

УТВЕРЖДЕНА

Приказом Ректора АНО ВО
«Центральный университет»
Ивашкевич Е.В.
от «19» января 2024 г. № 0119.37

**Рабочая программа дисциплины (модуля)
«Natural Language Processing (Обработка естественного языка)»
дополнительной профессиональной программы – программы
профессиональной переподготовки «Академия data science»**

Траектория: Backend-разработка

**Москва
2024**

Содержание

1. Краткая характеристика дисциплины (модуля)	3
2. Тематический план	4
3. Содержание дисциплины (модуля)	4
4. Учебно-методическое обеспечение	5
5. Материально-техническое обеспечение	5
6. Методические и оценочные материалы	7

1. Краткая характеристика дисциплины (модуля)

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.

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):

— формирование знаний и понимания по темам: основы предобработки текста и выделения признаков, принципы языкового моделирования и ключевые архитектуры нейронных сетей, применяемых в области NLP (RNN, Трансформер), подходы к машинному переводу, генерации текста и суммаризации, современные методы обучения и оптимизации языковых моделей (LLM), основы работы диалоговых систем и извлечения информации, применение мультимодальных моделей и технологий обработки речи;

— освоение умений: проводить предобработку и векторизацию текста, применять нейронные сети и LLM для задач NLP, обучать и оптимизировать языковые модели, разрабатывать и интегрировать NLP-приложения, оценивать и улучшать качество моделей;

— формирование навыков владения инструментами Python для NLP (HuggingFace, nltk, pytorch), методами обучения и оптимизации языковых моделей, навыками работы с мультимодальными системами и Text-to-speech.

2. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоя тельная работа	
Лекции	Семинары (практичес кие занятия)					
1	Основы обработки текста и векторные представления	8	8		23	Домашние задания, Тесты
2	Основные архитектуры нейронных сетей для задач NLP	8	8		23	Домашние задания, Тесты
3	Языковые модели на основе архитектуры Трансформер	9	9		24	Домашние задания, Тесты
4	Прикладные задачи NLP	8	9		24	Домашние задания, Тесты
	<i>Зачет с оценкой</i>			4		
	Итого:	33	34	4	94	
	Объем дисциплины (модуля) (в ак. ч.)	165				

3. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Основы обработки текста и векторные представления	Введение в анализ текстов, базовые методы предобработки и выделения признаков. Векторные представления слов.
2	Основные архитектуры нейронных сетей для задач NLP	Языковое моделирование. Рекуррентные нейронные сети (RNN). Машинный перевод и механизм внимания.
3	Языковые модели на основе архитектуры Трансформер	Архитектура Трансформер. языковые модели на основе архитектуры кодировщик Трансформера. Генеративные языковые модели на основе архитектуры декодеровщик Трансформера (LLM, prompt tuning, RLHF). Оптимизация языковых моделей (P-tuning, LoRA, квантизация).
4	Прикладные задачи NLP	RAG системы и ранжирование. Извлечение именованных сущностей и отношений. Диалоговые системы (intent detection, slot filling). Задача суммаризации. Мультимодальные модели. NLP для кода. Text-to-speech.

4. Учебно-методическое обеспечение

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

Каждый слушатель 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.

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

Основная литература:

1. 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.

2. Николенко С., Кадури А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. — 480 с.: ил. — (Серия «Библиотека программиста»). — ISBN 978-5-496-02536-2.

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.

5. 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

6. 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, практические занятия, домашние задания, тесты, 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 удастся, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Участие в семинаре (практическом занятии) – активная работа слушателя на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре слушатели 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.

Тест – особая форма проверки знаний. Проводится после освоения одной или нескольких тем и свидетельствует о качестве понимания основных понятий изучаемого материала. Тестовые задания составлены к ключевым понятиям, основным разделам, важным терминологическим категориям изучаемой дисциплины (модуля).

Для подготовки к тесту необходимо знать терминологический аппарат дисциплины (модуля), понимать смысл научных категорий и уметь их использовать в профессиональной

лексике. Владение понятийным аппаратом, включённым в тестовые задания, позволяет преподавателю быстро проверить уровень понимания слушателями важных методологических категорий.

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

В процессе самостоятельной работы слушателя взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи слушателя 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)

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 с оценкой*.

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	Слушатель 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. Слушатель хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
9	Отлично	Зачтено	
8	Отлично	Зачтено	
7	Хорошо	Зачтено	Слушатель обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на
6	Хорошо	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
			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. Слушатель хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
5	Удовлетворительно	Зачтено	Слушатель 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. Слушатель способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	Зачтено	
3	Не сдан	Не зачтено	Слушатель 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
Накопительная оценка			
Домашние задания	60%	2	Набор задач по темам недели
Тесты		7	Письменная работа с набором задач, которые нужно решить за ограниченное время
Промежуточная аттестация			
Зачет с оценкой	40%	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)»: « $0,6 \times$ накопительная оценка ($0,6 \times$ среднее за домашние задания + $0,4 \times$ среднее за тесты) + $0,4 \times$ зачет с оценкой».

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 задания по тестам

Тест 1.

Вопрос 1.

Какие из перечисленных видов векторных представлений по построению имеют разреженную структуру векторов?

- A. Fasttext
- B. One-hot
- C. word2vec
- D. Glove

Ответ: B.

Вопрос 2.

Какие преимущества имеет FastText перед word2vec? Выберите **ВСЕ** правильные варианты:

- A. Обработка OOV-слов
- B. Существенное сокращение потребления памяти
- C. Борьба с опечатками
- D. Ускоренное обучение за счет использования хэш-таблиц

Ответ: А, С.

Вопрос 3.

Сколько векторов поставит в соответствие многозначному слову (например, слову "замок") модель класса word2vec?

- А. один вектор
- В. столько векторов, сколько у слова смыслов
- С. количество векторов слова – это гиперпараметр, который можно задать
- Д. модель не сойдется, вектор многозначного слова будет случайно

Ответ: А.

Вопрос 4.

Сколько обучаемых матриц весов в простейшем варианте архитектуры CBOW?

- А. 1
- В. 2
- С. 3
- Д. 4

Ответ: В.

Вопрос 5.

Примените операцию Softmax к вектору (1, 2, 0, 1). Какое максимальное значение у полученного в результате применения операции вектора? Считать $\exp(1) = 2.7$, $\exp(2) = 7.4$.

Ответ укажите с точностью до 3-х знаков после запятой, прочие удалите без округления.

- А. 0,536
- В. 0,578
- С. 1
- Д. 7,4

Ответ: А.

Тест 2.

Вопрос 1.

Выберите **ВСЕ** верные утверждения об исходной модели Трансформер, используемой для машинного перевода:

- А. В модели Трансформер есть рекуррентные слои
- В. Модель Трансформер является моделью класса кодировщик-декодировщик
- С. В модели Трансформер есть механизм внимания
- Д. В модели Трансформер есть механизм памяти
- Е. Все утверждения верны

Ответ: В, С.

Вопрос 2.

Позиционные эмбединги в модели Трансформер...

- А. помогают учитывать порядок слов
- В. помогают учитывать текущий номер слоя
- С. помогают учитывать текущий номер головы модели
- Д. помогают учитывать частоту слова в большом корпусе

Ответ: А.

Вопрос 3.

Чем отличается слой Encoder-Decoder Attention в декодировщике от Self-Attention?

- А. Матрицы запроса (query) и ключа (key) берутся из кодировщика
- В. Матрицы запроса (query) и значения (value) берутся из кодировщика
- С. Матрицы ключа (key) и значения (value) берутся из кодировщика
- Д. Входная последовательность берется из кодировщика

Ответ: С.

Вопрос 4.

Сколько матриц весов будет суммарно в слоях self-attention кодировщика, если в модели 8

голов внимания и 6 слоев кодировщика?

- A. 3
- B. 8
- C. 24
- D. 25
- E. 144
- F. 150

Ответ: F.

Вопрос 5.

Если на вход трансформерному слою кодировщика с 8 головами внимания подается n векторов слов размерности m . Какого размера матрица получится на выходе слоя?

- A. $[n/8, m]$
- B. $[n, m/8]$
- C. $[n/8, m/8]$
- D. $[n, m]$

Ответ: D.

Тест 3.

Вопрос 1.

Допустим, модель ранжирования выдала 5 топовых документов на запрос q в следующем порядке: [doc1, doc3, doc2, doc4, doc5]. При этом идеальная выдача для запроса q вместе со скорями релевантности: [doc1 - 3, doc2 - 2, doc5 - 1, doc3 - 0, doc4 - 0]. Чему будет равна метрика $nDCG@3$ для выдачи модели? Логарифмы и результаты деления округляйте до 3 знака после запятой.

- A. 0.840
- B. 0.921
- C. 4
- D. 4.762

Ответ: A.

Вопрос 2.

Выберите все верные утверждения:

- A. Модель T5 имеет архитектуру кодировщик-декодировщик
- B. Модель FlanT5 – это инструктивно обученная модель T5
- C. Модель T5 предназначена исключительно для задачи суммаризации
- D. Модель BART предобучалась на задаче перевода
- E. В модели BART не используется маскированная языковая модель
- F. В модели BART больше параметров, чем в модели BERT

Ответ: A, B, F.

Вопрос 3.

Обратимся к алгоритму экстрактивной суммаризации TextRank. Упорядочьте его шаги в правильном порядке: а) Построение графа близости между предложениями; б) Разбиение текста на предложения; в) Вычисление меры центральности в графе предложений; г) Ранжирование предложений; д) Определение близости между предложениями; е) Векторизация предложений.

- A. a b c d e f
- B. f b e d c a
- C. b f e a c d
- D. b a f e c d

Ответ: C.

Вопрос 4.

В теме диалоговых систем были рассмотрены задачи intent detection и slot filling. К каким двум задачам NLP сводятся эти две задачи?

- A. Классификация текста
- B. Суммаризация
- C. Распознавание именованных сущностей
- D. Извлечение отношений

Ответ: A, C.

Вопрос 5.

В тексте необходимо выделить сущности: локация, персона, дата, организация. Если использовать схему IOB при разметке и модель с архитектурой BERT + полносвязный слой со скрытой размерностью векторов n , то какова будет размерность последнего слоя?

- A. $[n, 2]$
- B. $[n, 4]$
- C. $[n, 5]$
- D. $[n, 9]$

Ответ: D.

Примерный список вопросов к зачету с оценкой

1. Что такое предобработка текста и почему она важна в NLP?
2. Перечислите основные шаги предобработки текстовых данных.
3. Каковы различия между стеммингом и лемматизацией?
4. Что такое "мешок слов" (Bag of Words) и как он используется в NLP?
5. Каковы основные методы выделения признаков из текстов?
6. Что такое векторные представления слов и почему они важны для NLP?
7. Объясните, как работает метод Word2Vec.
8. Каково назначение модели GloVe и какие преимущества она имеет?
9. В чем разница между статическими и динамическими векторными представлениями слов?
10. Что такое языковая модель и как она используется в NLP?
11. Объясните принцип работы рекуррентных нейронных сетей (RNN).
12. Каковы основные проблемы, с которыми сталкиваются RNN при обучении?
13. Что такое LSTM и в чем его преимущества по сравнению с обычными RNN?
14. Как работает машинный перевод на основе нейронных сетей?
15. Объясните концепцию механизма внимания (attention mechanism).
16. В чем заключается отличие между механизмом внимания и традиционными методами перевода?
17. Какова основная архитектура модели Трансформер?
18. Что такое позиционное кодирование и зачем оно нужно в Трансформере?
19. Какова роль многоголового внимания (multi-head attention) в Трансформере?
20. Как работает кодировщик в архитектуре Трансформера?
21. Каковы основные применения языковых моделей на основе кодировщика Трансформера?
22. Что такое генеративные языковые модели и как они работают?
23. Объясните концепцию prompt tuning и его применение в LLM.
24. Как работает RLHF (Reinforcement Learning from Human Feedback) в контексте языковых моделей?
25. Что такое P-tuning и как он улучшает производительность языковых моделей?
26. Объясните, как работает метод LoRA (Low-Rank Adaptation).
27. Что такое квантизация и как она помогает оптимизировать языковые модели?
28. Как работают RAG (Retrieval-Augmented Generation) системы?
29. Каковы основные подходы к ранжированию результатов в NLP?
30. Что такое извлечение именованных сущностей и как оно применяется в реальных задачах?