ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Приложение № 1
к приказу ректора
АНО ВО «Центральный университет»
от «15» августа 2024 г. № 0815.50

Автономная некоммерческая организация высшего образования

«Центральный университет»

УТВЕРЖДАЮ

Ректор АНО ВО «Центральный университет»

Е.В. Ивашкевич

ДОПОЛНИТЕЛЬНАЯ ПРОФЕССИОНАЛЬНАЯ ПРОГРАММА – ПРОГРАММА ПОВЫШЕНИЯ КВАЛИФИКАЦИИ «ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА»

Трудоемкость обучения: 112 ак. часов

Москва

2024

Электронный документ

1. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

1.1. Общие положения

Настоящая дополнительная профессиональная программа – программа повышения квалификации «Обработка естественного языка» (далее – программа повышения квалификации) разработана на основании Федерального закона от 29 декабря 2012 г. № 273-ФЗ «Об образовании в Российской Федерации», приказа Минобрнауки России от 01 июля 2013 г. № 499 «Об утверждении Порядка организации и осуществления образовательной деятельности по дополнительным профессиональным программам» и приказа Минобрнауки России от 11 октября 2023 г. № 1678 «Об утверждении Правил применения организациями, осуществляющими образовательную деятельность, электронного обучения, дистанционных образовательных технологий при реализации образовательных программ».

Программа повышения квалификации реализуется в Автономной некоммерческой организации высшего образования «Центральный университет» (далее – АНО ВО «Центральный университет»).

Разработчик программы: Алимова Ильсеяр Салимовна, преподаватель АНО ВО «Центральный университет».

Программа повышения квалификации разработана в инициативном порядке.

Программа реализуется на русском языке.

1.2. Цель реализации программы

Программа повышения квалификации нацелена на совершенствование и (или) получение слушателем новой компетенции, необходимой для профессиональной деятельности, и (или) повышение профессионального уровня знаний в рамках имеющейся у слушателя квалификации в сфере обработки естественного языка, включая методы машинного обучения, глубокого обучения, анализ текста и разработку интеллектуальных систем для обработки и понимания естественного языка.

1.3. Категории обучающихся

Основными категориями обучающихся, на которых рассчитана программа повышения квалификации, являются студенты 3 и 4 курса бакалавриата, стремящиеся к развитию карьеры в области data science; руководители и специалисты IT-подразделений, желающие углубить знания и навыки в области обработки естественного языка; специалисты с технической подготовкой из другой сферы, планирующие изменить свой карьерный трек.

1.4. Требования к уровню подготовки поступающего на обучение, необходимому для освоения программы

Лица, имеющие среднее профессиональное и (или) высшее образование. Наличие указанного образования должно подтверждаться документом государственного или установленного образца.

Слушатель программы может быть студентом старших курсов программ высшего образования – программ бакалавриата или специалитета. Данный факт подтверждается предоставлением справки с места обучения по программе высшего образования.

1.5. Перечень нормативных документов, определяющих квалификационные требования к выпускнику программы

— Федеральный государственный образовательный стандарт по направлению подготовки (специальности) 02.04.01 «Математика и компьютерные науки» и уровню высшего образования магистратура, утвержденный приказом Минобрнауки России от 23 августа 2017 г. № 810;

— Профессиональный стандарт «Программист», утвержденный приказом Министерства труда и социальной защиты Российской Федерации от 18 ноября 2013 г. № 679н.

1.6. Планируемые результаты обучения

а) Перечень компетенций, совершенствование или получение которых осуществляется в результате обучения:

- Способность определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки;
- Способность создавать и исследовать новые математические модели в естественных науках, совершенствовать и разрабатывать концепции, теории и методы;
- Способность решать задачи профессиональной деятельности, формулировать результат, увидеть следствия полученного результата;
- Способность передавать результат решенных прикладных задач в виде конкретных рекомендаций, выраженных в терминах области машинного обучения.

б) Квалификационные требования к выпускнику программы:

Слушатель в результате освоения программы будет способен выполнять следующие должностные обязанности, приведённые в «Квалификационном справочнике должностей руководителей, специалистов и других служащих», утверждённом постановлением Минтруда России от 21 августа 1998 г. № 37.

Инженер-программист (Программист)

На основе анализа математических моделей и алгоритмов решения экономических и других задач разрабатывает программы, обеспечивающие возможность выполнения алгоритма и соответственно поставленной задачи средствами вычислительной техники, проводит их тестирование и отладку. Разрабатывает технологию решения задачи по всем этапам обработки информации. Осуществляет выбор языка программирования для описания алгоритмов и структур данных. Определяет информацию, подлежащую обработке средствами вычислительной техники, ее объемы, структуру, макеты и схемы ввода, обработки, хранения и вывода, методы ее контроля. Выполняет работу по подготовке программ к отладке и проводит отладку. Определяет объем

и содержание данных контрольных примеров, обеспечивающих наиболее полную проверку соответствия программ их функциональному назначению. Осуществляет запуск отлаженных программ и ввод исходных данных, определяемых условиями поставленных задач. Проводит корректировку разработанной программы на основе анализа выходных данных. Разрабатывает инструкции по работе с программами, оформляет необходимую техническую документацию. Определяет возможность использования готовых программных продуктов. Осуществляет сопровождение внедренных программ и программных средств. Разрабатывает и внедряет системы автоматической проверки правильности программ, типовые и стандартные программные средства, составляет технологию обработки информации. Выполняет работу по унификации и типизации вычислительных процессов. Принимает участие в создании каталогов и картотек стандартных программ, в разработке форм документов, подлежащих машинной обработке, в проектировании программ, позволяющих расширить область применения вычислительной техники.

в) Идентификаторы достижения компетенций:

Слушатель, освоивший программу повышения квалификации, должен:

знать:

— 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.

уметь:

— 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;

— применять мультимодальные модели и технологии обработки речи.

владеть:

— инструментами Python для NLP (HuggingFace, nltk, pytorch);

— методами обучения и оптимизации языковых моделей;

— навыками работы с мультимодальными системами и Text-to-speech.

1.7. Трудоемкость программы

Нормативная трудоемкость обучения по данной программе – 112 академических часов, включая все виды контактной и самостоятельной работы слушателя.

1.8. Форма и сроки обучения

Обучение по программе повышения квалификации осуществляется в очной форме с использованием дистанционных образовательных технологий и (или) электронного обучения.

Формат обучения на программе – гибридный. Учебные занятия проходят в очном формате, при этом слушателям предоставляется возможность участвовать как очно, так и дистанционно в режиме онлайн.

Минимальный срок обучения на программе составляет 3 месяца.

1.9. Режим занятий

Длительность одного занятия – 2 академических часа.

Для всех занятий академический час устанавливается продолжительностью 45 минут.

2. ПРОГРАММА УЧЕБНОГО КУРСА

2.1. Учебный план программы повышения квалификации «Обработка естественного языка»

Продолжительность обучения – 112 ак. часов.

Форма обучения – очная.

№ п/п	Наименование дисциплин учебного курса	Всего, академ. ч.	Контактная работа (академ. ч.)		Самост. работа (академ. ч.)	Формы аттестации
			лекции	семинары (практич. занятия)		
1.	Основы обработки текста и векторные представления	14	4	4	6	Зачет
2.	Основные архитектуры нейронных сетей для задач NLP	21	6	6	9	Зачет
3.	Языковые модели на основе архитектуры Трансформер	21	6	6	9	Зачет
4.	Прикладные задачи NLP	49	14	14	21	Зачет
	ВСЕГО:	**105**	**30**	**30**	**45**	
	Итоговая аттестация	**7**		**7**		**Экзамен**
	ИТОГО:	**112**	**30**	**37**	**45**	

2.2. Учебно-тематический план программы повышения квалификации «Обработка естественного языка»

№ п/п	Наименование дисциплин и тем	Всего, академ. ч.	Контактная работа (академ. ч.)		Самост. работа (академ. ч.)	Формы аттестации и контроля знаний
			лекции	семинары (практич. занятия)		
1.	**Основы обработки текста и векторные представления**	**14**	**4**	**4**	**6**	**Зачет**
1.1.	Введение в анализ текстов, базовые методы предобработки и выделения признаков	7	2	2	3	Домашнее задание
1.2.	Векторные представления слов	7	2	2	3	Домашнее задание
2.	**Основные архитектуры нейронных сетей для задач NLP**	**21**	**6**	**6**	**9**	**Зачет**
2.1.	Языковое моделирование	7	2	2	3	Домашнее задание
2.2.	Рекуррентные нейронные сети (RNN)	7	2	2	3	Домашнее задание
2.3.	Машинный перевод и механизм внимания	7	2	2	3	Домашнее задание
3.	**Языковые модели на основе архитектуры Трансформер**	**21**	**6**	**6**	**9**	**Зачет**
3.1.	Архитектура «Трансформер». Языковые модели на основе архитектуры кодировщик Трансформера	7	2	2	3	Домашнее задание
3.2.	Генеративные языковые модели на основе архитектуры декодеровщик Трансформера (LLM, prompt tuning, RLHF)	7	2	2	3	Домашнее задание
3.3.	Оптимизация языковых моделей (P-tuning, LoRA, квантизация)	7	2	2	3	Домашнее задание
4.	**Прикладные задачи NLP**	**49**	**14**	**14**	**21**	**Зачет**
4.1.	RAG системы и ранжирование	7	2	2	3	Домашнее задание
4.2.	Диалоговые системы (intent detection, slot filling)	7	2	2	3	Домашнее задание
4.3.	Извлечение именованных сущностей и отношений	7	2	2	3	Домашнее задание

4.4.	Задача суммаризации	7	2	2	3	Домашнее задание
4.5.	Мультимодальные модели	7	2	2	3	Домашнее задание
4.6.	NLP для кода	7	2	2	3	Домашнее задание
4.7.	Text-to-speech	7	2	2	3	Домашнее задание
	ВСЕГО:	105	30	30	45	
	Итоговая аттестация	7		7		Экзамен
	ИТОГО:	112	30	37	45	

2.3. Календарный учебный график

Дисциплины учебного курса	**Наименование темы дисциплин**	**Академ. часов**	**Учебные месяцы**		
			1	**2**	**3**
Основы обработки текста и векторные представления	Введение в анализ текстов, базовые методы предобработки и выделения признаков	7	7		
	Векторные представления слов	7	7		
Основные архитектуры нейронных сетей для задач NLP	Языковое моделирование	7	7		
	Рекуррентные нейронные сети (RNN)	7	7		
	Машинный перевод и механизм внимания	7	7		
Языковые модели на основе архитектуры Трансформер	Архитектура «Трансформер». Языковые модели на основе архитектуры кодировщик Трансформера	7		7	
	Генеративные языковые модели на основе архитектуры декодеровщик Трансформера (LLM, prompt tuning, RLHF)	7		7	
	Оптимизация языковых моделей (P-tuning, LoRA, квантизация)	7		7	
Прикладные задачи NLP	RAG системы и ранжирование	7		7	
	Диалоговые системы (intent detection, slot filling)	7		7	
	Извлечение именованных сущностей и отношений	7		7	

	Задача суммаризации	7			7
	Мультимодальные модели	7			7
	NLP для кода	7			7
	Text-to-speech	7			7
Итоговая аттестация		7			7

3. РАБОЧАЯ ПРОГРАММА КУРСА

3.1. Содержание курса

№ п/п	Наименование дисциплины	Наименование темы	Содержание темы (тезисно)
1.	Основы обработки текста и векторные представления	Введение в анализ текстов, базовые методы предобработки и выделения признаков	Токенизация текста. Стемминг и лемматизация. TF-IDF векторизация.
		Векторные представления слов	Word embeddings. Word2Vec модель. Контекстные векторы.
2.	Основные архитектуры нейронных сетей для задач NLP	Языковое моделирование	Вероятностное предсказание. N-граммы. Оценка перплексии.
		Рекуррентные нейронные сети (RNN)	Последовательная обработка. Проблема исчезающего градиента. LSTM и GRU архитектуры.
		Машинный перевод и механизм внимания	Encoder-Decoder модель. Attention механизм. Seq2Seq обучение.
3.	Языковые модели на основе архитектуры Трансформер	Архитектура «Трансформер». Языковые модели на основе архитектуры кодировщик Трансформера	Self-attention механизм. BERT модель. Fine-tuning задач.
		Генеративные языковые модели на основе архитектуры декодеровщик Трансформера (LLM, prompt tuning, RLHF)	GPT архитектура. Prompt engineering. Reinforcement learning from human feedback.
		Оптимизация языковых моделей (P-tuning, LoRA, квантизация)	Parameter-efficient tuning. Low-rank adaptation. Model compression.
4.	Прикладные задачи NLP	RAG системы и ранжирование	Retrieval-Augmented Generation. Информационный поиск. Ранжирование релевантности.
		Диалоговые системы (intent detection, slot filling)	Классификация намерений. Извлечение слотов. Контекстное управление диалогом.
		Извлечение именованных сущностей и отношений	Named Entity Recognition. Relation extraction. Аннотация сущностей.
		Задача суммаризации	Extractive summarization. Abstractive summarization. Оценка качества резюме.
		Мультимодальные модели	Текст-изображение интеграция. CLIP модель. Multimodal embeddings.

		NLP для кода	Code tokenization. Syntax-aware модели. Автодополнение кода.
		Text-to-speech	Синтез речи. WaveNet архитектура. Голосовые модели.

3.2. Методические указания для обучающихся по освоению курса

В процессе изучения программы повышения квалификации «Обработка естественного языка» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, практические занятия, 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 (практическое занятие) — это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где слушатели 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.

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 слушателей, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины.

В процессе самостоятельной работы слушатели 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.

3.3. Текущий контроль успеваемости обучающихся по дисциплинам

Домашние задания по дисциплине

«Основы обработки текста и векторные представления»

1. Возьмите небольшой текстовый датасет (например, отзывы из IMDB). Реализуйте токенизацию с помощью NLTK и spaCy, затем очистите текст от стоп-слов и пунктуации. Сравните результаты и визуализируйте частоту слов с помощью wordcloud. Ожидаемый результат: отчет с примерами и графиками.

2. Используя NLTK, примените стемминг (PorterStemmer) и лемматизацию (WordNetLemmatizer) к корпусу новостных статей. Оцените влияние на качество предобработки и сравните с spaCy. Ожидаемый результат: анализ различий на примерах текста.

3. На датасете текстовых документов (например, 20 Newsgroups) постройте TF-IDF векторы с помощью scikit-learn. Найдите топ-10 наиболее информативных слов для каждой категории и визуализируйте их. Ожидаемый результат: код, векторы и heatmap.

4. Обучите модель Word2Vec на корпусе (например, Wikipedia dumps) с помощью Gensim. Найдите наиболее близкие слова к заданному термину

(например, "king") и визуализируйте эмбеддинги с t-SNE. Ожидаемый результат: модель, примеры аналогий и график.

5. Сравните TF-IDF и Word2Vec на задаче классификации текста (используйте LogisticRegression). Оцените точность на тестовом наборе и объясните различия. Ожидаемый результат: сравнительный анализ с метриками и выводами.

Домашние задания по дисциплине

«Основные архитектуры нейронных сетей для задач NLP»

1. Реализуйте простую RNN с PyTorch для генерации текста на основе корпуса (например, Shakespeare). Обучите модель и сгенерируйте 100 символов. Ожидаемый результат: код, сгенерированный текст и график потерь.

2. Постройте LSTM-модель с TensorFlow для классификации настроений на датасете IMDb. Включите embedding слой и сравните с базовой RNN. Ожидаемый результат: модель, точность на тесте и анализ overfitting.

3. Используйте CNN с PyTorch для задачи классификации новостей (AG News). Экспериментируйте с фильтрами и pooling. Ожидаемый результат: архитектура, результаты и визуализация фильтров.

4. Реализуйте BiLSTM с CRF для named entity recognition на датасете CoNLL-2003. Оцените F1-score. Ожидаемый результат: модель, предсказания и метрики.

5. Обучите все три архитектуры на задаче предсказания следующего слова. Сравните их производительность по perplexity и времени обучения. Ожидаемый результат: сравнительный отчет с графиками и выводами.

Домашние задания по дисциплине

«Языковые модели на основе архитектуры Трансформер»

1. Реализуйте простой Transformer с PyTorch для перевода английского на французский (на датасете WMT). Обучите на подмножестве и оцените BLEU. Ожидаемый результат: модель, примеры переводов и метрики.

2. Используйте предобученный BERT (из Hugging Face) для задачи классификации эмоций в твитах. Fine-tune модель и сравните с baseline. Ожидаемый результат: fine-tuned модель, accuracy и анализ attention.

3. С помощью GPT-2 (Hugging Face) сгенерируйте продолжения текстов на заданные промпты. Экспериментируйте с temperature и length. Ожидаемый результат: примеры генераций и анализ качества.

4. Реализуйте self-attention слой в PyTorch и примените к задаче summarization. Визуализируйте attention weights. Ожидаемый результат: код, summaries и heatmap attention.

5. Возьмите RoBERTa и fine-tune на датасете SQuAD для QA. Оцените EM и F1. Ожидаемый результат: модель, примеры ответов и метрики.

Домашние задания по дисциплине

«Прикладные задачи NLP»

1. Постройте модель для классификации спама в SMS с помощью BERT. Оцените precision и recall. Ожидаемый результат: модель, confusion matrix и отчет.

2. Используйте spaCy или fine-tune BERT для извлечения сущностей из новостных статей. Визуализируйте результаты. Ожидаемый результат: tagged текст и анализ ошибок.

3. Реализуйте seq2seq модель с attention для перевода с английского на русский. Оцените BLEU на тесте. Ожидаемый результат: модель, переводы и метрики.

4. С помощью GPT-3 API или локальной модели сгенерируйте описания изображений на основе промптов. Оцените coherence. Ожидаемый результат: примеры и субъективный анализ.

5. Примените VADER или модель на основе LSTM к датасету отзывов. Создайте дашборд с графиками распределения настроений. Ожидаемый результат: анализ, графики и insights.

4. ОРГАНИЗАЦИОННО-ПЕДАГОГИЧЕСКИЕ УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ

4.1.Требования к кадровым условиям реализации программы

Реализация программы обеспечивается штатными руководящими и научно-педагогическими работниками АНО ВО «Центральный университет», а также внешними совместителями, работающими по договорам гражданско-правового характера. Научно-педагогические работники, осуществляющие преподавание данной программы, имеют образование, соответствующее профилю курса, или конкретный опыт реализации разработок и иной формы практической деятельности по направлению курса.

4.2. Требования к материально-техническим условиям реализации программы

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов учебной деятельности, предусмотренных учебным планом.

Помещения 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 обеспечивается в учебных аудиториях, оснащенных:

— столами и стульями;

— компьютерной техникой;

— 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 к 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, 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

4.3. Учебно-методическое обеспечение программы

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., Кадурин 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. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

5.1. Формы контроля

Критерии получения уровня и оценивания сформированности компетенций по программе повышения квалификации «Обработка естественного языка»

Оценивание уровня учебных достижений обучающихся по программе осуществляется в виде текущего контроля успеваемости, промежуточных аттестаций и итоговой аттестации.

Промежуточные аттестации проводятся в форме ***зачета***. Формат проведения – тестирование.

К итоговой аттестации допускается слушатель, не имеющий задолженности и в полном объеме выполнивший учебный план по программе повышения квалификации.

Итоговая аттестация по программе осуществляется в форме ***экзамена,*** при этом проводится оценка компетенций, сформированных по курсу.

5.2. Система оценивания результатов обучения по курсу

Для оценивания текущего контроля успеваемости, промежуточных аттестаций и итоговой 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 курсу
10	Отлично	Зачтено	Обучающийся полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет курс. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и
9	Отлично	Зачтено	
8	Отлично	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по курсу
			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. Обучающийся хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты курса с практическими задачами.
7	Хорошо	Зачтено	Обучающийся 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. Обучающийся хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	Зачтено	
5	Удовлетворительно	Зачтено	Обучающийся обладает базовыми знаниями по курсу, но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути
4	Удовлетворительно	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по курсу
			вопросов. Обучающийся способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
3	Не сдан	Не зачтено	Обучающийся 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	

5.3. Примеры оценочных материалов по промежуточным аттестациям

Промежуточные аттестации проходят в формате тестирования, состоящего из 20 закрытых вопросов. Ниже приведены критерии оценивания, а также примерные оценочные материалы по промежуточным аттестациям.

Критерии оценивания:

Всего: 20 вопросов.

Максимальная сумма баллов за тест – 20.

Соотношение баллов за тест с оценками:

Баллы	Оценка (по 10-балльной шкале)	Оценка за зачет
19 – 20	10	Зачтено
17 – 18	9	Зачтено
15 – 16	8	Зачтено
13 – 14	7	Зачтено
11 – 12	6	Зачтено
9 – 10	5	Зачтено
7 – 8	4	Зачтено
5 – 6	3	Зачтено
3 – 4	2	Зачтено
0 – 2	1	Не зачтено

Тестирование по дисциплине

«Основы обработки текста и векторные представления»

1. Что такое токенизация в NLP?

A) Процесс приведения слов к их базовой форме

B) Разделение текста на отдельные единицы, такие как слова или предложения

C) Преобразование текста в числовые векторы

D) Удаление стоп-слов из текста

Правильный ответ: B

2. Какой метод используется для приведения слов к их корню, игнорируя грамматические окончания?

A) Лемматизация

B) Стемминг

C) TF-IDF

D) Word2Vec

Правильный ответ: B

3. Что представляет собой TF-IDF в векторных представлениях?

A) Модель, которая обучает эмбеддинги на основе контекста слов

B) Метод, который вычисляет важность слова в документе относительно корпуса

C) Процесс генерации текста с помощью нейронных сетей

D) Архитектура для обработки последовательностей

Правильный ответ: B

4. В чем основное отличие Word2Vec от TF-IDF?

A) TF-IDF не учитывает контекст, а Word2Vec создает плотные векторы на основе семантического сходства

B) Word2Vec используется только для классификации текста

C) TF-IDF обучает эмбеддинги, а Word2Vec – нет

D) Они идентичны по принципу работы

Правильный ответ: A

5. Что такое стоп-слова в предобработке текста?

A) Слова с высокой частотой, которые часто удаляют для снижения шума

B) Специальные символы, такие как пунктуация

C) Векторы, представляющие слова в многомерном пространстве

D) Метод для генерации новых слов

Правильный ответ: A

Тестирование по дисциплине

«Основные архитектуры нейронных сетей для задач NLP»

1. Что такое RNN и в чем ее основное преимущество для обработки последовательностей?

A) Сеть, которая обрабатывает данные параллельно без учета порядка

B) Рекуррентная сеть, способная учитывать последовательную зависимость в данных

C) Сеть с сверточными слоями для анализа изображений

D) Модель, основанная исключительно на attention-механизме

Правильный ответ: B

2. В чем отличие LSTM от базовой RNN?

A) LSTM не имеет памяти для долгосрочных зависимостей

B) LSTM использует ворота для контроля потока информации и борьбы с vanishing gradient

C) LSTM предназначена только для генерации текста

D) Они идентичны по архитектуре

Правильный ответ: B

3. Как CNN применяется в NLP?

A) Для обработки последовательностей с помощью рекурсии

B) Для извлечения локальных паттернов в тексте через фильтры и pooling

C) Для генерации текста на основе трансформеров

D) Только для задач машинного зрения

Правильный ответ: B

4. Что такое BiLSTM и зачем она используется?
A) Однонаправленная LSTM для быстрого обучения

B) Двунаправленная LSTM, которая обрабатывает последовательность в обоих направлениях для лучшего контекста

C) Сеть для классификации изображений

D) Модель без рекуррентных связей

Правильный ответ: B

5. Какая проблема часто возникает в RNN при обучении на длинных последовательностях?

A) Overfitting

B) Vanishing gradient problem

C) Слишком быстрая сходимость

D) Отсутствие параллелизации

Правильный ответ: B

Тестирование по дисциплине

«Языковые модели на основе архитектуры Трансформер»

1. Что является ключевым компонентом архитектуры Transformer?

A) Рекуррентные слои

B) Attention-механизм для моделирования зависимостей между словами

C) Сверточные фильтры

D) Только энкодер без декодера

Правильный ответ: B

2. В чем основное отличие BERT от GPT?

A) BERT – это только декодер, GPT – энкодер

B) BERT использует masked language modeling, GPT - generative pre-training

C) Они идентичны по архитектуре

D) BERT предназначен только для генерации текста

Правильный ответ: B

3. Как работает self-attention в Transformer?

A) Он фокусируется только на предыдущих словах в последовательности

B) Он вычисляет веса внимания для всех пар слов в предложении одновременно

C) Он игнорирует контекст и работает с отдельными словами

D) Он используется только в декодере

Правильный ответ: B

4. Что такое fine-tuning в контексте моделей вроде BERT?

A) Обучение модели с нуля на большом корпусе

B) Донастройка предобученной модели на специфической задаче с меньшим датасетом

C) Удаление attention-слоев для ускорения

D) Преобразование модели в CNN

Правильный ответ: B

5. Какой тип модели Transformer используется для машинного перевода?

A) Только энкодер (как BERT)

B) Только декодер (как GPT)

C) Полный Transformer с энкодером и декодером

D) Без attention-механизма

Правильный ответ: C

Тестирование по дисциплине «Прикладные задачи NLP»

1. Что такое named entity recognition (NER)?

A) Генерация нового текста на основе промпта

B) Извлечение и классификация именованных сущностей, таких как имена, даты в тексте

C) Перевод текста с одного языка на другой

D) Анализ настроений в отзывах

Правильный ответ: B

2. Какая задача NLP включает классификацию текста на категории, такие как спам или не-спам?

A) Машинный перевод

B) Text classification

C) Question answering

D) Text summarization

Правильный ответ: B

3. Что такое sentiment analysis?

A) Извлечение ключевых фраз из текста

B) Определение эмоционального тона текста (положительный, отрицательный)

C) Генерация вопросов на основе текста

D) Преобразование речи в текст

Правильный ответ: B

4. В чем заключается задача machine translation?

A) Анализ настроений

B) Автоматический перевод текста с одного языка на другой

C) Извлечение сущностей

D) Классификация документов

Правильный ответ: B

5. Что такое text summarization?

A) Генерация длинного текста из короткого

B) Создание краткого резюме основного содержания текста

C) Перевод текста в аудио

D) Распознавание речи

Правильный ответ: B

5.4. Примеры оценочных материалов по итоговой аттестации

Форма итоговой аттестации: экзамен.

Примерный перечень вопросов для подготовки к экзамену:

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?

6. ВЫХОДНЫЕ ДОКУМЕНТЫ

Лицам, успешно освоившим соответствующую программу повышения квалификации и прошедшим итоговую аттестацию, выдается удостоверение о повышении квалификации. Удостоверение выдается на бланке, являющемся защищенной от подделок полиграфической продукцией, образец которого самостоятельно установлен образовательным учреждением.

Лицам, не прошедшим итоговую аттестацию или получившим на итоговой аттестации неудовлетворительные результаты, а также лицам, освоившим часть программы повышения квалификации и (или) отчисленным из организации, выдается справка об обучении или о периоде обучения, по образцу, самостоятельно устанавливаемому организацией.

При освоении программы параллельно с получением среднего профессионального образования и (или) высшего образования удостоверение о повышении квалификации выдается одновременно с получением соответствующего документа об образовании и о квалификации. На момент завершения программы лицам, получающим среднее профессиональное и (или) высшее образование, успешно освоившим соответствующую программу повышения квалификации и прошедшим итоговую аттестацию, выдается справка об обучении или о периоде обучения, по образцу, самостоятельно устанавливаемому организацией.