

Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Ястребов Олег Александрович
Должность: Ректор
Дата подписания: 02.06.2025 12:22:08
Уникальный программный ключ:
sa953a01204891083f939673078ef1a989dae18a

Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов имени Патриса Лумумбы»
Факультет искусственного интеллекта
(наименование основного учебного подразделения (ОУП)-разработчика ОП ВО)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ПРАКТИКУМ ПО ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА (NLP)

(наименование дисциплины/модуля)

Рекомендована МССН для направления подготовки/специальности:

**02.03.02 ФУНДАМЕНТАЛЬНАЯ ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ,
09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА**

(код и наименование направления подготовки/специальности)

Освоение дисциплины ведется в рамках реализации основной профессиональной образовательной программы высшего образования (ОП ВО):

**ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: РАЗРАБОТКА И ОБУЧЕНИЕ
ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ**

(наименование (профиль/специализация) ОП ВО)

2025 г.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Дисциплина «Практикум по обработке естественного языка (NLP)» входит в программу бакалавриата «Искусственный интеллект: разработка и обучение интеллектуальных систем» по направлению 02.03.02 «Фундаментальная информатика и информационные технологии» и изучается в 6, 7, 8 семестрах 3, 4 курсов. Дисциплину реализует Кафедра прикладного искусственного интеллекта. Дисциплина состоит из 8 разделов и 72 тем и направлена на изучение формирования у студентов глубоких практических навыков самостоятельной разработки, настройки, доработки, интеграции и тестирования решений по анализу, генерации и пониманию текстовых данных на русском и английском языке. Курс построен как серия лабораторных и проектных задач разного уровня сложности — от базовой предобработки до промышленных сценариев применения современных моделей глубинного обучения, что обеспечивает освоение студентами индустриальных стандартов и best-practice NLP.

Целью освоения дисциплины является научить студентов самостоятельно строить end-to-end решения по обработке текстов, проводить оценку и сравнение различных методов и инструментов NLP, интегрировать различные библиотеки и фреймворки, использовать современные approaches (от классических ML до DL и трансформеров), а также внедрять свои наработки как в исследовательских, так и в продуктовых задачах.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины «Практикум по обработке естественного языка (NLP)» направлено на формирование у обучающихся следующих компетенций (части компетенций):

Таблица 2.1. Перечень компетенций, формируемых у обучающихся при освоении дисциплины (результаты освоения дисциплины)

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
ПК-1	Способен создавать и оценивать различные модели машинного обучения, архитектуру нейронных сетей и алгоритмы искусственного интеллекта с целью выбора наиболее эффективных решений для конкретных профессиональных задач	ПК-1.1 Может выбирать подходящий алгоритм машинного обучения и архитектуру нейронных сетей для конкретной задачи, учитывая особенности данных и требования к решению; ПК-1.2 Демонстрирует навыки обработки, представления и анализа данных для построения моделей машинного обучения; ПК-1.3 Владеет методами создания и обучения моделей с использованием различных алгоритмов и архитектур; ПК-1.4 Умеет оценивать соблюдение методологии разработки различных моделей машинного обучения, архитектур нейронных сетей и алгоритмов, анализировать качество моделей и разрабатывать стратегии для улучшения качества моделей;
ПК-2	Способен эффективно работать с большими объемами данных, включая их предварительную обработку, анализ и визуализацию, с целью извлечения полезной информации для обучения моделей искусственного интеллекта	ПК-2.2 Демонстрирует навыки анализа данных с использованием статистических методов и инструментов; ПК-2.3 Владеет методами работы с различными алгоритмами машинного обучения и глубокого обучения для решения различных задач;

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

Дисциплина «Практикум по обработке естественного языка (NLP)» относится к блоку по выбору блока образовательной программы высшего образования.

В рамках образовательной программы высшего образования обучающиеся также осваивают другие дисциплины и/или практики, способствующие достижению запланированных результатов освоения дисциплины «Практикум по обработке естественного языка (NLP)».

Таблица 3.1. Перечень компонентов ОП ВО, способствующих достижению запланированных результатов освоения дисциплины

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
ПК-1	Способен создавать и оценивать различные модели машинного обучения, архитектуру нейронных сетей и алгоритмы искусственного интеллекта с целью выбора наиболее эффективных решений для конкретных профессиональных задач	Методы машинного обучения; Анализ естественного языка с помощью методов искусственного интеллекта; Параллельное и распределенное программирование; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; <i>Цифровые двойники**</i> ; <i>Основы больших языковых моделей**</i> ; <i>Основы робототехники**</i> ;	
ПК-2	Способен эффективно работать с большими объемами данных, включая их предварительную обработку, анализ и визуализацию, с целью извлечения полезной информации для обучения моделей искусственного интеллекта	Статистические методы и первичный анализ данных; <i>Цифровые двойники**</i> ; <i>Основы больших языковых моделей**</i> ; Введение в базы данных; Программирование на языке Python; Hadoop, SPARK; Анализ естественного языка с помощью методов искусственного интеллекта; Лингвистические основы анализа естественного языка; Введение в компьютерное зрение; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Программирование на языке C++; <i>Программирование на языке NodeJS**</i> ; <i>Программирование на языке Go**</i> ; <i>Основы робототехники**</i> ; Эксплуатационная практика (учебная); Технологическая (проектно-технологическая) практика (учебная);	

* - заполняется в соответствии с матрицей компетенций и СУП ОП ВО

** - элективные дисциплины /практики

4. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Общая трудоемкость дисциплины «Практикум по обработке естественного языка (NLP)» составляет «12» зачетные единицы.

Таблица 4.1. Виды учебной работы по периодам освоения образовательной программы высшего образования для очной формы обучения.

Вид учебной работы	ВСЕГО, ак.ч.		Семестр(-ы)		
			6	7	8
Контактная работа, ак.ч.	144		68	40	36
Лекции (ЛК)	0		0	0	0
Лабораторные работы (ЛР)	144		68	40	36
Практические/семинарские занятия (СЗ)	0		0	0	0
Самостоятельная работа обучающихся, ак.ч.	198		49	77	72
Контроль (экзамен/зачет с оценкой), ак.ч.	90		27	27	36
Общая трудоемкость дисциплины	ак.ч.	432	144	144	144
	зач.ед.	12	4	4	4

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Таблица 5.1. Содержание дисциплины (модуля) по видам учебной работы

Номер раздела	Наименование раздела дисциплины	Содержание раздела (темы)		Вид учебной работы*
Раздел 1	Вводная обработка текстов и подготовка корпуса	1.1	Импорт, очистка и базовый анализ текстовых данных (русский/английский). Регулярные выражения	ЛР
		1.2	Токенизация: sentence, word, субсловная, сравнение библиотек (NLTK, spaCy, razdel)	ЛР
		1.3	Нормализация, стемминг, лемматизация: сравнение и выбор стратегий	ЛР
		1.4	Удаление стоп-слов, фильтрация чисел, символов, повторов	ЛР
		1.5	Работа с морфологией для русского языка: pymorphy2, natasha	ЛР
		1.6	Очистка шума (html, emoji, unicode, спецсимволы)	ЛР
		1.7	Извлечение метаданных и разметка датасетов (жанр, источник, размер)	ЛР
		1.8	Формирование текстовых корпусов для задачи: агрегация, сэмплирование, недопустимые записи	ЛР
		1.9	Мини-проект: подготовка собственного тематического корпуса (новости, отзывы, научные тексты)	ЛР
Раздел 2	Представление текста: векторизация и эмбединги	2.1	Bag-of-Words, CountVectorizer и TF-IDF: принцип и реализация	ЛР
		2.2	N-граммы, анализ контекстных окон, генерация биграмм/триграмм	ЛР
		2.3	Работа с FastText и Word2Vec: получение и визуализация эмбедингов	ЛР
		2.4	Сравнение классических векторных представлений для классификации	ЛР
		2.5	Использование предобученных эмбедингов rusvectors, GloVe	ЛР
		2.6	Получение и анализ Doc2Vec/Paragraph2Vec для длинных текстов	ЛР
		2.7	Визуализация эмбедингов (TSNE, UMAP, PCA)	ЛР
		2.8	Использование sentence-transformers, SBERT	ЛР
		2.9	Мини-проект: создание собственного эмбединга-корпуса для узкопредметной задачи	ЛР
Раздел 3	Классификация и регрессия текстов	3.1	Базовые ML-модели классификации: Naïve Bayes, Logistic Regression	ЛР
		3.2	Классификация отзывов и новостей с помощью ML (scikit-learn)	ЛР
		3.3	Построение метрик: accuracy, F1, confusion matrix	ЛР
		3.4	Оценка важности признаков для текстовых задач	ЛР
		3.5	Классификация длинных текстов (документы, статьи)	ЛР
		3.6	Тематическая классификация: мультитематика, multilabel	ЛР
		3.7	Регрессия по тексту: тональность, рейтинг, балл по отзыву	ЛР
		3.8	Использование sklearn pipeline для автоматизации работы	ЛР
		3.9	Мини-проект: построение полного рабочего классификатора (выбор задачи)	ЛР
Раздел 4	Извлечение информации: NER, морфология,	4.1	Named Entity Recognition (NER) с помощью spaCy/huggingface	ЛР

Номер раздела	Наименование раздела дисциплины	Содержание раздела (темы)		Вид учебной работы*
	тематическое моделирование	4.2	Извлечение дат, временных интервалов, имен, организаций	ЛР
		4.3	Morphological tagging, PoS-теггинг, разметка UD формата	ЛР
		4.4	Тематическое моделирование: LDA, LSI, раскраска тем в тексте	ЛР
		4.5	Выделение ключевых слов и фраз: TextRank, RAKE, yake	ЛР
		4.6	Построение семантических сетей по результатам NER/PoS	ЛР
		4.7	Визуализация результатов извлечения информации	ЛР
		4.8	Практика по ошибкам и конфликтам в результатах NER	ЛР
		4.9	Мини-проект: комплексное извлечение информации по кейсу (напр. автоматизация документооборота)	ЛР
Раздел 5	Анализ и обработка структур сложных текстов	5.1	Синтаксический парсинг и построение деревьев зависимостей (UDPipe, SyntaxNet)	ЛР
		5.2	Фрагментация сложных документов: абзацы, главы, рубрики	ЛР
		5.3	Определение связей между абзацами/частями текста (coreference, entity linking)	ЛР
		5.4	Построение фреймов и шаблонов событий из текста	ЛР
		5.5	Анализ аргументации и риторических структур (RST анализ)	ЛР
		5.6	Обработка диалогов и чатов: выделение ролей, turn-taking	ЛР
		5.7	Построение графов текста для анализа тематических линий	ЛР
		5.8	Сравнение различных синтаксических анализаторов	ЛР
		5.9	Мини-проект: полноценный парсинг и анализ разделённого на главы документа	ЛР
Раздел 6	Поколение текста и генеративные задачи	6.1	Генерация текста по шаблону, последовательная генерация цепочек	ЛР
		6.2	Использование seq2seq моделей (OpenNMT, Fairseq, huggingface)	ЛР
		6.3	Summarization: составление аннотаций к текстам	ЛР
		6.4	Question Generation и Question Answering: вводные практики	ЛР
		6.5	Простое машинное перевод: en-ru, ru-en, библиотека Marian/T5	ЛР
		6.6	Диалоговые системы: генерация реплик, rule-based и ML	ЛР
		6.7	Саммаризация больших текстов	ЛР
		6.8	Практика генерации FAQ по корпусу	ЛР
		6.9	Мини-проект: генератор аннотаций/ответов/диалогов или переводчик	ЛР
Раздел 7	Модели глубокого обучения и трансформеры для NLP	7.1	Fine-tuning bert-like моделей (BERT, RuBERT, DistilBERT, RoBERTa)	ЛР
		7.2	Классификация, NER, QA через Huggingface Transformers	ЛР
		7.3	Практика Zero-shot и Few-shot обучения	ЛР
		7.4	Использование современных эмбеддеров (sentence-transformers, Universal Sentence Encoder)	ЛР

Номер раздела	Наименование раздела дисциплины	Содержание раздела (темы)		Вид учебной работы*
		7.5	Extractive vs Generative QA	ЛР
		7.6	Prompt Engineering для LLM API (ChatGPT, GPT-4, YandexGPT)	ЛР
		7.7	Оценка и интерпретация результатов DL-моделей для NLP	ЛР
		7.8	Мониторинг и ускорение вывода больших моделей (ONNX, quantization)	ЛР
		7.9	Мини-проект: узкоспециализированный пайплайн на BERT/LLM под выбранную задачу	ЛР
Раздел 8	Интеграция, деплой и поддержка NLP-решений	8.1	Интеграция ML-моделей в веб-интерфейсы и chatbots (FastAPI, Flask, Bot frameworks)	ЛР
		8.2	Построение REST API для выдачи NLP-результатов	ЛР
		8.3	Docker-контейнеризация и упаковка пайплайна	ЛР
		8.4	Автоматизация экспериментов и мониторинга (MLflow, wandb)	ЛР
		8.5	Управление версиями моделей и текстовых данных	ЛР
		8.6	Реализация защищенного вывода (фильтрация, moderation, toxicity detection)	ЛР
		8.7	Прототипирование и A/B тестирование NLP-сервисов	ЛР
		8.8	Поддержка и обновление моделей на production	ЛР
		8.9	Итоговый мини-проект: развёртывание end-to-end NLP-решения для приложения или бизнеса, подготовка инструкции и кейс-документации	ЛР

* - заполняется только по **ОЧНОЙ** форме обучения: ЛК – лекции; ЛР – лабораторные работы; СЗ – практические/семинарские занятия.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Таблица 6.1. Материально-техническое обеспечение дисциплины

Тип аудитории	Оснащение аудитории	Специализированное учебное/лабораторное оборудование, ПО и материалы для освоения дисциплины (при необходимости)
Компьютерный класс	Компьютерный класс для проведения занятий, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная персональными компьютерами (в количестве 25 шт.), доской (экраном) и техническими средствами мультимедиа презентаций.	
Для самостоятельной работы	Аудитория для самостоятельной работы обучающихся (может использоваться для проведения семинарских занятий и консультаций), оснащенная комплектом специализированной мебели и компьютерами с доступом в ЭИОС.	

* - аудитория для самостоятельной работы обучающихся указывается **ОБЯЗАТЕЛЬНО!**

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература:

1. Гольдберг Й. Нейросетевые методы в обработке естественного языка / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2019. – 282 с.: ил. ISBN 978-5-97060-754-1

Дополнительная литература:

1. Лейн Хобсон, Хапке Ханнес, Ховард Коул. Обработка естественного языка в действии. — СПб.: Питер, 2021. ISBN 978-5-4461-1371-2

Ресурсы информационно-телекоммуникационной сети «Интернет»:

1. ЭБС РУДН и сторонние ЭБС, к которым студенты университета имеют доступ на основании заключенных договоров

- Электронно-библиотечная система РУДН – ЭБС РУДН

<https://mega.rudn.ru/MegaPro/Web>

- ЭБС «Университетская библиотека онлайн» <http://www.biblioclub.ru>

- ЭБС «Юрайт» <http://www.biblio-online.ru>

- ЭБС «Консультант студента» www.studentlibrary.ru

- ЭБС «Знаниум» <https://znanium.ru/>

2. Базы данных и поисковые системы

- Sage <https://journals.sagepub.com/>

- Springer Nature Link <https://link.springer.com/>

- Wiley Journal Database <https://onlinelibrary.wiley.com/>

- Научометрическая база данных Lens.org <https://www.lens.org>

Учебно-методические материалы для самостоятельной работы обучающихся при освоении дисциплины/модуля:*

1. Курс лекций по дисциплине «Практикум по обработке естественного языка (NLP)».

* - все учебно-методические материалы для самостоятельной работы обучающихся размещаются в соответствии с действующим порядком на странице дисциплины **в ТУИС!**

РАЗРАБОТЧИК:

Заведующий кафедрой
прикладного искусственного
интеллекта

Должность, БУП

Подпись

Подолько Павел
Михайлович

Фамилия И.О.

РУКОВОДИТЕЛЬ БУП:

Заведующий кафедрой
прикладного искусственного
интеллекта

Должность, БУП

Подпись

Подолько Павел
Михайлович

Фамилия И.О.

РУКОВОДИТЕЛЬ ОП ВО:

Заведующий кафедрой
прикладного искусственного
интеллекта

Должность, БУП

Подпись

Подолько Павел
Михайлович

Фамилия И.О.