ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«24» июня 2025 г.
Протокол № 2

Рабочая программа дисциплины (модуля)
«Natural Language Processing (Обработка естественного языка)»

Направление подготовки: *02.04.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Продуктовая аналитика*

Квалификация (степень) выпускника: *магистр*

Форма обучения: *очная*

Срок освоения программы: *2 года*

Год набора: *2025*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Москва
2025

Содержание

1. **Краткая характеристика дисциплины (модуля)** 3
2. **Перечень планируемых результатов обучения** 5
3. **Тематический план** 7
4. **Содержание дисциплины (модуля)** 7
5. **Учебно-методическое обеспечение** 8
6. **Материально-техническое обеспечение** 8
7. **Методические и оценочные материалы** 10

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Natural Language Processing (Обработка естественного языка)» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – магистратура по специальности 02.04.01 Математика и компьютерные науки, профиль Продуктовая аналитика, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 810 от 23.08.2017 года.

Изучение дисциплины (модуля) «Natural Language Processing (Обработка естественного языка)» играет ключевую роль в создании систем, которые могут понимать и генерировать человеческий язык, что открывает новые возможности для взаимодействия между людьми и машинами. Эта область обеспечивает развитие технологий, таких как чат-боты, автоматический перевод и анализ настроений, которые становятся все более важными в различных сферах, включая бизнес, медицину и образование.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки магистратуры по направлению 02.04.01 Математика и компьютерные науки, профиль Продуктовая аналитика и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений, как дисциплина по выбору.

Дисциплина (модуль) изучается на 1 или 2 курсе во 2, 3 или 4 семестре на выбор, доступна для прохождения при условии успешного завершения дисциплины (модуля) «Machine Learning (Машинное обучение)» и общеуниверситетского факультатива «Deep Learning (Глубокое обучение).

Цель изучения дисциплины (модуля): приобретение знаний и навыков, необходимых для разработки и применения технологий, позволяющих компьютерам анализировать, понимать и взаимодействовать с человеческим языком в различных контекстах.

Задачи изучения дисциплины (модуля):

- изучить фундаментальные приемы подготовки текстовых данных для машинного анализа и преобразования их в числовые представления.
- освоить подходы к решению основных проблем в области обработки языков, включая категоризацию документов, выделение ключевых фактов, маркировку элементов и синтез новых текстов;
- познакомиться с технологиями адаптации готовых языковых моделей для специфических задач анализа естественного языка;
- развить навыки измерения эффективности и объяснения поведения моделей в контексте различных сценариев обработки текста;
- исследовать способы объединения языковых моделей с внешними источниками информации для улучшения генерации ответов.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

- основные способы предобработки и векторизации текста;
- методы решения ключевых задач обработки естественного языка: классификация текстов, извлечение информации, тегирование последовательностей, генерация текста, суммаризация;
- принципы построения и методы дообучения предобученных языковых моделей (BERT, GPT и др.) для задач NLP;
- базовые концепции оценки качества и интерпретации работы языковых моделей в различных задачах обработки естественного языка;

— подходы к интеграции внешних знаний в LLM с использованием RAG (Retrieval Augmented Generation);
— методы построения агентных систем с применением LLM;
— принципы работы мультимодальных языковых моделей, способных обрабатывать и интегрировать различные модальности, такие как текст и звук.

уметь:

— конструировать и реализовывать пайплайны обработки текста для решения типовых и прикладных задач NLP;
— применять и дообучать предобученные языковые модели (например, BERT, GPT и др.) на собственные датасеты для решения задач обработки естественного языка;
— оптимизировать производительность и ускорять работу NLP-систем, используя методы сокращения размера моделей, квантизацию, дистилляцию и внедрение эффективных вычислительных решений;
— осуществлять комплексную оценку качества разработанных моделей с использованием метрик (accuracy, F1-score, BLEU, ROUGE, LLM-as-a-judge и др.), а также применять методы повышения качества (тюннинг гиперпараметров, аугментация данных и др.).

владеть:

— основными инструментами и библиотеками для NLP (NLTK, spaCy, HuggingFace Transformers,HuggingFace Datasets, vllm, PEFT, SentenceTransformers и др.);
— приемами работы с предобученными языковыми моделями, включая их применение, дообучение и оптимизацию;
— умением самостоятельно разрабатывать, оптимизировать и оценивать NLP-решения для прикладных задач, а также интерпретировать их результаты.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
УК-6.	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1.	Знает основные методы самооценки и анализа своей деятельности, а также принципы управления временем и целеполагания
		УК-6.2.	Умеет ставить реалистичные и достижимые цели, определять приоритеты в своей деятельности, а также разрабатывать и внедрять планы по совершенствованию своих навыков и компетенций на основе полученной самооценки
		УК-6.3.	Имеет практический опыт применения методов самооценки в своей профессиональной деятельности, включая участие в тренингах, семинарах и проектах, направленных на развитие личной эффективности и профессионального роста
ОПК-2.	Способен создавать и исследовать новые математические модели в естественных науках, совершенствовать и разрабатывать концепции, теории и методы	ОПК-2.1.	Знает основные математические модели и методы, используемые в естественных науках, включая статистическое моделирование, дифференциальные уравнения и численные методы, а также современные подходы к исследованию и анализу данных
		ОПК-2.2.	Умеет разрабатывать и адаптировать математические модели для решения конкретных проблем в естественных науках, проводить их анализ и верификацию, а также интерпретировать полученные результаты в контексте научных исследований
		ОПК-2.3.	Имеет практический опыт создания и исследования математических моделей в

			рамках научных проектов или исследований, включая участие в публикациях, конференциях или коллаборациях, где были разработаны и апробированы новые концепции и методы
ПК-3.	Способен решать задачи профессиональной деятельности в области продуктовой аналитики, формулировать результаты анализа и выявлять последствия полученных данных для принятия обоснованных решений и оптимизации продуктов	ПК-3.1.	Знает методы и инструменты продуктовой аналитики
		ПК-3.2.	Умеет применять аналитические инструменты и программное обеспечение для обработки и визуализации данных, а также формулировать выводы на основе проведенного анализа
		ПК-3.3.	Имеет опыт работы над реальными проектами в области продуктовой аналитики, включая анализ пользовательского поведения и оптимизацию продуктов на основе полученных данных
ПК-4.	Способен публично представлять собственные и известные научные результаты	ПК-4.1.	Знает основные принципы эффективного публичного выступления, методы визуализации данных и основные требования к научным презентациям, включая структуру и содержание
		ПК-4.2.	Умеет четко и логично формулировать свои научные результаты, адаптируя их для различных аудиторий, а также использовать визуальные средства для улучшения восприятия информации
		ПК-4.3.	Имеет практический опыт участия в научных конференциях, семинарах или других мероприятиях, где успешно представлял свои и известные научные результаты, получая обратную связь и взаимодействуя с аудиторией

3. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары (практические занятия)			
1	Encoder-based модели	8	8		34	Подготовка к семинару, Домашние задания
2	Большие языковые модели (LLM)	10	10		42	Подготовка к семинару, Домашние задания
3	Современные возможности и расширения больших языковых моделей (LLM)	12	12		50	Подготовка к семинару, Домашние задания
	Зачет			4		
	Итого:	***30***	***30***	***4***	***126***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Encoder-based модели	Векторизация и классификация текстов Предобученные языковые модели на основе архитектуры Кодировщик Трансформера (BERT и др.) Ранжирование и методы информационного поиска Мультилингвальность и доменная адаптация предобученных языковых моделей
2	Большие языковые модели (LLM)	Языковое моделирование и обучение больших языковых моделей (LLM) Дообучение и fine-tuning LLM Efficient Transformers, PEFT и оценка эффективности LLM Сжатие и оптимизация LLM Mixture-of-Experts и параллельные вычисления
3	Современные возможности и расширения больших языковых моделей (LLM)	Retrieval Augmented Generation (RAG) LLM-агенты Мультимодальные языковые модели Кодовые LLM Альтернативные архитектуры LLM Обработка речи и аудиосигналов

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Риз, Р. Обработка естественного языка на Java : практическое руководство / Р. Риз ; пер. с англ. А. В. Снастина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 266 с. - ISBN 978-5-89818-333-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2102620.

2. Хобсон, Л. Обработка естественного языка в действии : практическое руководство / Л. Хобсон, Х. Ханнес, Х. Коул. - Санкт-Петербург : Питер, 2020. - 576 с. - (Серия «Для профессионалов»). - ISBN 978-5-4461-1371-2. - Текст : электронный. - URL: https://znanium.com/catalog/product/1733506.

3. Васильев, Ю. Обработка естественного языка. Python и spaCy на практике / Ю. Васильев. - Санкт-Петербург : Питер, 2021. - 256 с. - ISBN 978-5-4461-1506-8. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2141633.

Дополнительная литература:

1. Бенгфорт, Б. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка : практическое руководство / Б. Бенгфорт, Р. Билбро, Т. Охеда. - Санкт-Петербург : Питер, 2020. - 368 с. - (Серия «Бестселлеры O'Reilly»). - ISBN 978-5-4461-1153-4. - Текст : электронный. - URL: https://znanium.com/catalog/product/1733693.

2. Николенко, С. И. Глубокое обучение : практическое руководство / С. И. Николенко, А. Кадурин, Е. Архангельская. - Санкт-Петербург : Питер, 2019. - 480 с. - (Серия «Библиотека программиста»). - ISBN 978-5-496-02536-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/1760785.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

— столами и стульями;

— компьютерной техникой;

— механическими калькуляторами;

— специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование"	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное

Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое
Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Natural Language Processing (Обработка естественного языка)» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, аудиторная работа, домашние задания, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Аудиторная работа – активная работа студента на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре студентам рекомендуется заранее ознакомиться с

темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения и аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным, опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал. Использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Natural Language Processing (Обработка естественного языка)»

Оценивание уровня учебных достижений обучающихся по дисциплине (модулю) осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***зачета***, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом:

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Зачтено	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину (модуль). Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами
9	Отлично	Зачтено	
8	Отлично	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
			исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Зачтено	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	Зачтено	
5	Удовлетворительно	Зачтено	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	Зачтено	
3	Не сдан	Не зачтено	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	Не зачтено	
1	Не сдан	Не зачтено	

Дисциплина (модуль) «Natural Language Processing (Обработка естественного языка)» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	60%	14	Набор задач по темам недели
Аудиторная работа	10%	15	Активная работа студента на семинаре

Активность	Вес	Количество	Описание
Зачет	30%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Natural Language Processing (Обработка естественного языка)»: «0,6 × среднее за домашние задания + 0,1 × аудиторная работа + 0,3 × зачет».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1.

Задание 1.

1. Выберите текстовый корпус (например, набор новостных статей или отзывов о продуктах).
2. Выполните предобработку текста, включая:
   - Удаление пунктуации и специальных символов.
   - Приведение текста к нижнему регистру.
   - Стемминг или лемматизацию.
   - Удаление стоп-слов.
3. Реализуйте метод "мешок слов" (Bag of Words) для выделения признаков из вашего корпуса.
4. Создайте таблицу, отображающую частоту слов в вашем наборе данных.

Ожидаемый результат: Отчет, содержащий результаты предобработки, таблицу частоты слов и краткий анализ полученных данных.

Задание 2.

1. Используя библиотеку Gensim, обучите модель Word2Vec на выбранном текстовом корпусе (можно использовать корпус из первого задания).
2. Проведите анализ векторных представлений, выбрав несколько слов и найдите их ближайшие "соседи" в векторном пространстве.
3. Объясните, как векторные представления слов помогают в решении задач NLP.
4. Создайте простую языковую модель на основе N-грамм, используя ваш текстовый корпус, и оцените ее производительность на тестовом наборе данных.

Ожидаемый результат: Отчет с результатами обучения модели Word2Vec, списком ближайших соседей для выбранных слов и кратким анализом языковой модели.

Задание 3.

1. Реализуйте простую рекуррентную нейронную сеть (RNN) для задачи классификации текста (например, классификация отзывов как положительных или отрицательных). Используйте фреймворк TensorFlow или PyTorch.
2. Обучите модель на размеченном наборе данных (например, IMDB для отзывов о фильмах).
3. Сравните производительность RNN с более сложной архитектурой, такой как LSTM или GRU, и проанализируйте результаты.
4. Реализуйте механизм внимания для улучшения производительности вашей модели. Опишите, как механизм внимания помогает в контексте вашей задачи.

Ожидаемый результат: Отчет, содержащий результаты обучения RNN и LSTM/GRU, графики производительности и анализ влияния механизма внимания на результаты.

Домашнее задание 2.

Задание 1.

1. Изучите архитектуру Трансформер, включая механизмы внимания и позиционное кодирование. Создайте визуальную схему, иллюстрирующую основные компоненты модели.
2. Реализуйте простую языковую модель на основе кодировщика Трансформера (например, BERT) с использованием библиотеки Hugging Face Transformers. Обучите модель на задаче классификации текстов (например, классификация отзывов).
3. Проведите оценку производительности модели на тестовом наборе данных и проанализируйте, какие факторы могут влиять на результаты.

Ожидаемый результат: Отчет с визуальной схемой архитектуры, кодом реализации модели, результатами оценки производительности и анализом факторов.

Задание 2.

1. Выберите генеративную языковую модель на основе декодера Трансформера (например, GPT-2 или GPT-3). Изучите, как работает prompt tuning и реализуйте его для вашей модели.
2. Проведите эксперимент, сравнив производительность модели с и без prompt tuning на задаче генерации текста (например, продолжение предложений).
3. Исследуйте методы оптимизации, такие как P-tuning и LoRA. Реализуйте один из этих методов для улучшения производительности вашей модели и оцените результаты.

Ожидаемый результат: Отчет, включающий результаты экспериментов с prompt tuning, сравнение производительности, реализацию метода оптимизации и анализ полученных данных.

Задание 3.

1. Реализуйте RAG (Retrieval-Augmented Generation) систему, используя предобученные модели и набор данных для извлечения информации (например, Wikipedia). Объясните, как система использует механизмы извлечения и генерации.
2. Создайте простую диалоговую систему, реализовав задачи intent detection и slot filling. Используйте библиотеку Rasa или аналогичную для реализации.
3. Проведите тестирование вашей диалоговой системы на наборе пользовательских запросов и проанализируйте результаты, включая точность определения намерений и заполнения слотов.

Ожидаемый результат: Отчет, содержащий описание RAG системы, код реализации диалоговой системы, результаты тестирования и анализ производительности.

Примерные вопросы для подготовки к семинарам

Подготовка к семинару 1.

1. Как происходит процесс векторизации текстов в моделях на основе архитектуры Кодировщика Трансформера, и какие преимущества это дает для задач классификации?
2. В чем основные отличия между BERT и другими предобученными моделями, такими как RoBERTa или DistilBERT, в контексте обработки естественного языка?

3. Как применяются методы ранжирования в информационном поиске с использованием encoder-based моделей, и какие метрики используются для оценки их эффективности?

4. Каким образом обеспечивается мультилингвальность в предобученных языковых моделях, и как это влияет на их производительность при работе с несколькими языками?

5. Какие стратегии доменной адаптации применяются для предобученных моделей типа BERT при переносе на специфические области, такие как медицина или финансы?

Подготовка к семинару 2.

1. Каковы ключевые принципы языкового моделирования в больших языковых моделях, и какие данные используются для их обучения?

2. В чем разница между дообучением (fine-tuning) и полным переобучением больших языковых моделей, и когда предпочтительнее использовать каждый подход?

3. Как работают методы Efficient Transformers и PEFT для оптимизации больших моделей, и как они влияют на оценку эффективности?

4. Какие техники сжатия и оптимизации применяются к LLM, чтобы снизить их размер и ускорить работу, без значительной потери качества?

5. Как архитектура Mixture-of-Experts интегрируется с параллельными вычислениями в больших языковых моделях, и какие преимущества это дает?

Подготовка к семинару 3.

1. Как работает Retrieval Augmented Generation (RAG) в больших языковых моделях, и в каких сценариях это улучшает генерацию ответов?

2. Какие принципы лежат в основе создания LLM-агентов, и как они могут автономно выполнять задачи в реальных приложениях?

3. Как мультимодальные языковые модели интегрируют текст с другими модальностями, такими как изображения или видео, и какие вызовы при этом возникают?

4. В чем особенности кодовых LLM, и как они используются для генерации и анализа программного кода?

5. Как альтернативные архитектуры LLM, такие как те, основанные на обработке речи, расширяют возможности моделей, и какие инструменты применяются для анализа аудиосигналов?

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Укажите способ определения приоритетов в изучении обработки естественного языка.	Анализ	УК-6
2.	Укажите метод обучения больших языковых моделей.	Fine-tuning	ОПК-2
3.	Назовите технику NLP для выявления тем в отзывах продуктов.	Topic modeling	ПК-3
4.	Назовите способ публичного представления результатов NLP.	Доклад	ПК-4
5.	Назовите метод совершенствования собственной деятельности в NLP.	Практика	УК-6
6.	Назовите технику оптимизации трансформеров.	PEFT	ОПК-2
7.	Назовите применение LLM в продуктовой оптимизации.	RAG	ПК-3

8.	Назовите инструмент для представления мультимодальных моделей.	Демонстрация	ПК-4
9.	Укажите аспект самооценки в работе с языковыми моделями.	Точность	УК-6
10.	Укажите модель для мультимодальных задач.	CLIP	ОПК-2
11.	Укажите метод анализа текстов для принятия решений.	Sentiment analysis	ПК-3
12.	Укажите формат демонстрации языковых моделей.	Презентация	ПК-4
13.	Назовите приоритет в обучении больших языковых моделей.	Эффективность	УК-6
14.	Назовите архитектуру для параллельных вычислений в LLM.	Mixture-of-Experts	ОПК-2
15.	Укажите задачу для выявления последствий данных в NLP.	Named entity recognition	ПК-3
16.	Назовите метод презентации научных результатов в NLP.	Визуализация	ПК-4
17.	Укажите способ реализации приоритетов в самообучении NLP.	Планирование	УК-6
18.	Назовите архитектуру трансформера для кодирования текстов.	BERT	ОПК-2
19.	Назовите задачу NLP для оптимизации продуктовой аналитики.	Классификация	ПК-3
20.	Укажите задачу для публичного показа в обработке речи.	Транскрибация	ПК-4