

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«Луганский государственный университет
имени Владимира Даля»

(ФГБОУ ВО «ЛГУ им. В. Даля»)

Краснодонский факультет инженерии и менеджмента (филиал)

Кафедра информационных технологий и транспорта



УТВЕРЖДАЮ:

Директор

Панайотов К.К.

«21» апреля 2023 года

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

По дисциплине Технологии работы с естественным языком

(название дисциплины по учебному плану)

По направлению подготовки 09.04.01 – Информатика и вычислительная техника

(код, название без кавычек)

Профиль подготовки Интеллектуальные системы в производственно-транспортных комплексах

Краснодон 2023

Лист согласования РПУД

Рабочая программа учебной дисциплины «Технологии работы с естественным языком» по направлению подготовки 09.04.01 – Информатика и вычислительная техника. – 15 с.

Рабочая программа учебной дисциплины «Технологии работы с естественным языком» составлена с учетом Федерального государственного образовательного стандарта высшего образования по направлению подготовки 09.04.01 – Информатика и вычислительная техника, утвержденного приказом Министерства образования и науки Российской Федерации от 19 сентября 2017 года № 918.

СОСТАВИТЕЛЬ (СОСТАВИТЕЛИ):

к.т.н., доц. Бихдрикер А.С.

(ученая степень, ученое звание, должность фамилия, инициалы)

Рабочая программа дисциплины утверждена на заседании кафедры информационных технологий и транспорта «_15_» __марта__ 2023 г., протокол № _7_

Заведующий кафедрой



Бихдрикер А.С.

Рекомендована на заседании учебно-методической комиссии факультета «20» марта 2023 г., протокол № 8.

Председатель учебно-методической
комиссии факультета



Замота О.Н.

Структура и содержание дисциплины

1. Цели и задачи дисциплины, ее место в учебном процессе

Целями освоения дисциплины «Технологии работы с естественным языком» изучение передовых методов, моделей, средств и технологий компьютерной обработки текстов на естественных языках, дать умение представлять в алгоритмическом виде процессы анализа и синтеза текста.

Задачами освоения дисциплины «Технологии работы с естественным языком» являются:

дать студентам представление о современных методах компьютерной обработки текстов на естественных языках;

ознакомление студентов с состоянием и перспективами развития технического обеспечения автоматизированных систем компьютерной обработки текстов на естественных языках.

2. Место дисциплины в структуре ООП ВО

Курс входит в часть, формируемую участниками образовательных отношений учебного плана.

Основывается на базе дисциплин: «Системы обработки изображений и распознавания образов», «Методы и алгоритмы принятия решений».

Является основой для изучения следующих дисциплин: выполнения магистерской работы.

3. Требования к результатам освоения содержания дисциплины

Код и наименование компетенции	Индикаторы достижений компетенции (по реализуемой дисциплине)	Перечень планируемых результатов
ПК-2 Способен разрабатывать и применять методы и алгоритмы машинного обучения для решения задач	ПК-2.2 Знает классы методов и алгоритмов машинного обучения применением новых методов и алгоритмов машинного обучения	Знать: процедуры выявления, формирования и согласования требований к результатам аналитических работ с применением технологий обработки естественного языка;; Уметь: подготавливать данные — для проведения аналитических работ по исследованию текстовых массивов данных методами обработки естественного языка; Владеть: методами планирования и организации аналитических работ с использованием технологий обработки естественного языка;;

4. Структура и содержание дисциплины

4.1. Объем учебной дисциплины и виды учебной работы

Вид учебной работы	Объем часов (зач. ед.)	
	Очная форма	Заочная форма
Общая учебная нагрузка (всего)	180 (5 зач. ед)	180 (5 зач. ед)
Обязательная аудиторная учебная нагрузка (всего) в том числе:	60	12
Лекции	24	6
Семинарские занятия	-	-
Практические занятия	36	6
Лабораторные работы	-	-
Курсовая работа (курсовой проект)	-	-
Другие формы и методы организации образовательного процесса (<i>расчетно-графические работы, групповые дискуссии, ролевые игры, тренинг, компьютерные симуляции, интерактивные лекции, семинары, анализ деловых ситуаций и т.п.</i>)	-	-
Самостоятельная работа студента (всего)	120	168
Форма аттестации	экзамен	экзамен

4.2. Содержание разделов дисциплины

ТЕМА 1. ВВЕДЕНИЕ, ИСТОРИЯ РАЗВИТИЯ ДИСЦИПЛИНЫ, РЕШАЕМЫЕ ЗАДАЧИ, ПОДХОДЫ, МЕТОДЫ И ИНСТРУМЕНТЫ

ТЕМА 2. ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ТЕКСТОВЫХ ДАННЫХ

ТЕМА 3. РАЗМЕТКА ПОСЛЕДОВАТЕЛЬНОСТЕЙ

ТЕМА 4. ВЕКТОРИЗАЦИЯ ТЕКСТОВЫХ ДАННЫХ

ТЕМА 5. ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

ТЕМА 6. ЯЗЫКОВЫЕ МОДЕЛИ

ТЕМА 7. АНАЛИЗ ТОНАЛЬНОСТИ

ТЕМА 8. МОДЕЛИ SEQUENCE-TO-SEQUENCE И МЕХАНИЗМ ВНИМАНИЯ

4.3. Лекции

№ п/п	Название темы	Объем часов	
		Очная форма	Заочная форма
1	Введение, история развития дисциплины, решаемые задачи, подходы, методы и инструменты	2	0,5
2	Предварительная обработка текстовых данных	2	0,5
3	Разметка последовательностей	2	0,5
4	Векторизация текстовых данных	4	1
5	Тематическое моделирование	4	1
6	Языковые модели	2	0,5
7	Анализ тональности	4	1
8	Модели sequence-to-sequence и механизм внимания	4	1
Итого:		24	6

4.4. Практические (семинарские) занятия

№ п/п	Название темы	Объем часов	
		Очная форма	Заочная форма
1	Введение, история развития дисциплины, решаемые задачи, подходы, методы и инструменты	2	0,5
2	Предварительная обработка текстовых данных	2	0,5
3	Разметка последовательностей	6	0,5
4	Векторизация текстовых данных	6	1
5	Тематическое моделирование	6	1
6	Языковые модели	6	0,5
7	Анализ тональности	4	1
8	Модели sequence-to-sequence и механизм внимания	4	1
Итого:		36	6

4.4. Лабораторные работы

Лабораторные работы рабочим учебным планом не предусмотрены.

4.6. Самостоятельная работа студентов

№ п/п	Название темы	Вид СРС	Объем часов	
			Очная форма	Заочная форма
1	Введение, история развития дисциплины, решаемые задачи, подходы, методы и инструменты	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	10	16
2	Предварительная обработка текстовых данных	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	10	16
3	Разметка последовательностей	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	10	16
4	Векторизация текстовых данных	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	10	16
5	Тематическое моделирование	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	11	17
6	Языковые модели	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	11	7
7	Анализ тональности	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	11	17
8	Модели sequence-to-sequence и механизм внимания	Подготовка к практическим занятиям, к текущему и промежуточному контролю знаний и умений.	11	17
9	Подготовка к экзамену	Повтор теоретического материала. Прохождение теста для самопроверки.	36	36
Итого:			120	168

4.7. Курсовые работы/проекты.

Курсовая работа рабочим учебным планом не предусмотрена.

5. Образовательные технологии

Преподавание дисциплины ведется с применением следующих видов образовательных технологий:

традиционные объяснительно-иллюстративные технологии, которые обеспечивают доступность учебного материала для большинства студентов, системность, отработанность организационных форм и привычных методов, относительно малые затраты времени;

технологии проблемного обучения, направленные на развитие познавательной активности, творческой самостоятельности студентов и предполагающие последовательное и целенаправленное выдвижение перед студентом познавательных задач, разрешение которых позволяет студентам активно усваивать знания (используются поисковые методы; постановка познавательных задач);

технологии развивающего обучения, позволяющие ориентировать учебный процесс на потенциальные возможности студентов, их реализацию и развитие;

технологии концентрированного обучения, суть которых состоит в создании максимально близкой к естественным психологическим особенностям человеческого восприятия структуры учебного процесса и которые дают возможность глубокого и системного изучения содержания учебных дисциплин за счет объединения занятий в тематические блоки;

технологии модульного обучения, дающие возможность обеспечения гибкости процесса обучения, адаптации его к индивидуальным потребностям и особенностям обучающихся (применяются, как правило, при самостоятельном обучении студентов по индивидуальному учебному плану);

технологии дифференцированного обучения, обеспечивающие возможность создания оптимальных условий для развития интересов и способностей студентов, в том числе и студентов с особыми образовательными потребностями, что позволяет реализовать в культурно-образовательном пространстве университета идею создания равных возможностей для получения образования

технологии активного (контекстного) обучения, с помощью которых осуществляется моделирование предметного, проблемного и социального содержания будущей профессиональной деятельности студентов (используются активные и интерактивные методы обучения) и т.д.

Максимальная эффективность педагогического процесса достигается путем конструирования оптимального комплекса педагогических технологий и (или) их элементов на личностно-ориентированной, деятельностной, диалогической основе и использования необходимых современных средств обучения.

6. Учебно-методическое и информационное обеспечение дисциплины:

а) основная литература:

1. Хобсон Лейн, Ханнес Хапке, Коул Ховард Обработка естественного языка в действии. — СПб.: Питер, 2020. — 576 с.: ил. — (Серия «Для профессионалов»).
2. ГрасД. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2021. - 416 с.: ил.
3. Масленникова О.Е., Основы искусственного интеллекта : учеб. пособие / О.Е. Масленникова, И.В. Гаврилова - М. : ФЛИНТА, 2019. - 283 с. - ISBN 978-5-9765-1602-1 - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : <http://www.studentlibrary.ru/book/ISBN9785976516021.html>

б) дополнительная литература:

1. Пойнтер Ян. Программируем с PyTorch: Создание приложений глубокого обучения. — СПб.: Питер, 2020. — 256 с.: ил.

2. Гифт Ной Прагматичный ИИ. Машинное обучение и облачные технологии. - СПб.: Питер, 2019. - 304 с.: ил.
3. Ashish Bansal. Advanced Natural Language Processing with TensorFlow 2 – BIRMINGHAM – MUMBAI: Packt Publishing, 2021. – 381 с.: ил.

в) Интернет-ресурсы:

1. Министерство образования и науки Российской Федерации – <http://минобрнауки.рф/>
2. Федеральная служба по надзору в сфере образования и науки – <http://obrnadzor.gov.ru/>
3. Министерство образования и науки Луганской Народной Республики – <https://minobr.su>
4. Министерство промышленности и торговли Луганской Народной Республики – <https://www.minpromlnr.su/main.php/>
5. Министерство экономического развития Луганской Народной Республики – <https://merlnr.su/>
6. Министерство финансов Луганской Народной Республики – <https://minfinlnr.su/>
7. Народный совет Луганской Народной Республики – <https://nslnr.su>
8. Государственный комитет статистики Луганской Народной Республики – <https://www.gkslnr.su/>
9. Портал Федеральных государственных образовательных стандартов высшего образования – <http://fgosvo.ru/>
10. Федеральный портал «Российское образование» – <http://www.edu.ru/>
11. Информационная система «Единое окно доступа к образовательным ресурсам» – <http://window.edu.ru/>
12. Федеральный центр информационно-образовательных ресурсов – <http://fcior.edu.ru/>
13. Статистические сборники ВШЭ - <https://www.hse.ru/org/hse/primarydata/>
14. OpenOffice.org: Теория и практика - <https://www.altlinux.org/Books/Openoffice>
15. Базовый курс по OpenOffice - <https://4creates.com/training/49-bazovyy-kurs-po-openoffice.html>

Электронные библиотечные системы и ресурсы

16. Электронно-библиотечная система «Консультант студента» – <http://www.studentlibrary.ru/>
17. Электронно-библиотечная система «StudMed.ru» – <https://www.studmed.ru/>

Информационный ресурс библиотеки образовательной организации

18. Научная библиотека имени А. Н. Коняева – <http://biblio.dahluniver.ru/>

7. Материально-техническое и программное обеспечение дисциплины

Для проведения лекционных занятий требуется аудитория, оборудованная мультимедийным проектором с экраном. Для проведения практических занятий необходим компьютерный класс и презентационная техника (проектор, экран, компьютер/ноутбук)

Программное обеспечение:

Функциональное назначение	Бесплатное программное обеспечение	Ссылки
Офисный пакет	OpenOffice 4.3.7	https://www.openoffice.org/
Операционная система	UBUNTU 19.04	https://ubuntu.com/
Браузер	Firefox Mozilla	http://www.mozilla.org/ru/firefox/fx

Графический редактор	GIMP (GNU Image Manipulation Program)	http://www.gimp.org/ http://gimp.ru/viewpage.php?page_id=8 http://ru.wikipedia.org/wiki/GIMP
Редактор PDF	Adobe Acrobat Reader	https://get.adobe.com/ru/reader/
Аудиоплеер	VLC	http://www.videolan.org/vlc/

8. Оценочные средства по дисциплине

Паспорт фонда оценочных средств по учебной дисциплине «Технологии работы с естественным языком»

Перечень компетенций (элементов компетенций), формируемых в результате освоения учебной дисциплины (модуля) или практики

№ п/п	Код контролируемой компетенции	Формулировка контролируемой компетенции	Индикаторы достижений компетенции (по реализуемой дисциплине)	Контролируемые темы учебной дисциплины, практики	Этапы формирования (семестр изучения)
1	ПК-2	Способен разрабатывать и применять методы и алгоритмы машинного обучения для решения задач	ПК-2.2 Знает классы методов и алгоритмов машинного обучения применением новых методов и алгоритмов машинного обучения	Тема 1. Введение, история развития дисциплины, решаемые задачи, подходы, методы и инструменты	4
				Тема 2. Предварительная обработка текстовых данных	4
				Тема 3. Разметка последовательностей	4
				Тема 4. Векторизация текстовых данных	4
				Тема 5. Тематическое моделирование	4
				Тема 6. Языковые модели	4
				Тема 7. Анализ тональности	4
				Тема 8. Модели sequence-to-sequence и механизм внимания	4

Показатели и критерии оценивания компетенций, описание шкал оценивания

№ п/п	Код контролируемой компетенции	Индикаторы достижений компетенции (по реализуемой дисциплине)	Перечень планируемых результатов	Контролируемые темы учебной дисциплины	Наименование оценочного средства
1	ПК-2	ПК-2.2	Знать: процедуры выявления, формирования и согласования требований к результатам аналитических работ с применением технологий обработки естественного языка;; Уметь: подготавливать данные — для проведения аналитических работ по исследованию текстовых массивов данных методами обработки естественного языка; Владеть: методами планирования и организации аналитических работ с использованием технологий обработки естественного языка;;	Тема 1. Тема 2. Тема 3. Тема 4. Тема 5. Тема 6. Тема 7 Тема 8	Устный опрос, контрольная работа (по вариантам), тесты

Перечень вопросов (для проведения собеседования (устный или письменный опрос))

Перечень примерных вопросов для экзамена:

- 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 стоп-слов/наиболее и наименее частых слов.
- 5 Токенизация, вывод, лемматизация текста.
- 6 Методы векторизации текста: построение словаря, мешок слов.
- 7 Методы векторизации текста: TF-IDF.
- 8 Методы векторизации текста: word2vec.
- 9 Методы векторизации текста: fasttext
- 10 Методы векторизации текста: GloVe.
- 11 Классические методы машинного обучения для решения задач классификации текста.
- 12 Классические методы машинного обучения для решения задачи определения тональности текста.
- 13 Архитектуры нейронных сетей для обработки текста: LSTM.
- 14 Архитектуры нейронных сетей для обработки текста: GRU.
- 15 Архитектуры нейронных сетей для обработки текста: одномерные сверточные сети.
- 16 Классификация текста с использованием нейронных сетей.
- 17 Определение тональности текста с помощью нейронных сетей.
- 18 Языковая модель.
- 19 Обучение языковой модели.
- 20 Основные подходы к генерации текста.
- 21 Задача поиска именованных сущностей в тексте.
- 22 Использование нейронных сетей для поиска именованных сущностей.
- 23 Механизм внимания в нейронных сетях.
- 24 Применение механизма внимания для обработки текста.
- 25 Архитектура трансформаторной нейронной сети.
- 26 Предварительно обученные нейронные сети для обработки текста BERT.
- 27 Предварительно обученные нейронные сети для обработки текста GPT.
- 28 Передача обучения задачам обработки текстов.
- 29 Классификация текста с использованием сетей с трансформаторной архитектурой.
- 30 Генерация текста с использованием сетей с трансформаторной архитектурой.
- 31 Поиск именованных объектов в тексте с использованием сетей с архитектурой трансформатора.

Критерии и шкала оценивания по оценочному средству собеседование (устный или письменный опрос)

Шкала оценивания (интервал баллов)	Критерий оценивания
5	собеседование (устный или письменный опрос) прошел на высоком уровне (студент в полном объеме осветил рассматриваемый вопрос, владеет профильным понятийным (категориальным) аппаратом и т.п.)
4	собеседование (устный или письменный опрос) прошел на среднем уровне (студент в целом осветил рассматриваемый вопрос, владеет профильным понятийным (категориальным) аппаратом и т.п.)
3	собеседование (устный или письменный опрос) на низком уровне (студент допустил существенные неточности, изложил материал с ошибками, не владеет в достаточной степени профильным категориальным аппаратом и т.п.)

2	собеседование (устный или письменный опрос) прошел на неудовлетворительном уровне или не представлен (студент не готов, не выполнил задание и т.п.)
---	---

ФОНДЫ ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Вопросы для подготовки к экзамену

1. Трудности автоматизации обработки естественного языка в интеллектуальных системах.
2. Этапы анализа предложений на естественном языке.
3. Синтаксически-ориентированный и семантически-ориентированный анализ ЕЯ.
4. Основные методы анализа ЕЯ: шаблоны.
5. Основные методы анализа ЕЯ: семантические грамматики.
6. Основные методы анализа ЕЯ: падежные фреймы.
7. Деревья анализа и свободно-контекстные грамматики.
10. Синтаксический анализ естественного языка.
8. Основные подходы к решению задач в интеллектуальных системах. Поиск в пространстве состояний.
9. Основные подходы к решению задач в интеллектуальных системах – логический вывод.
10. Основные подходы к решению задач в интеллектуальных системах – сопоставление с образцом и ассоциативный поиск.
11. Психолингвистический подход к анализу естественного языка.
12. Интеллектуальные системы, использующие естественный язык.
13. Задачи морфологического анализа, морфологический разбор, стемминг, лемматизация.
14. Понятия лексемы, словоформы, леммы, морфемы, псевдо-основы и псевдо-окончания.
15. Грамматические категории. Словоизменительная парадигма. Морфотактика. Структура данных морфологического словаря, лексикона.
16. Грамматические модели русского языка в контексте автоматической обработки.
17. Минимальное расстояние редактирования. Алгоритм подсчета расстояния Левенштейна.
18. Понятие статистической языковой модели. Области применения. N-граммы.

Критерии и шкала оценивания к промежуточной аттестации «экзамен»

Национальная шкала	Характеристика знания предмета и ответов
отлично (5)	Студент глубоко и в полном объеме владеет программным материалом. Грамотно, исчерпывающе и логично его излагает в устной или письменной форме. При этом знает рекомендованную литературу, проявляет творческий подход в ответах на вопросы и правильно обосновывает принятые решения, хорошо владеет умениями и навыками при выполнении практических задач.
хорошо (4)	Студент знает программный материал, грамотно и по сути излагает его в устной или письменной форме, допуская незначительные неточности в утверждениях, трактовках, определениях и категориях или незначительное количество ошибок. При этом владеет необходимыми умениями и навыками при выполнении практических задач.
удовлетворительно (3)	Студент знает только основной программный материал, допускает неточности, недостаточно четкие формулировки,

	непоследовательность в ответах, излагаемых в устной или письменной форме. При этом недостаточно владеет умениями и навыками при выполнении практических задач. Допускает до 30% ошибок в излагаемых ответах.
неудовлетворительно (2)	Студент не знает значительной части программного материала. При этом допускает принципиальные ошибки в доказательствах, в трактовке понятий и категорий, проявляет низкую культуру знаний, не владеет основными умениями и навыками при выполнении практических задач. Студент отказывается от ответов на дополнительные вопросы

9. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

При необходимости рабочая программа учебной дисциплины может быть адаптирована для обеспечения образовательного процесса инвалидов и лиц с ограниченными возможностями здоровья, в том числе с применением электронного обучения и дистанционных образовательных технологий.

Для этого требуется заявление студента (его законного представителя) и заключение психолого-медико-педагогической комиссии (ПМПК). В случае необходимости обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося), а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей:

- создание текстовой версии любого нетекстового контента для его возможного преобразования в альтернативные формы, удобные для различных пользователей;
- создание контента, который можно представить в различных видах без потери данных или структуры, предусмотреть возможность масштабирования текста и изображений без потери качества, предусмотреть доступность управления контентом с клавиатуры;
- создание возможностей для обучающихся воспринимать одну и ту же информацию из разных источников, например, так, чтобы лица с нарушениями слуха получали информацию визуально, с нарушениями зрения – аудиально;
- применение программных средств, обеспечивающих возможность освоения навыков и умений, формируемых дисциплиной (модулем), за счёт альтернативных способов, в том числе виртуальных лабораторий и симуляционных технологий;
- применение электронного обучения, дистанционных образовательных технологий для передачи информации, организации различных форм интерактивной контактной работы обучающегося с преподавателем, в том числе вебинаров, которые могут быть использованы для проведения виртуальных лекций с возможностью взаимодействия всех участников дистанционного обучения, проведения семинаров, выступления с докладами и защиты выполненных работ, проведения тренингов, организации коллективной работы;
- применение электронного обучения, дистанционных образовательных технологий для организации форм текущего и промежуточного контроля;
- увеличение продолжительности сдачи обучающимся инвалидом или лицом с ограниченными возможностями здоровья форм промежуточной аттестации по отношению к установленной продолжительности их сдачи:
 - продолжительность сдачи зачёта или экзамена, проводимого в письменной форме, – не более чем на 90 минут;
 - продолжительность подготовки обучающегося к ответу на зачёте или экзамене, проводимом в устной форме, – не более чем на 20 минут; – продолжительность выступления обучающегося при защите курсовой работы – не более чем на 15 минут.

Форма листа изменений и дополнений, внесенных в ФОС

Лист изменений и дополнений

№ п/п	Виды дополнений и изменений	Дата и номер протокола заседания кафедры (кафедр), на котором были рассмотрены и одобрены изменения и дополнения	Подпись (с расшифровкой) заведующего кафедрой (заведующих кафедрами)