

Министерство науки и высшего образования Российской Федерации
(МИНОБРНАУКИ РОССИИ)
Федеральное государственное бюджетное образовательное учреждение высшего образования
«ДОНБАССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ДонГТУ»)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

(наименование дисциплины)

(код, наименование специальности)

(магистерская программа)

магистр

(бакалавр/специалист/магистр)

очная

(очная, очно-заочная, заочная)

Алчевск, 2024

1 Цели и задачи изучения дисциплины

Цели дисциплины. Целью изучения дисциплины «Автоматическая обработка текста» является дать студентам теоретические сведения и практические навыки по созданию основных принципов и методов автоматической обработки текстов на естественном языке.

Задачи изучения дисциплины:

- изучение методов обработки естественного языка, применение междисциплинарных методов в обработке исследовательских данных;
- научиться применять понятийный математический аппарат в области лингвистики для решения практических задач профессиональной деятельности;
- приобрести навыки хранения, структуризации, анализа и визуализации текстового массива данных.

Дисциплина направлена на формирование общепрофессиональной (ОПК-2) компетенции выпускника.

2 Место дисциплины в структуре образовательной программы

Логико-структурный анализ дисциплины – курс входит обязательную часть БЛОКА 1 «Дисциплины (модули)» подготовки студентов по специальности 09.04.01 Информатика и вычислительная техника (Искусственный интеллект и цифровые двойники предприятий).

Дисциплина реализуется кафедрой интеллектуальных систем и информационной безопасности. Основывается на базе дисциплин из цикла подготовки бакалавров и специалистов.

Является основой для изучения следующих дисциплин: «Математические модели технологических процессов», «Выполнение и защита выпускной квалификационной работы».

Для изучения дисциплины необходимы компетенции, сформированные у студента для решения профессиональных задач деятельности, связанных с научно-исследовательской работы.

Курс является фундаментом для ориентации студентов в сфере научных исследований.

Общая трудоемкость освоения дисциплины составляет 6 зачетных единицы, 216 ак.ч. Программой дисциплины предусмотрены лекционные (36 ч.), лабораторные (36 ч.) и практические (18 ч.) занятия и самостоятельная работа студента (126 ч.).

Дисциплина изучается на 2 курсе во 4 семестре. Форма промежуточной аттестации – экзамен.

3 Перечень результатов обучения по дисциплине, соотнесённых с планируемыми результатами освоения образовательной программы

Процесс изучения дисциплины «Автоматическая обработка текста» направлен на формирование компетенции, представленной в таблице 1.

Таблица 1 –Компетенции, обязательные к освоению

Содержание компетенции	Код компетенции	Код и наименование индикатора достижения компетенции
Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач	ОПК-2	ОПК-2.1. Применяет современные интеллектуальные технологии для решения профессиональных задач

4 Объём и виды занятий по дисциплине

Общая трудоёмкость учебной дисциплины составляет 6 зачётных единицы, 216 ак.ч.

Самостоятельная работа студента (СРС) включает проработку материалов лекций, подготовку к практическим занятиям, текущему контролю, выполнение индивидуального задания, самостоятельное изучение материала и подготовку к экзамену.

При организации внеаудиторной самостоятельной работы по данной дисциплине используются формы и распределение бюджета времени на СРС для очной формы обучения в соответствии с таблицей 2.

Таблица 2 – Распределение бюджета времени на СРС

Вид учебной работы	Всего ак.ч.	Ак.ч. по семест рам
		5
Аудиторная работа, в том числе:	90	90
Лекции (Л)	36	36
Практические занятия (ПЗ)	18	18
Лабораторные работы (ЛР)	36	36
Курсовая работа/курсовой проект	-	-
Самостоятельная работа студентов (СРС), в том числе:	126	126
Подготовка к лекциям	9	9
Подготовка к лабораторным работам	18	18
Подготовка к практическим занятиям / семинарам	18	18
Выполнение курсовой работы / проекта	-	-
Расчетно-графическая работа (РГР)	-	-
Реферат (индивидуальное задание)	12	12
Домашнее задание	-	-
Подготовка к контрольным работам	-	-
Подготовка к коллоквиуму	-	-
Аналитический информационный поиск	18	18
Работа в библиотеке	18	18
Подготовка к экзамену	33	33
Промежуточная аттестация – экзамен (Э)	Э	Э
Общая трудоемкость дисциплины		
ак.ч.	144	144
з.е.	4	4

5 Содержание дисциплины

С целью освоения компетенции, приведенной в п.3 дисциплина разбита на 3 темы:

- тема 1 (Начальные этапы обработки текста);
- тема 2 (Подходы к автоматическому анализу синтаксиса и семантики текстов);
- тема 3 (Лингвистические ресурсы и прикладные задачи автоматической обработки текста).

Виды занятий по дисциплине и распределение аудиторных часов для очной формы приведены в таблице 3.

Таблица 3 – Виды занятий по дисциплине и распределение аудиторных часов (очная форма обучения)

№ п/п	Наименование темы (раздела) дисциплины	Содержание лекционных занятий	Трудоемкость в ак.ч.	Темы практических занятий	Трудоемкость в ак.ч.	Тема лабораторных занятий	Трудоемкость в ак.ч.
1	2	3	4	5	6	7	8
1	Начальные этапы обработки текста	Введение Начальные этапы обработки текста Статистические характеристики текстов и корпусная лингвистика	2 2 4	Подбор и предварительная подготовка текстов	4	Предварительная подготовка и анализ текстов	4
2	Подходы к автоматическому анализу синтаксиса и семантики текстов	Дискурсивный анализ текста. Имплиcitная информация. Клауза. Кореферентность. Корпус. Корпусная лингвистика. Лемма. Несвободные словосочетания. Онтология предметной области. Семантика текста. Стемминг. Токенизация. Эксплицитная информация. Эллиptические конструкции. Эмбеdдинг.	2 2 2 2 2 2	Изучение и тестирование методов автоматического анализа синтаксиса и семантики текстов	8	Дискурсивный анализ, имплиcitная информация, клауза, кореферентность. Корпус. Корпусная лингвистика. Лемма. Несвободные сло- восочетания. Онтология пред- метной области. Семантика текста. Стемминг. Токе- низация. Экспли- цитная инфор- мация. Эллипти- ческие конструк- ции. Эмбеdдинг.	4 6 4 6 6 6

Завершение таблицы 3

1	2	3	4	5	6	7	8
3	Лингвистические ресурсы и прикладные задачи автоматической обработки текста	Мировые лингвистические ресурсы Отечественные лингвистические ресурсы Прикладные задачи автоматической обработки текста	4 4 8	Изучение доступных лингвистических ресурсов	6		
Всего аудиторных часов			36	18		36	

6 Фонд оценочных средств для проведения текущего контроля успеваемости и промежуточной аттестации студентов по дисциплине

6.1 Критерии оценивания

В соответствии с Положением о кредитно-модульной системе организации образовательного процесса ФГБОУ ВО «ДонГТУ» (https://www.dstu.education/images/structure/license_certificate/polog_kred_modul.pdf) при оценивании сформированности компетенций по дисциплине используется 100-балльная шкала.

Перечень компетенций по дисциплине и способы оценивания знаний приведены в таблице 4.

Таблица 4 – Перечень компетенций по дисциплине и способы оценивания знаний

Код и наименование компетенции	Способ оценивания	Оценочное средство
ОПК-2	Экзамен	Комплект контролирующих материалов для экзамена

Всего по текущей работе в семестре студент может набрать 100 баллов, в том числе:

- лабораторные работы – всего 60 баллов;
- выполнение практических заданий – 40 баллов.

Экзаменационная оценка проставляется автоматически, если студент набрал в течении семестра не менее 60 баллов и отчитался за каждую контрольную точку. Минимальное количество баллов по каждому из видов текущей работы составляет 60% от максимального.

Экзамен по дисциплине «Автоматическая обработка текста» проводится по результатам работы в семестре. В случае, если полученная в семестре сумма баллов не устраивает студента, во время сессии студент имеет право повысить итоговую оценку либо в форме устного собеседования по приведенным ниже вопросам (п.п. 6.5), либо в результате тестирования.

Шкала оценивания знаний при проведении промежуточной аттестации приведена в таблице 5.

Таблица 5 – Шкала оценивания знаний

Сумма баллов за все виды учебной деятельности	Оценка по национальной шкале зачёт/экзамен
0-59	Не зачтено/неудовлетворительно
60-73	Зачтено/удовлетворительно
74-89	Зачтено/хорошо
90-100	Зачтено/отлично

6.2 Домашнее задание

Домашнее задание не предусмотрено.

6.3 Темы для рефератов (презентаций) – индивидуальное задание

1. Сферы применения автоматической обработки текста.
2. Теоретико-практические проблемы снятия лингвистических неоднозначностей.
3. Сегментация устного и письменного текста: проблемы и решения.
4. Современные модули преобразования устной речи в письменную.
5. Современные программы синтеза речи по тексту.
6. Обработка текстовой информации средствами свободного программного обеспечения
7. Автоматическое реферирование текстов на основе ранжирования связанных структур
8. Технология обработки текстовой информации
 - тема 1 (Начальные этапы обработки текста);
 - тема 2 (Подходы к автоматическому анализу синтаксиса и семантики текстов);
 - тема 3 (Лингвистические ресурсы и прикладные задачи автоматической обработки текста).

6.4 Оценочные средства для самостоятельной работы и текущего контроля успеваемости

Тема 1. Начальные этапы обработки текста.

1. Фрагмент текста:
 - а) слово
 - б) предложение
 - в) непрерывная часть текста
2. В виде чего хранится на внешнем запоминающем устройстве текст, который был набран в текстовом редакторе:
 - а) файла
 - б) папки
 - в) каталога
3. Какой из представленных форматов не относится к форматам файлов, в которых сохраняют текстовые документы:
 - а) TXT

б) ODT

в) PPT

4. Что называется систематизацией информации:

а) обработка документа с целью получения новых данных

б) разделение информации по определенному признаку +

в) кодирование данных

5. Первоначально специализированное устройство, позже компьютерная программа, используемая для набора, сохранения, редактирования и печати текста:

а) текстовый процесс

б) текстовый процессор +

в) текстовый файл

Тема 2. Подходы к автоматическому анализу синтаксиса и семантики текстов.

1. Раздел науки о языке, в котором изучается словосочетание, предложение, текст, а также правила их построения – это :

а) синтаксис

б) пунктуация

в) фонетика

2. Пунктуация:

а) Раздел науки о языке, в котором изучается система знаков препинания и правила их построения

б) Раздел науки о языке, в котором изучается словосочетание, предложение, текст, а также правила их построения.

в) раздел лингвистики, изучающий звуки речи и звуковое строение языка.

3. Словосочетание состоит из ... частей.

а) 2+

б) 1

в) 3

4. Грамматическая основа предложения - это...

а) главные члены предложения

б) второстепенные члены предложения

в) словосочетания

5. Может ли грамматическая основа предложения состоять из одного члена:

а) да

б) нет

в) не знаю

Тема 3. Лингвистические ресурсы и прикладные задачи автоматической обработки текста.

1. База данных - это:

а) совокупность данных, организованных по определенным правилам

б) совокупность программ для хранения и обработки больших массивов информации

в) интерфейс, поддерживающий наполнение и манипулирование

данными

г) определенная совокупность информации

2. Наиболее распространенными в практике являются:

а) распределенные базы данных

б) иерархические базы данных

в) сетевые базы данных

г) реляционные базы данных

3. Наиболее точным аналогом реляционной базы данных может служить:

а) неупорядоченное множество данных

б) вектор

в) генеалогическое дерево;

г) двумерная таблица.

4. Таблицы в базах данных предназначены:

а) для хранения данных базы;

б) для отбора и обработки данных базы;

в) для ввода данных базы и их просмотра;

г) для автоматического выполнения группы команд;

д) для выполнения сложных программных действий.

5. В каком режиме работает с базой данных пользователь:

а) в проектировочном

б) в любительском

в) в заданном

г) эксплуатационном

6.5 Вопросы для подготовки к экзамену

1) Какие Вы знаете основные особенности и сложности естественных языков?

2) В чем суть явления полисемии? омонимии? Приведите примеры.

3) Перечислите основные этапы обработки текста в системах АОТ.

4) Какие лингвистические ресурсы используются в лингвистических процессорах?

5) Какие Вы знаете типичные приложения методов автоматической обработки текстов?

6) В чем заключается этап графематического анализа текста?

7) Что такое морфема? аффикс?

8) Чем основа слова отличается от корня? Приведите примеры.

9) Что такое словоизменительная парадигма?

10) В чем заключается лемматизация?

11) Приведите пример морфологической омонимии.

12) Чем лемма отличается от лексемы?

13) Какие Вы знаете виды морфологического анализа?

14) Как определяется статистика словоупотреблений в текстах?

15) Что такое биграмма? триграмма?

16) Объясните смысл закона Ципфа-Мальдельброта.

17) В чем отличие коллекции текстов от корпуса?

- 18) Какие бывают типы разметки в корпусе текстов?
- 19) Что такое статистическая языковая модель?
- 20) Какие статистические меры применяются для извлечения словосочетаний?
- 21) Что такое синтаксическое дерево?
- 22) В чем отличие деревьев составляющих от деревьев зависимостей?
- 23) Что такое модель управления слова-предиката? Приведите примеры.
- 24) Какие методы синтаксического разбора Вы знаете?
- 25) В чем состоит синтаксическая сегментация текста?
- 26) Укажите особенности семантической сети как способа представления смысла текста.
- 27) Что такое семантический падеж?
- 28) Как семантические падежи используются при анализе текста?
- 29) Какие Вы знаете отличительные характеристики связного текста?
- 30) Что такое анафорическая ссылка?
- 31) Какие виды смысловых связей лексических единиц Вы знаете?
- 32) Что такое тезаурус?
- 33) Что такое онтология? Приведите пример. Какие виды онтологий бывают?
- 34) Какие семантические связи представлены в системе WordNet?
- 35) Какие Вы знаете приложения АОТ, в которых нужен морфологический анализ?
- 36) В каких приложениях АОТ применяется синтаксический анализ?
- 37) Что такое индексация текста?
- 38) Какие модели информационного поиска вы знаете?
- 39) В чем заключается задача классификации текстов?
- 40) Чем классификация текстов отличается от кластеризации?
- 41) Что такое рубрицирование текстов?
- 42) Какие бывают стратегии машинного перевода?
- 43) Какие Вы знаете особенности задачи извлечения информации из текстов?

7. Учебно-методическое и информационное обеспечение дисциплины

7.1 Рекомендуемая литература

Основная литература

1. Нугуманова, А. Б. Автоматизированная обработка текстовых массивов : учебник и практикум для вузов / А. Б. Нугуманова. — 2-е изд. — Москва : Издательство Юрайт, 2024. — 82 с. — (Высшее образование). — ISBN 978-5-534-20738-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/558668> (дата обращения: 26.08.2024).

Дополнительная литература

1. Е.И. Большакова. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011 — 272 с. URL: [http://peskova.ru/books/NLP_miem_2011\(textbook\).pdf](http://peskova.ru/books/NLP_miem_2011(textbook).pdf) (Дата обращения 26.08.2024).

Учебно-методические материалы и пособия

1. Методические указания к выполнению лабораторной работы «Автоматизированная обработка текстовой информации» по курсам «Современные технологии и способы обработки информации» и «Информационные технологии в лингвистике» : (для студ. напр. подготовки 38.04.05 «Бизнес-информатика» II курса и 45.03.02 «Лингвистика» I курса всех форм обуч.) / сост. Н.А. Подгорная, Н.В. Клочко ; Каф. Экономической кибернетики и информационных технологий . — Алчевск : ГОУ ВПО ЛНР ДонГТУ, 2018 . — 17 с. <https://library.dstu.education/download.php?rec=107619>.

7.2 Базы данных, электронно-библиотечные системы, информационно-справочные и поисковые системы

1. Научная библиотека ДонГТУ : официальный сайт.— Алчевск. — URL: library.dstu.education .— Текст : электронный.
2. Научно-техническая библиотека БГТУ им. Шухова : официальный сайт. — Белгород. — URL: <http://ntb.bstu.ru/jirbis2/> .— Текст : электронный.
3. Консультант студента : электронно-библиотечная система.— Москва. — URL: <http://www.studentlibrary.ru/cgi-bin/mb4x> .— Текст : электронный.
4. Университетская библиотека онлайн : электронно-библиотечная система.— URL: http://biblioclub.ru/index.php?page=main_ub_red .— Текст : электронный.
5. Сайт кафедры ИСИБ <http://scs.dstu.education>

8 Материально-техническое обеспечение дисциплины

Материально-техническая база обеспечивает проведение всех видов деятельности в процессе обучения, соответствует требованиям ФГОС ВО.

Материально-техническое обеспечение представлено в таблице 6.

Таблица 6 – Материально-техническое обеспечение

Наименование оборудованных учебных кабинетов	Адрес (местоположение) учебных кабинетов
Специальные помещения: <i>Мультимедийная аудитория. (60 посадочных мест), оборудованная специализированной (учебной) мебелью (скамья учебная – 20 шт., стол – 1 шт., доска аудиторная – 1 шт.), учебное ПК (монитор + системный блок), мультимедийная стойка с оборудованием – 1 шт., широкоформатный экран.</i> Аудитории для проведения лекций:	ауд. <u>207</u> корп. <u>4</u>
<i>Компьютерные классы (22 посадочных места), оборудованный учебной мебелью, компьютерами с неограниченным доступом к сети Интернет, включая доступ к ЭБС:</i> ПК – 11 шт.; интерактивная доска – 1 шт.	ауд. <u>208</u> корп. <u>4</u>

Лист согласования рабочей программы дисциплины

Разработал

и.о. заведующего кафедрой
интеллектуальных систем
и информационной безопасности
(должность)


(подпись)

Е.Е. Бизянов
Ф.И.О.)

(должность)

(подпись)

Ф.И.О.)

И.о. заведующего кафедрой
интеллектуальных систем
и информационной безопасности


(подпись)

Е.Е. Бизянов
Ф.И.О.)

Протокол № 1 заседания кафедры ИСИБ от 27.08.2024 г

И.о. декана факультета


(подпись)

В.В. Дьячкова
Ф.И.О.)

Согласовано:

Председатель методической
комиссии по направлению 09.04.01
«Информатика и вычислительная техника»


(подпись)

Е.Е. Бизянов
Ф.И.О.)

Начальник учебно-методического центра


(подпись)

О.А. Коваленко
Ф.И.О.)

Лист изменений и дополнений

Номер изменения, дата внесения изменения, номер страницы для внесения изменений	
ДО ВНЕСЕНИЯ ИЗМЕНЕНИЙ:	ПОСЛЕ ВНЕСЕНИЯ ИЗМЕНЕНИЙ:
Основание:	
Подпись лица, ответственного за внесение изменений	