систем

Специализация №5 «Безопасность открытых информационных систем»

Формы обучения: **очная**

Трудоемкость дисциплины: 4 ЗЕ (144 академических часа)
Семестры: 11-й
Форма промежуточной аттестации: зачет

Содержание дисциплины

Теоретические аспекты обработки естественного языка. Предварительная обработка текста. Векторизация текста. Машинное обучение для обработки текстов. Нейронные сети в решении задач текстовой обработки. Языковая модель. Поиск именованных сущностей. Механизм внимания. Трансформер.