

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
ИМЕНИ В.Ф. УТКИНА»

Кафедра «Вычислительная и прикладная математика»

«СОГЛАСОВАНО»

Заведующий кафедрой ВПМ
/ Г.В. Овечкин

27.01 2023 г

«УТВЕРЖДАЮ»

Проректор по учебной работе
/ А.В. Корячко

* 27.01 2023 г



РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

МАШИННАЯ ОБРАБОТКА ТЕКСТОВ НА ЕЯ

Направление подготовки
09.03.04 Программная инженерия

Направленность (профиль) подготовки
Программное обеспечение систем искусственного интеллекта

Квалификация выпускника — бакалавр

Форма обучения — очная

Рязань 2023 г

Программу составил(и):

д.техн.н., проф., Каширин Игорь Юрьевич

Рабочая программа дисциплины

Машинная обработка текстов на ЕЯ

разработана в соответствии с ФГОС ВО:

ФГОС ВО - бакалавриат по направлению подготовки 09.03.04 Программная инженерия (приказ Минобрнауки России от 19.09.2017 г. № 920)

составлена на основании учебного плана:

09.03.04 Программная инженерия

утвержденного учёным советом вуза от 27.01.2023 протокол № 6.

Рабочая программа одобрена на заседании кафедры

Вычислительной и прикладной математики

Протокол от 29.12.2022 г. № 4

Срок действия программы: 2023-2027 уч.г.

Зав. кафедрой Овечкин Геннадий Владимирович

Общая трудоемкость

3 ЗЕТ

Распределение часов дисциплины по семестрам

Семестр (<Курс>.<Семестр на курсе>)	8 (4.2)		Итого	
Неделя	8			
Вид занятий	УП	РП	УП	РП
Лекции	24	24	24	24
Лабораторные	24	24	24	24
Иная контактная работа	0,35	0,35	0,35	0,35
Консультирование перед экзаменом и практикой	2	2	2	2
Итого ауд.	50,35	50,35	50,35	50,35
Контактная работа	50,35	50,35	50,35	50,35
Сам. работа	4	4	4	4
Часы на контроль	53,65	53,65	53,65	53,65
Итого	108	108	108	108

Визирование РПД для исполнения в очередном учебном году

Рабочая программа пересмотрена, обсуждена и одобрена для
исполнения в 2024-2025 учебном году на заседании кафедры

Вычислительной и прикладной математики

Протокол от _____ 2024 г. № ____

Зав. кафедрой _____

Визирование РПД для исполнения в очередном учебном году

Рабочая программа пересмотрена, обсуждена и одобрена для
исполнения в 2025-2026 учебном году на заседании кафедры

Вычислительной и прикладной математики

Протокол от _____ 2025 г. № ____

Зав. кафедрой _____

Визирование РПД для исполнения в очередном учебном году

Рабочая программа пересмотрена, обсуждена и одобрена для
исполнения в 2026-2027 учебном году на заседании кафедры

Вычислительной и прикладной математики

Протокол от _____ 2026 г. № ____

Зав. кафедрой _____

Визирование РПД для исполнения в очередном учебном году

Рабочая программа пересмотрена, обсуждена и одобрена для
исполнения в 2027-2028 учебном году на заседании кафедры

Вычислительной и прикладной математики

Протокол от _____ 2027 г. № ____

Зав. кафедрой _____

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)	
1.1	Целью освоения дисциплины является: изучение основных принципов и методов автоматической обработки текстов на естественном языке (ЕЯ).
1.2	Задачи:
1.3	- освоить основные понятия и концепции в области обработки текста на естественном языке;
1.4	- изучить различные алгоритмы и модели, используемые для обработки текста на естественном языке, включая методы машинного обучения, глубокое обучение и статистические модели;
1.5	- изучить вопросы применения и возможностей машинной обработки текста на естественном языке в различных областях.

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ	
Цикл (раздел) ОП:	Б1.В.ДВ.02
2.1	Требования к предварительной подготовке обучающегося:
2.1.1	Основы работы с большими данными (Big Data)
2.1.2	Функциональное и логическое программирование
2.2	Дисциплины (модули) и практики, для которых освоение данной дисциплины (модуля) необходимо как предшествующее:
2.2.1	Выполнение и защита выпускной квалификационной работы

3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)	
ПК-8: Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	
ПК-8.1. Осуществляет поиск данных в открытых источниках, специализированных библиотеках и репозиториях	
Знать методы поиска данных. Уметь отделять достоверные источники данных от сомнительных, осуществлять критический отбор данных, проверять их на целостность и непротиворечивость. Владеть методологией поиска данных из разных источников.	
ПК-8.2. Выполняет подготовку и разметку структурированных и неструктурированных данных для машинного обучения	
Знать методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки структурированных и неструктурированных данных; методы планирования вычислительного эксперимента, формирования обучающей и контрольной выборок. Уметь выявлять и исключать из массива данных ошибочные данные и выбросы; выделять входные и выходные переменные с целью использования предиктивных моделей. Владеть основами подготовки структурированных и неструктурированных данных для машинного обучения.	

ПК-10: Способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	
ПК-10.2. Участвует в реализации проектов в области сквозной цифровой субтехнологии «Обработка естественного языка»	
Знать принципы построения систем обработки естественного языка, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка». Уметь решать задачи по выполнению коллективной проектной деятельности для создания, поддержки и использования систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка». Владеть методологией сквозной цифровой субтехнологии «Обработка естественного языка».	

В результате освоения дисциплины (модуля) обучающийся должен

3.1 Знать:	
3.1.1	- подходы в обработке естественного языка;
3.1.2	- графематический, морфологический, синтаксический, семантический этапы анализа текстов;
3.1.3	- особенности текстов, такие как омонимия и сочетаемость слов;
3.1.4	- грамматики Хомского в синтаксическом этапе обработки текста.
3.2 Уметь:	
3.2.1	- обрабатывать текст на ЕЯ;
3.2.2	- выполнять синтаксический анализ текста с использованием лингвистического взгляда и грамматик Хомского;
3.2.3	- анализировать текст методом "мешка слов" и использованием N-грамм.
3.3 Владеть:	
3.3.1	- в обработке и анализе текстов на ЕЯ;
3.3.2	- работы с различными этапами обработки текста;
3.3.3	- с методами TF-IDF, "мешок слов".

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Компетенции	Литература	Форма контроля
	Раздел 1. Машинная обработка текстов на ЕЯ					
1.1	Аналитический подход в обработке естественного языка. Графематический, морфологический, синтаксический, семантический этапы анализа текстов. Статистический подход к анализу текстов на основе мешка слов. /Тема/	8	0			
1.2	Обработка текста на ЕЯ. Этапы и принципы. Особенности текстов. /Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2/Л2.1/Л3.1 Э1	экзамен
1.3	Графематический и морфологический этапы обработки текста. Омонимия. Сочетаемость слов. /Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2/Л2.1/Л3.1 Э1	экзамен
1.4	Синтаксический этап обработки текста. Лингвистический взгляд. Грамматики Хомского. /Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2/Л2.1/Л3.1 Э1	экзамен
1.5	Семантический этап обработки текста. Онтологии и тезаурусы. /Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2/Л2.1/Л3.1 Э1	экзамен

1.6	Корпуса текстов. TF-IDF. Поиск терминов. //Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.7	Анализ текстов методом "мешка слов". N- граммы //Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.8	Графематический и морфологический этапы обработки текста на ЕЯ. //Лаб/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.9	Синтаксический этап обработки текста. Лингвистический взгляд. Грамматики Хомского. //Лаб/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.10	Семантический этап обработки текста. Онтологии и тезаурусы. //Лаб/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.11	Корпуса текстов. TF-IDF. Поиск терминов. //Лаб/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.12	Анализ текстов методом «Мешок слов». N- граммы. //Лаб/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен

1.13	Анализ текстов методом «Мешок слов». N- граммы. /Ср/	8	1	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.14	Корпуса текстов. TF-IDF. Поиск терминов. /Ср/	8	1	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.15	Семантический этап обработки текста. Онтологии и тезаурусы. /Ср/	8	1	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.16	Синтаксический этап обработки текста. Лингвистический взгляд. Грамматики Хомского. /Ср/	8	1	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.17	Нейросетевой подход в обработке естественного языка. Предобученные модели. /Тема/	8	0			
1.18	Статистический анализ слов. Word2Vec для хранения контекста. /Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.19	Автоматическое выделение ключевых элементов текста. /Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.20	Обработка разговорного ЕЯ. /Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен

1.21	Чат-боты. Тест Тьюринга и "китайская комната". //Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.22	Краткое содержание текстов. //Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.23	Генерация текстов, модели GPT и YaLM. //Лек/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.24	Статистический анализ слов. Word2Vec для хранения контекста. //Лаб/	8	2	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.25	Автоматическое выделение ключевых элементов текста. //Лаб/	8	3	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.26	Обработка разговорного ЕЯ. //Лаб/	8	3	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
1.27	Чат-боты. Тест Тьюринга и "китайская комната". //Лаб/	8	3	ПК-8.1-3 ПК-8.1-У ПК-8.1-В ПК-8.2-3 ПК-8.2-У ПК-8.2-В ПК-10.2-3 ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен

1.28	Краткое содержание текстов. /Лаб/	8	3	ПК-8.1-З ПК-8.1-У ПК-8.1-В ПК-8.2-З ПК-8.2-У ПК-8.2-В ПК-10.2-З ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
	Раздел 2. Промежуточная аттестация					
2.1	Экзамен /Тема/	8	0			
2.2	Приём экзамена /ИКР/	8	0,35	ПК-8.1-З ПК-8.1-У ПК-8.1-В ПК-8.2-З ПК-8.2-У ПК-8.2-В ПК-10.2-З ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
2.3	Консультация перед экзаменом /Кнс/	8	2	ПК-8.1-З ПК-8.1-У ПК-8.1-В ПК-8.2-З ПК-8.2-У ПК-8.2-В ПК-10.2-З ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен
2.4	Подготовка к экзамену /Экзамен/	8	53,65	ПК-8.1-З ПК-8.1-У ПК-8.1-В ПК-8.2-З ПК-8.2-У ПК-8.2-В ПК-10.2-З ПК-10.2-У ПК-10.2-В	Л1.1 Л1.2Л2.1Л3.1 Э1	экзамен

5. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине представлен в виде оценочных материалов и приведен в Приложении к рабочей программе дисциплины (см. документ «Оценочные материалы по дисциплине «Машинная обработка текстов на ЕЯ»).

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

6.1. Рекомендуемая литература

6.1.1. Основная литература

№	Авторы, составители	Заглавие	Издательство, год	Количество/ название ЭБС
Л1.1	Батура Т. В.	Математическая лингвистика и автоматическая обработка текстов на естественном языке : учебное пособие	Новосибирск: Новосибирский государственный университет, 2016, 166 с.	978-5-4437- 0548-4, http://www.ipr- bookshop.ru/9 3489.html

№	Авторы, составители	Заглавие	Издательство, год	Количество/ название ЭБС
Л1.2	Редкол.:Твердохлеб Н.Г.(отв.ред.) и др.;Киев.ин-т нар.хоз- ва	Машинная обработка информации : Респ.межвед.сборник	Киев:Лыбидь, 1990, 132с.	, 1

6.1.2. Дополнительная литература

№	Авторы, составители	Заглавие	Издательство, год	Количество/ название ЭБС
Л12.1	Батура, Т. В.	Математическая лингвистика и автоматическая обработка текстов на естественном языке : учебное пособие	Новосибирск: Новосибирский государственный университет, 2016, 166 с.	978-5-4437-0548-4, https://www.iprbookshop.ru/93489.html

6.1.3. Методические разработки

№	Авторы, составители	Заглавие	Издательство, год	Количество/ название ЭБС
Л13.1	Пруцков А.В.	Машинная обработка знаний и текстов на естественных языках : Метод.указ.к лаб.раб.и курс.проекту	Рязань, 2005, 28с.	, 1

6.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

Э1	Электронная библиотека РГРТУ https://elibr.rsreu.ru/ebs
----	--

6.3 Перечень программного обеспечения и информационных справочных систем

6.3.1 Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства

Наименование	Описание
PyCharm Community	Свободное ПО
Python	Свободно распространяемое программное обеспечение под лицензиями
Qt Creator Community	Свободное ПО
Операционная система Ubuntu	Свободное ПО

6.3.2 Перечень информационных справочных систем

6.3.2.1	Справочная правовая система «КонсультантПлюс» (договор об информационной поддержке №1342/455-100 от 28.10.2011 г.)
6.3.2.2	Система КонсультантПлюс http://www.consultant.ru
6.3.2.3	Информационно-правовой портал ГАРАНТ.РУ http://www.garant.ru

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

1	206-1 учебно-административный корпус. Учебная аудитория для проведения занятий лекционного и семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации 42 мест, 1 ПК: ЦП: Intel Pentium 4 class 3200 ОЗУ: 1 Гб ПЗУ: 80 Гб Телевизор: PHILIPS U7PEL4606H/60 документ-камера: AVER Media POB3 (AverVision 330)
---	---

2	206-2 учебно-административный корпус. Аудитория для самостоятельной работы 18 мест, Телевизор PHILIPS 46PFL3208T/60; документ-камера: AverVisionF33 POE7D; 20 ПК с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно- образовательную среду: ЦП: Intel Pentium II/III class 2327 ОЗУ: 2 Гб ПЗУ: 80 Гб (1 шт.) ЦП: Intel Pentium III 2992 ОЗУ: 1,5 Гб ПЗУ: 150 Гб (1 шт.) ЦП: Intel Pentium III 2660 ОЗУ: 2 Гб ПЗУ: 80 Гб (9 шт.) ЦП: Intel Pentium III 2793 ОЗУ: 2 Гб ПЗУ: 100 Гб (1 шт.) ЦП: Intel Pentium II/III class 2660 ОЗУ: 1 Гб ПЗУ: 50 Гб (1 шт.) ЦП: Intel Pentium III 2527 ОЗУ: 2 Гб ПЗУ: 100 Гб (1 шт.) ЦП: Intel Pentium III 3158 ОЗУ: 2 Гб ПЗУ: 50 Гб (3 шт.) ЦП: Intel Pentium III 2826 ОЗУ: 2 Гб ПЗУ: 100 Гб (2 шт.) ЦП: Intel Pentium III 2693 ОЗУ: 1,5 Гб ПЗУ: 100 Гб (1 шт.)
3	206-5 учебно-административный корпус. Учебная аудитория для проведения практический занятий, лабораторных работ 24 ПК с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду: ЦП: Intel Pentium II/III class 2394 ОЗУ: 2 Гб ПЗУ: 70 Гб (17 шт.) ЦП: Intel Pentium II/III class 2327 ОЗУ: 2 Гб ПЗУ: 100 Гб (1 шт.) ЦП: Intel Pentium III Xeon 3093 ОЗУ: 4 Гб ПЗУ: 300 Гб (6 шт.)
4	106 учебно-административный корпус. Учебная аудитория для проведения занятий лекционного и семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации 30 мест проектор BENQ 11 ПК с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно- образовательную среду: ЦП: AMD 3411, ОЗУ: 4Гб, ПЗУ:780 Гб (4 штук); ЦП: AMD 3013, ОЗУ: 4 Гб, ПЗУ: 780 Гб (3 штук); ЦП: Intel Pentium 4 class 2659, ОЗУ: 1 Гб, ПЗУ: 50 Гб (4 штук).
5	106а учебно-административный корпус. Учебная аудитория для проведения занятий лекционного и семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации 42 мест проектор BENQ 15 ПК с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно- образовательную среду: ЦП: 2x Intel Pentium II/III class 2126, ОЗУ: 2 Гб, ПЗУ: 74 Гб (1 шт) ЦП: Intel Pentium II/III class 3192, ОЗУ: 4 Гб, ПЗУ: 200 Гб (13 шт.) ЦП: Intel Pentium II/III class 2128, ОЗУ: 2 Гб ПЗУ: 74 Гб (1 шт.)

8. МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Методические указания для студентов по освоению дисциплины

Приступая к работе, каждый студент должен принимать во внимание нижеследующие положения.

Дисциплина построена по модульному принципу, каждый модуль представляет собой логически завершённый раздел дисциплины.

В первом семестре четыре модуля (включая экзамен). Во втором семестре четыре модуля (включая экзамен).

На первом занятии студент получает информацию для доступа к комплексу методических материалов по дисциплине.

Лекционные занятия посвящены рассмотрению ключевых, базовых положений курса и разъяснению учебных заданий, выносимых на самостоятельную проработку.

Лабораторные работы предназначены для приобретения опыта практической реализации основной профессиональной образовательной программы.

Методические документы к лабораторным работам прорабатываются студентами во время самостоятельной подготовки. Необходимый уровень подготовки контролируется перед сдачей лабораторной работы.

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ В.Ф. УТКИНА»**

Кафедра «Вычислительная и прикладная математика»

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ
«Машинная обработка текстов на ЕЯ»**

Направление подготовки
09.03.04 «Программная инженерия»

Направленность (профиль) подготовки
«Программное обеспечение систем искусственного интеллекта»

Уровень подготовки – бакалавриат

Квалификация выпускника – бакалавр

Форма обучения – очная

Срок обучения – 4 года

Рязань 2023 г.

1. ОБЩИЕ ПОЛОЖЕНИЯ

Оценочные материалы – это совокупность учебно-методических материалов и процедур, предназначенных для оценки качества освоения обучающимися данной дисциплины как части основной образовательной программы.

Цель – оценить соответствие знаний, умений и уровня приобретенных компетенций, обучающихся целям и требованиям основной образовательной программы в ходе проведения текущего контроля и промежуточной аттестации.

Основная задача – обеспечить оценку уровня сформированности компетенций и индикаторов их достижения, приобретаемых обучающимся в соответствии с этими требованиями.

Контроль знаний обучающихся проводится в форме текущего контроля и промежуточной аттестации.

Текущий контроль успеваемости и промежуточная аттестация проводятся с целью определения степени усвоения учебного материала, своевременного выявления и устранения недостатков в подготовке обучающихся, организации работы обучающихся в ходе учебных занятий и оказания им индивидуальной помощи.

К контролю текущей успеваемости относятся проверка знаний, умений и навыков обучающихся на практических занятиях по результатам выполнения и защиты обучающимися индивидуальных заданий, по результатам выполнения контрольных работ и тестов, по результатам проверки качества конспектов лекций и иных материалов.

В качестве оценочных средств на протяжении семестра используется устные и письменные ответы студентов на индивидуальные вопросы, письменное тестирование по теоретическим разделам курса, реферат. Дополнительным средством оценки знаний и умений студентов является отчет о выполнении практических заданий и его защита.

По итогам курса обучающиеся сдают экзамен. Форма проведения – устный ответ с письменным подкреплением по утвержденным билетам, сформулированным с учетом содержания дисциплины. В билет для экзамена включается два теоретических вопроса и задача. В процессе подготовки к устному ответу студент должен составить в письменном виде план ответа.

1. Перечень компетенций с указанием этапов их формирования

При освоении дисциплины формируются следующие компетенции: ПК-8 (индикаторы ПК-8.1, ПК-8.2), ПК-10 (индикаторы ПК-10.2).

Указанные компетенции формируются в соответствии со следующими этапами:

- формирование и развитие теоретических знаний, предусмотренных указанными компетенциями (лекционные занятия, самостоятельная работа студентов);
- приобретение и развитие практических умений, предусмотренных компетенциями (практические занятия, самостоятельная работа студентов);
- закрепление теоретических знаний, умений и практических навыков, предусмотренных компетенциями, в ходе решения конкретных задач на занятиях, выполнения индивидуальных заданий на практических занятиях и их защиты, а также в процессе сдачи экзамена.

2 Показатели и критерии оценивания компетенций (результатов) на различных этапах их формирования, описание шкал оценивания

Сформированность каждой компетенции в рамках освоения данной дисциплины оценивается по трехуровневой шкале:

- пороговый уровень является обязательным для всех обучающихся по завершении освоения дисциплины;

- продвинутый уровень характеризуется превышением минимальных характеристик сформированности компетенций по завершении освоения дисциплины;
- эталонный уровень характеризуется максимально возможной выраженностью компетенций и является важным качественным ориентиром для самосовершенствования.

При достаточном качестве освоения более 80% приведенных знаний, умений и навыков преподаватель оценивает освоение данной компетенции в рамках настоящей дисциплины на эталонном уровне, при освоении более 60% приведенных знаний, умений и навыков – на продвинутом, при освоении более 40% приведенных знаний, умений и навыков – на пороговом уровне. При освоении менее 40% приведенных знаний, умений и навыков компетенция в рамках настоящей дисциплины считается неосвоенной.

Уровень сформированности каждой компетенции на различных этапах ее формирования в процессе освоения данной дисциплины оценивается в ходе текущего контроля успеваемости и представлено различными видами оценочных средств.

Оценке сформированности в рамках данной дисциплины подлежат компетенции/индикаторы:

Показатели достижения планируемых результатов обучения и критерии их оценивания на разных уровнях формирования компетенций приведены в таблице 1.

Таблица 1. Показатели достижения индикаторов компетенции

1	2	3	4
Компетенция: код по ФГОС 3++, формулировка	Индикаторы	Этап	Наименование оценочного средства
ПК-8 (09.03.04/02 Программное обеспечение систем искусственного интеллекта) Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	ПК-8.1 Осуществляет поиск данных в открытых источниках, специализированных библиотеках и репозиториях <u>Знать</u>: методы поиска данных. <u>Уметь</u>: отделять достоверные источники данных от сомнительных, осуществлять критический отбор данных, проверять их на целостность и непротиворечивость. <u>Владеть</u>: методологией поиска данных из разных источников. ПК-8.2 Выполняет подготовку и разметку структурированных и неструктурированных данных для машинного обучения <u>Знать</u>: методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки структурированных и неструктурированных данных; методы планирования вычислительного экспе-	1	Экзамен

1	2	3	4
	римента, формирования обучающей и контрольной выборок. <u>Уметь</u>: выявлять и исключать из массива данных ошибочные данные и выбросы; выделять входные и выходные переменные с целью использования предиктивных моделей. <u>Владеть</u>: основами подготовки структурированных и неструктурированных данных для машинного обучения.		
ПК-10 (09.03.04/02 Программное обеспечение систем искусственного интеллекта) Способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий ИИ	ПК-10.2 Участвует в реализации проектов в области сквозной цифровой субтехнологии "Обработка естественного языка" <u>Знать</u>: принципы построения систем обработки естественного языка, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка». <u>Уметь</u>: решать задачи по выполнению коллективной проектной деятельности для создания, поддержки и использования систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка». <u>Владеть</u>: методологией сквозной цифровой субтехнологии «Обработка естественного языка».	1	Экзамен

Преподавателем оценивается содержательная сторона и качество материалов, приведенных в отчетах студента по практическим занятиям. Кроме того, преподавателем учитываются ответы студента на вопросы по соответствующим видам занятий при текущем контроле:

- контрольные опросы;
- задания для практических занятий.

Принимается во внимание **знания** обучающимися:

- подходов в обработке естественного языка;
- графематический, морфологический, синтаксический, семантический этапов анализа текстов;
- особенностей текстов, таких как омонимия и сочетаемость слов;

- грамматики Хомского в синтаксическом этапе обработки текста.

наличие **умений**:

- обрабатывать текст на ЕЯ;
- выполнять синтаксический анализ текста с использованием лингвистического взгляда и грамматик Хомского;
- анализировать текст методом "мешка слов" и использованием N-грамм.

обладание навыками:

- обработки и анализа текстов на ЕЯ;
- работы с различными этапами обработки текста;
- работы с методами TF-IDF, "мешок слов".

Критерии оценивания уровня сформированности компетенции в процессе выполнения практических работ:

41%-60% правильных ответов соответствует пороговому уровню сформированности компетенции на данном этапе ее формирования;

61%-80% правильных ответов соответствует продвинутому уровню сформированности компетенции на данном этапе ее формирования;

81%-100% правильных ответов соответствует эталонному уровню сформированности компетенции на данном этапе ее формирования.

Сформированность уровня компетенций не ниже порогового является основанием для допуска обучающегося к промежуточной аттестации по данной дисциплине.

Формой промежуточной аттестации по данной дисциплине является экзамен, оцениваемый по принятой в ФГБОУ ВО «РГРТУ» четырехбалльной системе: «неудовлетворительно», «удовлетворительно», «хорошо» и «отлично».

Критерии оценивания промежуточной аттестации представлены в таблице.

Шкала оценивания	Критерии оценивания
«отлично»	студент должен: продемонстрировать глубокое и прочное усвоение знаний материала; исчерпывающе, последовательно, грамотно и логически стройно изложить теоретический материал; правильно формулировать определения; уметь сделать выводы по излагаемому материалу; безупречно ответить не только на вопросы билета, но и на дополнительные вопросы в рамках рабочей программы дисциплины; выполнить все практические задания, предусмотренные программой
«хорошо»	студент должен: продемонстрировать достаточно полное знание материала; продемонстрировать знание основных теоретических понятий; достаточно последовательно, грамотно и логически стройно излагать материал; уметь сделать достаточно обоснованные выводы по излагаемому материалу; ответить на все вопросы билета; выполнить все практические задания, предусмотренные программой.
«удовлетворительно»	студент должен: продемонстрировать общее знание изучаемого материала; знать основную рекомендуемую программой дисциплины учебную литературу; уметь строить ответ в соответствии со структурой излагаемого вопроса; показать общее владение понятийным аппаратом дисциплины; уметь устранить допущенные погрешности в ответе на теоретические вопросы; выполнить все практические задания, предусмотренные программой.

«неудовлетворительно»	ставится в случае: невыполнения практических занятий; незнания значительной части пройденного материала; не владения понятийным аппаратом дисциплины; существенных ошибок при изложении учебного материала; неумения строить ответ в соответствии со структурой излагаемого вопроса; неумения делать выводы по излагаемому материалу. Как правило, оценка «неудовлетворительно» ставится студентам, которые не могут продолжить обучение по образовательной программе без дополнительных занятий по соответствующей дисциплине (формирования и развития компетенций, закрепленных за данной дисциплиной). Оценка «неудовлетворительно» выставляется также, если студент после начала экзамена отказался его сдавать или нарушил правила сдачи экзамена (списывал, подсказывал, обманом пытался получить более высокую оценку и т.д.).
------------------------------	--

3. Типовые контрольные задания или иные материалы

ФОС по дисциплине содержит следующие оценочные средства, позволяющие оценить знания, умения и уровень приобретенных компетенций при текущем контроле и промежуточной аттестации, разбитые по модулям дисциплины:

- перечни экзаменационных вопросов;
- макеты билетов к экзамену.

Средства для оценки различных уровней формирования компетенций по категориям «знать», «уметь», «владеть» обеспечивают реализацию основных принципов контроля, таких, как объективность и независимость, практико-ориентированность, междисциплинарность.

С учетом этого, контрольные вопросы (задания, задачи,) входящие в ФОС, для различных категорий и уровней освоения компетенций имеют следующий вид:

Уровень ЗНАТЬ

Дескрипторы	Пример задания из оценочного средства
виды представления данных, методы поиска и парсинга данных	1. Опишите этапы обработки текстов на ЕЯ, особенности каждого. Приведите примеры используемых инструментов.
уровни представления данных (ODS, DDL, семантический слой, модель данных)	1. Опишите этапы обработки текстов на ЕЯ, особенности каждого. Приведите примеры используемых инструментов.
основные инструменты, библиотеки и технологии Data Science	1. Опишите этапы обработки текстов на ЕЯ, особенности каждого. Приведите примеры используемых инструментов.
методы редукции размерности элементов набора данных и их предварительной статистической обработки, разметки структурированных и неструктурированных данных	1. Вектор документа. Применение. Основные подходы к формированию. Приведите пример.
методы планирования вычислительного эксперимента, формирования обучающей и	1. Вектор документа. Применение. Основные подходы к формированию. Приведите пример. 2. Корпуса текстов, назначение. Назовите особенно-

контрольной выборки	сти создания, приведя примеры использования.
принципы построения систем обработки ЕЯ, методы и технологии ИИ для анализа ЕЯ, методы и подходы к планированию и реализации проектов по созданию СИИ на основе сквозной субтехнологии "Обработка естественного языка"	1. Опишите этапы обработки текстов на ЕЯ, особенности каждого. Приведите примеры используемых инструментов.

Уровень УМЕТЬ

Дескрипторы	Пример задания из оценочного средства
отделять достоверные источники данных от сомнительных, осуществлять критических отбор данных, проверять их на целостность и непротиворечивость	1. Лингвистические онтологии и тезаурусы. Что изучает онтология? Что такое тезаурус? Применение. 2. Граматики Хомского, особенности.
использовать инструменты и библиотеки для DataScience для поиска данных в открытых источниках, специализированных библиотеках и репозиториях	1. Назовите способы выделения терминов из текстов, приведя алгоритмы. Опишите особенности каждого.
выявлять и исключать из массива данных ошибочные данные и выбросы	1. Лингвистические онтологии и тезаурусы. Что изучает онтология? Что такое тезаурус? Применение.
выявлять входные и выходные переменные с целью использования предиктивных моделей	1. Опишите существующие архитектуры для нейросетевых моделей ЕЯ. Приведите характерные особенности для каждой из них.
осуществлять разметку структурированных и неструктурированных данных	1. Отличительные особенности текстов на русском языке. Приведите пример.
использовать инструменты, библиотеки и технологии Data Science для подготовки и разметки структурированных и неструктурированных данных для машинного обучения	1. Опишите алгоритмы получения краткого содержания текстов на ЕЯ. 2. Алгоритмы генерирования текстов. Опишите существующие модели и приведите примеры возможного их использования, учитывая особенности.
использовать методы и технологии массово параллельной обработки и анализа данных	1. Опишите существующие архитектуры для нейросетевых моделей ЕЯ. Приведите характерные особенности для каждой из них.
применять методы и подходы к планированию и реализации проектов по созданию систем ИИ на основе сквозной цифровой субтехнологии "Обработка ЕЯ"	1. Чат-боты, назначение, примеры использования. Опишите алгоритм взаимодействия составляющих элементов, этапы обработки.

Перечень вопросов к экзамену и макет экзаменационного билета

1. Этапы обработки текстов на ЕЯ и их особенности.
2. Основные подходы к формированию вектора документа.
3. Способы выделения терминов из текстов
4. Корпуса текстов, примеры, назначение, особенности создания
5. Лингвистические онтологии и тезаурусы
6. Грамматики Хомского
7. Достоинства и недостатки аналитического и нейросетевого подходов к анализу текстов на ЕЯ
8. Существующие архитектуры для нейросетевых моделей ЕЯ
9. Тест Тьюринга сильные и слабые стороны
10. Чат-боты, назначение, примеры использования
11. Алгоритмы получения краткого содержания текстов на ЕЯ
12. Отличительные особенности текстов на русском языке
13. Алгоритмы генерирования текстов
14. Задача Коши для ОДУ. Метод Рунге - Кутты 2-го порядка точности. Оценка точности.

Макет экзаменационного билета

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Рязанский государственный радиотехнический университет имени В.Ф. Уткина» (РГРТУ)

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №1

по дисциплине «**Машинная обработка текстов на ЕЯ**»

1. Этапы обработки текстов на ЕЯ и их особенности.
2. Достоинства и недостатки аналитического и нейросетевого подходов к анализу текстов на ЕЯ.
3. Алгоритмы генерирования текстов

Билет рассмотрен и утвержден на заседании кафедры «__» _____ 20 г.

Перечень лабораторных работ

Лабораторная работа 1.1. Графематический и морфологический этапы обработки текста на ЕЯ.

Цель работы. Знакомство с этапами и принципами обработки текста на ЕЯ, а также получение практических навыков использования инструментов. Знакомство с графематическим и морфологическим этапами обработки текстов, стеммингом, лемматизацией, омонимией.

Задание:

Необходимо настроить рабочее окружение. В лабораторной работе потребуется определить морфологические особенности слов, их склонение, а также описание правил сочетания в роде, числе и падеже, используя изученные инструменты (pymorphy2, msystem и др.)

Источник данных задается индивидуально по варианту.

Лабораторная работа 1.2. Синтаксический этап обработки текста. Лингвистический взгляд. Грамматики Хомского.

Цель работы. Знакомство с этапами и принципами обработки текста на ЕЯ, а также получение практических навыков использования инструментов. Знакомство с синтаксическим этапом обработки текста, лингвистической постановкой задачи.

Задание:

Необходимо настроить рабочее окружение. В лабораторной работе потребуется определять связи слов в предложениях, используя изученные инструменты (nltk, pymorphy2, msystem и др.) Используя GC-grammar, определить части предложения в тексте.

Источник данных задается индивидуально по варианту.

Лабораторная работа 1.3. Семантический этап обработки текста. Онтологии и тезаурусы.

Цель работы. Знакомство с этапами и принципами обработки текста на ЕЯ, а также получение практических навыков использования инструментов. Знакомство с семантическим этапом обработки текста, онтологией. Изучить применение тезаурусов.

Задание:

Необходимо настроить рабочее окружение. В лабораторной работе потребуется создавать онтологии, используя изученные инструменты. Используя семантические падежи, разметить части текста.

Источник данных задается индивидуально по варианту.

Лабораторная работа 1.4 Корпуса текстов. TF-IDF. Поиск терминов.

Цель работы. Изучить применение корпусов текстов, TF-IDF, а также поиск терминов на примере работы со статистическими инструментами.

Задание:

Необходимо настроить рабочее окружение. В лабораторной работе потребуется статистически определять термины в наборе документов, используя изученные инструменты (RusCoproга и др.). Использовать аналитическое выделение терминов в наборе документов.

Лабораторная работа 1.5. Анализ текстов методом «Мешок слов». N-граммы.

Цель работы. Знакомство с методом анализа текстов «Мешок слов». Изучить применение N-грамм.

Задание:

Необходимо настроить рабочее окружение. В лабораторной работе потребуется составить «Мешок слов», определить тематически сходные документы, используя изученные инструменты.

Источник данных задается индивидуально по варианту.

Лабораторная работа 2.1 Статистический анализ слов. Word2Vec для хранения контекста.

Цель работы. Практическое использование Word2Vec для хранения контекста при статистическом анализе слов в тексте. Изучить особенности выходных данных.

Задание:

Необходимо настроить рабочее окружение. Подготовить текстовые данные для дальнейшей обработки методом контекстного представления в модели Word2Vec. Полученную модель протестировать на корректность и адекватность.

Источник данных задается индивидуально по варианту.

Лабораторная работа 2.2 Автоматическое выделение ключевых элементов текста.

Цель работы. Применение алгоритмов выделения ключевых элементов к векторам документов. Изучить особенности выходных данных.

Задание:

Необходимо настроить рабочее окружение. Подготовить текстовые данные для дальнейшей обработки методом выделения ключевых элементов. Полученные выходные данные объяснить.

Источник данных задается индивидуально по варианту.

Лабораторная работа 2.3 Обработка разговорного ЕЯ.

Цель работы. Практическое использование инструментов для задач поиска шаблонов высказываний.

Задание: Необходимо настроить рабочее окружение. Подготовить текстовые данные для дальнейшего поиска шаблонов, используя изученные инструменты (SpaCy, OpenNLTK, StanfordNLP и др.).

Источник данных задается индивидуально по варианту.

Лабораторная работа 2.4 Чат-боты. Тест Тьюринга и «китайская комната».

Цель работы. Изучение облачных технологий Dialogflow. Реализация чат-бота для изучения экспериментов Тьюринга и «китайской комнаты»

Задание:

Разработать чат-бота Dialogflow. Изучить особенности реализации. Индивидуальное задание.

Лабораторная работа 2.5 Краткое содержание текстов.

Цель работы. Применение алгоритмов суммаризации текста.

Задание:

Необходимо настроить рабочее окружение. Подготовить текстовые данные для дальнейшей обработки методом суммаризации. В лабораторной работе потребуется реализовать алгоритмы суммаризации и получить содержание текста.

Источник данных задается индивидуально по варианту.