Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»
(ВлГУ)**

Институт информационных технологий и радиоэлектроники

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
«Анализ естественного языка»

направление подготовки / специальность
09.04.04 «Программная инженерия»

направленность (профиль) подготовки
Инженерия искусственного интеллекта

г. Владимир
2021

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью освоения дисциплины «Анализ естественного языка» является ознакомление студентов с современными методами анализа естественного языка, основанными на глубоких нейронных сетях и машинном обучении. Рассматриваются задачи классификации текста, автоматической генерации текста с использованием рекуррентных нейронных сетей, включая LSTM и GRU, одномерных сверточных сетей, а также сетей с архитектурой Transformer.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Анализ естественного языка» относится к части учебного плана, формируемой участниками образовательных отношений.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине, в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции *(код, содержание индикатора*	Результаты обучения по дисциплине	
ПК-7. Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых субтехнологий искусственного интеллекта в прикладных областях	ПК-7.1. Знать: ПК-7.1.1. принципы построения систем компьютерного зрения, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Компьютерное зрение» ПК-7.1.2. 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» ПК-7.1.3. современное состояние и перспективы развития новых направлений, методов и технологий в области	Знает: 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»; современное состояние и перспективы развития новых направлений, методов и технологий в области искусственного интеллекта Умеет: руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка»	1) Контрольные работы 2) Рейтинг-контроль 3) Выполнение лабораторных работ 4) Зачет с оценкой

	искусственного интеллекта ПК-7.2. Уметь: ПК-7.2.1. руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Компьютерное зрение» ПК-7.2.2. руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка» ПК-7.3. Иметь навыки: ПК-7.3.1. проведения анализа новых направлений, методов и технологий в области искусственного интеллекта и определения наиболее перспективных для различных областей применения	Имеет навыки: проведения анализа новых направлений, методов и технологий в области искусственного интеллекта и определения наиболее перспективных для различных областей применения	

4. ОБЪЕМ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 4 зачетных единицы, 144 часа

Тематический план
форма обучения – очная

№ п/п	Наименование тем и/или разделов/тем дисциплины	Семестр	Неделя семестра	Контактная работа обучающихся с педагогическим работником				Самостоятельная работа	Формы текущего контроля успеваемости, форма промежуточной аттестации *(по семестрам)*
				Лекции	Практические занятия	Лабораторные работы	*в форме практической подготовки*		
1	Теоретические аспекты обработки естественного языка.	3	1-3	3		3	1	15	

2	Предварительная обработка текста.	3	4-6	2		2	1	13	Рейтинг-контроль №1
3	Векторизация текста.	3	7-8	2		2	1	13	
4	Машинное обучение для обработки текстов.	3	9-10	3		3	2	15	
5	Нейронные сети в решении задач текстовой обработки.	3	11-12	2		2	1	13	Рейтинг-контроль №2
6	Языковая модель.	3	13-14	2		2	1	13	
7	Поиск именованных сущностей.	3	15-16	2		2	1	13	
8	Механизм внимания. Трансформер.	3	17-18	2		2	1	13	Рейтинг-контроль №3
Всего за 3 семестр:				18		18		108	Зачет с оценкой
Наличие в дисциплине КП/КР									
Итого по дисциплине				18		18		108	Зачет с оценкой

Содержание лекционных занятий по дисциплине

1. Теоретические аспекты обработки естественного языка.

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

2. Предварительная обработка текста.

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

3. Векторизация текста.

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

4. Машинное обучение для обработки текстов.

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

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.

6. Языковая модель.

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

7. Поиск именованных сущностей.

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

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.

Содержание лабораторных занятий по дисциплине

1. Предварительная обработка текста для анализа.
2. Векторизация текста.
3. Классификация текста с использованием классических методов машинного обучения.
4. Классификация текста с использованием глубоких нейронных сетей.
5. Языковая модель. Обучение языковой модели.
6. Автоматическая генерация текста.
7. Поиск именованных объектов в тексте.
8. Механизм внимания в нейронных сетях. Сети с трансформаторной архитектурой.
9. Передача обучения в задачах обработки текстов.

5. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

5.1. Текущий контроль успеваемости

Рейтинг-контроль №1

1. Синтаксический анализ
2. Морфологический анализ
3. Семантический анализ
4. Графематический анализ
5. Омонимия
6. Задачи лингвистического анализа
7. Очистка текста
8. Токенизация
9. Стемминг
10. Лемматизация
11. Удаление стоп-слов
12. Фильтрация наиболее частотных и наименее частотных слов

Рейтинг-контроль №2

1. Построение словаря
2. Мешок слов
3. TF-IDF
4. word2vec
5. fasttext
6. LDA
7. LSI
8. GloVe.
9. Решение задач классификации и определения тональности методами классического машинного обучения на основе векторных моделей.

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

Рейтинг-контроль №3

1. Языковая модель и дистрибутивная семантика
2. Обучение векторной модели
3. Задача генерации текста
4. Различные подходы к генерации текста
5. Задача поиска именованных сущностей в тексте
6. Применение нейронных сетей для поиска именованных сущностей.
7. Механизм внимания в нейронных сетях
8. Применение механизма внимания для обработки текста
9. Нейронные сети с архитектурой Transformer
10. Нейронные сети BERT, GPT
11. Перенос обучения.

5.2. Промежуточная аттестация по итогам освоения дисциплины *(зачет с оценкой)*

Перечень примерных вопросов для зачета с оценкой:

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.

5.3. Самостоятельная работа обучающегося

Самостоятельная работа обучающихся заключается в самостоятельном изучении отдельных тем, практической реализации заданий контрольных и самостоятельных работ по этим темам. Контроль выполнения самостоятельной работы проводится при текущих контрольных мероприятиях и на промежуточной аттестации по итогам освоения дисциплины. Учебно-методическое обеспечение самостоятельной работы – основная литература [1-2], дополнительная литература [1-2].

Примерная тематика контрольных работ:

Проектирование конвейера для задач обработки естественного языка.

Примерные задания в составе контрольных работ:

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

1. Способ подготовки текста к обработке.
2. Подход к маркировке текста.
3. Подход к векторизации текста.
4. Используемая модель машинного обучения.
5. Метод обучения модели.
6. Метод оценки качества модели.
7. Использование обученной модели для решения задачи анализа текста.
8. Другие шаги, которые могут потребоваться при решении проблемы.

Примеры задач обработки естественного языка, для которых необходимо создать конвейеры:

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

Примерная тематика СРС:

Самостоятельная работа №1:

«Обучение языковой модели для текстов на русском языке»

Примерные задания:

1. Обучите языковую модель русскому языку и используйте ее для создания текста. Для выполнения задачи вам необходимо выполнить следующее:

- Подготовьте набор данных с текстами на русском языке. Вы можете использовать готовые наборы данных или создать свои собственные.

- Обучите языковую модель на подготовленном наборе данных.

- Используя обученную языковую модель, сгенерируйте пять примеров текстов на русском языке.

- Разместите набор данных, код и обученную модель в открытом доступе на GitHub.

● Сделайте презентацию или технологическую статью о ходе работы, обосновании принятых решений и результатах работы.

* (Необязательное задание). Запишите видео, показывающее, как работает созданное решение.

Самостоятельная работа №2:

«Переподготовка предварительно подготовленной сети BERT»

Примерные задания:

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

● Подготовьте набор данных с текстами на русском языке для классификации. Вы можете использовать готовые наборы данных или создать свои собственные.

● Выберите предварительно обученную нейронную сеть с трансформаторной архитектурой, подходящую для задачи классификации текстов на русском языке.

● Выполните дополнительное обучение выбранной нейронной сети на подготовленном наборе данных.

● Выполните тестирование классификации текста с использованием обученной нейронной сети и оцените качество сети.

● Поместите набор данных, код и завершенную модель в открытый доступ на GitHub.

● Сделайте презентацию или технологическую статью о ходе работы, обосновании принятых решений и результатах работы.

* (Необязательное задание). Запишите видео, показывающее, как работает созданное решение.

Пример обучения нейронной сети BERT в тензорном потоке – https://www.tensorflow.org/text/tutorials/fine_tune_bert

Образец кода решения – https://colab.research.google.com/github/tensorflow/text/blob/master/docs/tutorials/fine_tune_bert.ipynb

Пример переподготовки нейронных сетей с трансформаторной архитектурой в Hugging Face – https://huggingface.co/transformers/training.html

Фонд оценочных материалов (ФОМ) для проведения аттестации уровня сформированности компетенций обучающихся по дисциплине оформляется отдельным документом.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Книгообеспеченность

Наименование литературы: автор, название, вид издания, издательство	Год издания	КНИГООБЕСПЕЧЕННОСТЬ
		Наличие в электронном каталоге ЭБС
Основная литература		
1. Цитульский Антон Максимович, Иванников	2020	https://cyberleninka.ru/article/n/nlp-

Александр Владимирович, Рогов Илья Сергеевич NLP - Обработка естественных языков // StudNet. 2020. №6		obrabotka-estestvennyh-yazykov
2. Чернобаев Игорь Дмитриевич, Суркова Анна Сергеевна, Панкратова Анна Зурабовна Моделирование текстов с использованием рекуррентных нейронных сетей // Труды НГТУ им. Р. Е. Алексеева. 2018. №1 (120).	2018	https://cyberleninka.ru/article/n/modelirovanie-tekstov-s-ispolzovaniem-rekurrentnyh-neyronnyh-setey
Дополнительная литература		
1. Дэвенпорт, Т. Внедрение искусственного интеллекта в бизнес-практику. Преимущества и сложности / Т. Дэвенпорт. - Москва : Альпина Паблишер, 2021. - 316 с. - ISBN 978-5-9614-3952-6. - Текст : электронный. Режим доступа : по подписке.	2021	https://www.studentlibrary.ru/book/ISBN9785961439526.html
2. Берджесс, Э. Искусственный интеллект - для вашего бизнеса : Руководство по оценке и применению / Э. Берджесс. - Москва : Интеллектуальная Литература, 2021. - 232 с. - ISBN 9-785-907274-81-5. - Текст : электронный. Режим доступа : по подписке.	2021	https://www.studentlibrary.ru/book/ISBN9785907274815.html

6.2. Периодические издания

1. Вестник компьютерных и информационных технологий ISSN 1810-7206.

2. Цифровая библиотека научно-технических изданий Института инженеров по электротехнике и радиоэлектронике (Institute of Electrical and Electronic Engineers (IEEE)) на английском языке – http://www.ieee.org/ieeexplore

6.3. Интернет-ресурсы

1. Academic Search Ultimate EBSCO publishing – http://search.ebscohost.com
2. eBook Collections Springer Nature – https://link.springer.com/
3. Гугл Академия – https://scholar.google.ru/
4. Электронно-библиотечная система «Лань» – https://e.lanbook.com/
5. Университетская библиотека ONLINE – https://biblioclub.ru/
6. Электронно-библиотечная система "Библиокомплектатор" (IPRbooks) http://www.bibliocomplectator.ru/available
7. Электронные информационные ресурсы Российской государственной библиотеки https://www.rsl.ru/
8. Научная электронная библиотека «КиберЛенинка» https://cyberleninka.ru/
9. Портал российского образования www.edu.ru
10. Портал российских электронных библиотек www.elbib.ru
11. Научная электронная библиотека www.eLibrary.ru
12. Научная библиотека ВлГУ library.vlsu.ru
13. Учебный сайт кафедры ИСПИ ВлГУ https://ispi.cdo.vlsu.ru
14. Электронная библиотечная система ВлГУ https://vlsu.bibliotech.ru/
15. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. URL: http://www.machinelearning.ru/
16. Онлайн курс “Программирование глубоких нейронных сетей на Python”. URL: https://openedu.ru/course/urfu/PYDNN/

17. Онлайн курс “Generating discrete sequences: language and music”. URL: https://www.edx.org/course/generating-discrete-sequences-language-and-music
18. Браславский П.И. Введение в обработку естественного языка. URL: https://stepik.org/course/1233/
19. Роман Суворов, Анастасия Янина, Алексей Сильвестров, Николай Капырин. Нейронные сети и обработка текста URL: https://stepik.org/course/54098

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Для реализации данной дисциплины имеются специальные помещения для проведения занятий: занятий лекционного типа, занятий лабораторного типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы. Лабораторные работы проводятся в компьютерном классе, оборудованном мультимедийным проектором с экраном и обеспеченным доступом в Интернет.

Перечень используемого лицензионного программного обеспечения:

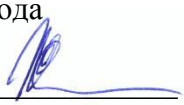
- Операционная система Microsoft Windows 10
- Офисный пакет Microsoft Office 2016
- Бесплатно-распространяемое программное обеспечение (Python – https://www.python.org/, TensorFlow – https://www.tensorflow.org/ , Hugging Face – https://huggingface.co/ , Веб - среда разработки для языка программирования Python: Google Colab – https://colab.research.google.com/).

Рабочую программу составил: зав. каф. ИСПИ И.Е. Жигалов 

Рецензент: к.т.н., ведущий специалист отдела ИТ ООО «Дау Изолан» Фадин Д.Н. 

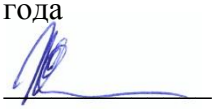
Программа рассмотрена и одобрена на заседании кафедры ИСПИ

Протокол № 5 от 15.12.21 года

Заведующий кафедрой И.Е. Жигалов 

Рабочая программа рассмотрена и одобрена на заседании учебно-методической комиссии направления 09.04.04 «Программная инженерия»

Протокол № 5 от 15.12.21 года

Председатель комиссии И.Е. Жигалов 

ЛИСТ ПЕРЕУТВЕРЖДЕНИЯ
РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

в рабочую программу дисциплины

Анализ естественного языка

образовательной программы направления подготовки *09.04.04«Программная инженерия»*,

направленность: *Инженерия искусственного интеллекта (магистратура*)

Номер изменения	Внесены изменения в части/разделы рабочей программы	Исполнитель ФИО	Основание (номер и дата протокола заседания кафедры)

Заведующий кафедрой _____/ _____

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»
(ВлГУ)**

Институт информационных технологий и радиоэлектроники
Кафедра информационных систем и программной инженерии

УТВЕРЖДАЮ
Заведующий кафедрой

_____ И.Е. Жигалов

«15» 12 2021

Основание:
решение кафедры

от «15» 12 2021

ФОНД ОЦЕНОЧНЫХ МАТЕРИАЛОВ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПРИ ИЗУЧЕНИИ УЧЕБНОЙ ДИСЦИПЛИНЫ

«Анализ естественного языка»

Направление подготовки: 09.04.04 «Программная инженерия»

Профиль подготовки: Инженерия искусственного интеллекта

Уровень высшего образования: магистратура

Владимир, 2021 г.

ПАСПОРТ ФОНДА ОЦЕНОЧНЫХ МАТЕРИАЛОВ

Фонд оценочных материалов (ФОМ) для текущего контроля успеваемости и промежуточной аттестации по дисциплине «Анализ естественного языка» разработан в соответствии с рабочей программой, входящей в ОПОП направления подготовки 09.04.04 «Программная инженерия», профиль подготовки «Инженерия искусственного интеллекта».

№ п/п	Контролируемые разделы (темы) дисциплины	Код контролируемой компетенции (или ее части)	Наименование оценочного материала
1	Теоретические аспекты обработки естественного языка.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой
2	Предварительная обработка текста.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой
3	Векторизация текста.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой
4	Машинное обучение для обработки текстов.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой
5	Нейронные сети в решении задач текстовой обработки.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой
6	Языковая модель.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой
7	Поиск именованных сущностей.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой
8	Механизм внимания. Трансформер.	ПК-7	Контрольные работы, выполнение лабораторных работ, рейтинг-контроль, зачет с оценкой

Комплект оценочных материалов по дисциплине «Анализ естественного языка» предназначен для аттестации обучающихся на соответствие их персональных достижений поэтапным требованиям образовательной программы, в том числе рабочей программы дисциплины «Анализ естественного языка», для оценивания результатов обучения: знаний, умений, владений и уровня приобретенных компетенций.

Комплект оценочных материалов по дисциплине «Анализ естественного языка» включает:

1. Оценочные материалы для проведения текущего контроля успеваемости:

- комплект вопросов рейтинг-контроля, позволяющих оценивать знание фактического материала (базовые понятия, алгоритмы, факты) и умение правильно использовать специальные термины и понятия, распознавание объектов изучения в рамках определенного раздела дисциплины;
- комплект вопросов для контроля самостоятельной работы обучающихся, позволяющих оценивать знание фактического материала.

2. Оценочные материалы для проведения промежуточной аттестации в форме

- контрольные вопросы для проведения зачета с оценкой, позволяющие провести процедуру измерения уровня знаний и умений обучающихся.

Перечень компетенций, формируемых в процессе изучения дисциплины «Анализ естественного языка» при освоении образовательной программы по направлению подготовки 09.04.04 «Программная инженерия»

ПК-7. Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых субтехнологий искусственного интеллекта в прикладных областях		
Знать	***Уметь***	***Иметь навыки***
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»; современное состояние и перспективы развития новых направлений, методов и технологий в области искусственного интеллекта	руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка»	проведения анализа новых направлений, методов и технологий в области искусственного интеллекта и определения наиболее перспективных для различных областей применения

Показатели, критерии и шкала оценивания компетенций <u>текущего</u> контроля знаний по учебной дисциплине «Анализ естественного языка»

Текущий контроль знаний, согласно «Положению о рейтинговой системе комплексной оценки знаний студентов в ВлГУ» (далее Положение) в рамках изучения дисциплины «Анализ естественного языка» предполагает письменный рейтинг-контроль, выполнение и защиту лабораторных работ, компьютерное тестирование.

Регламент проведения письменного рейтинг-контроля

№	Вид работы	Продолжительность
1	Предел длительности рейтинг-контроля	35-40 мин.
2	Внесение исправлений	до 5 мин.
	Итого	до 45 мин.

Критерии оценки письменного рейтинг-контроля

Результаты каждого письменного рейтинга оцениваются в баллах. Максимальная сумма, набираемая студентом на каждом письменном рейтинге, составляет 10 баллов.

Критерии оценки для письменного рейтинга:

- 9-10 баллов выставляется обучающемуся, если соблюдаются критерии: полное раскрытие темы, вопроса, указание точных названий и определений, правильная формулировка понятий и категорий, приведение формул и (в необходимых случаях) их вывода, приведение статистики, самостоятельность ответа, использование дополнительной литературы;

- 7-8 баллов выставляется обучающемуся, если соблюдаются критерии: недостаточно полное раскрытие темы, несущественные ошибки в определении понятий и категорий, формулах, выводе формул, статистических данных, кардинально не меняющих суть

изложения, наличие грамматических и стилистических ошибок, использование устаревшей учебной литературы;

- 6-7 баллов выставляется обучающемуся, если соблюдаются критерии: отражение лишь общего направления изложения лекционного материала и материала современных учебников, наличие достаточно количества несущественных или одной-двух существенных ошибок в определении понятий и категорий, формулах, их выводе, статистических данных, наличие грамматических и стилистических ошибок, использование устаревшей учебной литературы, неспособность осветить проблематику дисциплины;

- 1-6 выставляется обучающемуся, если соблюдаются критерии: нераскрытые темы; большое количество существенных ошибок, наличие грамматических и стилистических ошибок, отсутствие необходимых умений и навыков.

Регламент проведения лабораторных работ

В целях закрепления практического материала и углубления теоретических знаний по разделам дисциплины «Анализ естественного языка» предполагается выполнение лабораторных работ, что позволяет углубить процесс познания, раскрыть понимание прикладной значимости осваиваемой дисциплины.

Лабораторные работы выполняются на компьютерах.

Критерии оценки выполнения лабораторных работ

Результаты выполнения каждой лабораторной работы оцениваются в баллах. Максимальная сумма, набираемая студентом за выполнение каждой лабораторной работы, составляет 1 балл.

Критерии оценки для выполнения лабораторной работы:

- 0,9-1 балл выставляется обучающемуся, если соблюдаются критерии: представлен полный письменный отчет по лабораторной работе, содержащий описание всех этапов ее выполнения и надлежащим образом оформленный (в печатном или электронном виде - в соответствие с требованием преподавателя), полностью выполнено задание на лабораторную работу, обучающийся верно и полно ответил на все контрольные вопросы преподавателя по теоретической и практической части лабораторной работы, лабораторная работа выполнена самостоятельно и в определенный преподавателем срок;

- 0,7-0,8 баллов выставляется обучающемуся, если соблюдаются критерии: представлен недостаточно полный письменный отчет по лабораторной работе, содержащий описание всех этапов ее выполнения, имеющий, возможно, погрешности в оформлении (в печатном или электронном виде - в соответствие с требованием преподавателя), полностью выполнено задание на лабораторную работу, обучающийся преимущественно верно и полно ответил на контрольные вопросы преподавателя по теоретической и практической части лабораторной работы, лабораторная работа выполнена самостоятельно, возможно, с нарушением определенного преподавателем срока предоставления отчета, отчет содержит грамматические и стилистические ошибки;

- 0,6-0,7 баллов выставляется обучающемуся, если соблюдаются критерии: представлен недостаточно полный письменный отчет по лабораторной работе, содержащий описание не всех этапов ее выполнения, имеющий, возможно, погрешности в оформлении (в печатном или электронном виде - в соответствие с требованием преподавателя), в основном выполнено задание на лабораторную работу, обучающийся ответил на контрольные вопросы преподавателя по теоретической и практической части лабораторной работы с отражением лишь общего направления изложения материала, с наличием достаточно количества несущественных или одной-двух существенных ошибок, лабораторная работа выполнена самостоятельно, с нарушением определенного преподавателем срока предоставления отчета, отчет содержит грамматические и стилистические ошибки, при его составлении использована устаревшая учебная литература;

- 0,1-0,6 выставляется обучающемуся, если соблюдаются критерии: письменный отчет по лабораторной работе (в печатном или электронном виде - в соответствие с требованием преподавателя) не представлен или представлен неполный, отчет содержит описание не всех этапов выполнения работы, имеет погрешности в оформлении, задание на лабораторную работу выполнено не полностью, обучающийся ответил на контрольные вопросы преподавателя по теоретической и практической части лабораторной работы с большим количеством существенных ошибок, продемонстрировал неспособность осветить проблематику лабораторной работы, лабораторная работа выполнена несамостоятельно, с существенным нарушением определенного преподавателем срока предоставления отчета, отчет содержит грамматические и стилистические ошибки, при его составлении использована устаревшая учебная литература, обучающийся при выполнении работы продемонстрировал отсутствие необходимых умений и практических навыков.

При оценке за лабораторную работу менее 0,6 баллов, данная работа считается невыполненной и не зачитывается. При невыполнении лабораторной работы хотя бы по одной из изучаемых тем, обучающийся не получает положительную оценку при промежуточном контроле по дисциплине (зачете с оценкой).

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ ЗНАНИЙ ПО УЧЕБНОЙ ДИСЦИПЛИНЕ «Анализ естественного языка»

Перечень вопросов для текущего контроля знаний (письменный рейтинг-контроль)

Рейтинг-контроль №1

1. Синтаксический анализ
2. Морфологический анализ
3. Семантический анализ
4. Графематический анализ
5. Омонимия
6. Задачи лингвистического анализа
7. Очистка текста
8. Токенизация
9. Стемминг
10. Лемматизация
11. Удаление стоп-слов
12. Фильтрация наиболее частотных и наименее частотных слов

Рейтинг-контроль №2

1. Построение словаря
2. Мешок слов
3. TF-IDF
4. word2vec
5. fasttext
6. LDA
7. LSI
8. GloVe.
9. Решение задач классификации и определения тональности методами классического машинного обучения на основе векторных моделей.
10. Архитектуры нейронных сетей для обработки текстов: рекуррентные (LSTM, GRU), одномерные сверточные.
11. Применение нейронных сетей для обработки текстов.

Рейтинг-контроль №3

1. Языковая модель и дистрибутивная семантика
2. Обучение векторной модели
3. Задача генерации текста
4. Различные подходы к генерации текста
5. Задача поиска именованных сущностей в тексте
6. Применение нейронных сетей для поиска именованных сущностей.
7. Механизм внимания в нейронных сетях
8. Применение механизма внимания для обработки текста
9. Нейронные сети с архитектурой Transformer
10. Нейронные сети BERT, GPT
11. Перенос обучения.

Темы лабораторных работ:

1. Предварительная обработка текста для анализа.
2. Векторизация текста.
3. Классификация текста с использованием классических методов машинного обучения.
4. Классификация текста с использованием глубоких нейронных сетей.
5. Языковая модель. Обучение языковой модели.
6. Автоматическая генерация текста.
7. Поиск именованных объектов в тексте.
8. Механизм внимания в нейронных сетях. Сети с трансформаторной архитектурой.
9. Передача обучения в задачах обработки текстов.

Перечень вопросов для контроля самостоятельной работы обучающегося

Самостоятельная работа обучающихся заключается в самостоятельном изучении отдельных тем, практической реализации заданий контрольных и самостоятельных работ по этим темам. Контроль выполнения самостоятельной работы проводится при текущих контрольных мероприятиях и на промежуточной аттестации по итогам освоения дисциплины.

Примерная тематика контрольных работ:

Проектирование конвейера для задач обработки естественного языка.

Примерные задания в составе контрольных работ:

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

1. Способ подготовки текста к обработке.
2. Подход к маркировке текста.
3. Подход к векторизации текста.
4. Используемая модель машинного обучения.
5. Метод обучения модели.
6. Метод оценки качества модели.
7. Использование обученной модели для решения задачи анализа текста.
8. Другие шаги, которые могут потребоваться при решении проблемы.

Примеры задач обработки естественного языка, для которых необходимо создать конвейеры:

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

Примерная тематика СРС:

Самостоятельная работа №1:

«Обучение языковой модели для текстов на русском языке»

Примерные задания:

1. Обучите языковую модель русскому языку и используйте ее для создания текста. Для выполнения задачи вам необходимо выполнить следующее:

- Подготовьте набор данных с текстами на русском языке. Вы можете использовать готовые наборы данных или создать свои собственные.

- Обучите языковую модель на подготовленном наборе данных.

- Используя обученную языковую модель, сгенерируйте пять примеров текстов на русском языке.

- Разместите набор данных, код и обученную модель в открытом доступе на GitHub.

- Сделайте презентацию или технологическую статью о ходе работы, обосновании принятых решений и результатах работы.

* (Необязательное задание). Запишите видео, показывающее, как работает созданное решение.

Самостоятельная работа №2:

«Переподготовка предварительно подготовленной сети BERT»

Примерные задания:

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

- Подготовьте набор данных с текстами на русском языке для классификации. Вы можете использовать готовые наборы данных или создать свои собственные.

- Выберите предварительно обученную нейронную сеть с трансформаторной архитектурой, подходящую для задачи классификации текстов на русском языке.

- Выполните дополнительное обучение выбранной нейронной сети на подготовленном наборе данных.

- Выполните тестирование классификации текста с использованием обученной нейронной сети и оцените качество сети.

- Поместите набор данных, код и завершенную модель в открытый доступ на GitHub.

- Сделайте презентацию или технологическую статью о ходе работы, обосновании принятых решений и результатах работы.

* (Необязательное задание). Запишите видео, показывающее, как работает созданное решение.

Пример обучения нейронной сети BERT в тензорном потоке – https://www.tensorflow.org/text/tutorials/fine_tune_bert

Образец кода решения – https://colab.research.google.com/github/tensorflow/text/blob/master/docs/tutorials/fine_tune_bert.ipynb

Пример переподготовки нейронных сетей с трансформаторной архитектурой в Hugging Face – https://huggingface.co/transformers/training.html

Общее распределение баллов текущего и промежуточного контроля по видам учебных работ для студентов (в соответствие с Положением)

№	Пункт	Максимальное число баллов
1	Письменный рейтинг-контроль 1	10
2	Письменный рейтинг-контроль 2	10
3	Письменный рейтинг-контроль 3	10
4	Посещение занятий студентом	5
5	Дополнительные баллы (бонусы)	5
6	Выполнение лабораторных работ и семестрового плана самостоятельной работы	60
7	Всего	100

Показатели, критерии и шкала оценивания компетенций промежуточной аттестации знаний по учебной дисциплине «Анализ естественного языка» на зачете с оценкой

Регламент проведения промежуточного контроля (зачета с оценкой)

Промежуточная аттестация по итогам освоения дисциплины (зачет с оценкой) проводится перед экзаменационной сессией. Зачет проставляется студенту после выполнения студентом семестрового плана самостоятельной работы.

Критерии оценивания компетенций при проставлении зачета

Критерии оценки для промежуточного контроля (зачета с оценкой):

- оценка «отлично» (соответствует 91-100 баллов по шкале рейтинга) выставляется обучающемуся, если соблюдаются критерии: теоретическое содержание оцениваемой части дисциплины освоено полностью, необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены в установленные сроки, качество их выполнения оценено числом баллов, близким к максимальному;

- оценка «хорошо» (соответствует 74-90 баллов по шкале рейтинга) выставляется обучающемуся, если соблюдаются критерии: теоретическое содержание курса освоено полностью, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками или с нарушением установленных сроков;

- оценка «удовлетворительно» (соответствует 61-73 баллов по шкале рейтинга) выставляется обучающемуся, если соблюдаются критерии: теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий, возможно, содержат ошибки;

- оценка «неудовлетворительно» (соответствует менее 60 баллов по шкале рейтинга) выставляется обучающемуся, если соблюдаются критерии: теоретическое содержание курса не освоено, необходимые практические навыки работы не сформированы, выполненные учебные задания содержат грубые ошибки.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО УЧЕБНОЙ ДИСЦИПЛИНЕ «Анализ естественного языка»

Перечень примерных вопросов для промежуточной аттестации (зачет с оценкой)

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 оценивания компетенций при аттестации по дисциплине

Максимальная сумма баллов, набираемая студентом по дисциплине «Анализ естественного языка» в течение семестра равна 100

Оценка в баллах	Оценка по дисциплине	Критерии оценивания компетенций	Уровень сформированности компетенций
91 - 100	«Отлично»	Теоретическое содержание курса освоено полностью без пробелов, необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному.	Высокий
74 - 90	«Хорошо»	Теоретическое содержание курса освоено полностью без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками.	Продвинутый
61 - 73	«Удовлетворительно»	Теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий, возможно, содержат ошибки.	Пороговый
0 - 60	«Неудовлетворительно»	Теоретическое содержание курса не освоено, необходимые практические навыки работы не сформированы, выполненные учебные задания содержат грубые ошибки.	Компетенции не сформированы