

Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Ястребов Олег Александрович
Должность: Ректор
Дата подписания: 09.06.2025 13:08:32
Уникальный программный ключ:
sa953a0120d891083f939673078ef1a989dae18a

**Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов имени Патриса Лумумбы»**

Инженерная академия

(наименование основного учебного подразделения (ОУП)-разработчика ОП ВО)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ВВЕДЕНИЕ В ОБРАБОТКУ ЕСТЕСТВЕННОГО ЯЗЫКА

(наименование дисциплины/модуля)

Рекомендована МССН для направления подготовки/специальности:

27.04.04 УПРАВЛЕНИЕ В ТЕХНИЧЕСКИХ СИСТЕМАХ

(код и наименование направления подготовки/специальности)

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

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, МАШИННОЕ ОБУЧЕНИЕ И КОСМИЧЕСКИЕ НАУКИ

(наименование (профиль/специализация) ОП ВО)

2025 г.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Дисциплина «Introduction to Natural Language Processing» входит в программу магистратуры «Искусственный интеллект, машинное обучение и космические науки» по направлению 27.04.04 «Управление в технических системах» и изучается во 2 семестре 1 курса. Дисциплину реализует Кафедра Вуза-Партнёра. Дисциплина состоит из 8 разделов и 16 тем и направлена на изучение основных методов и подходов к ОЕЯ, принципов оценки качества методов ОЕЯ.

Целью освоения дисциплины является знакомство с основными методами и приложениями автоматической обработки естественного языка (ОЕЯ), получение практических навыков работы с инструментами ОЕЯ.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины «Введение в обработку естественного языка» 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 формулировать цели, задачи научных исследований в области управления аэрокосмическими системами, выбирать методы и средства решения профессиональных задач	ПК-1.1 Знает методы и средства решения задач научных исследований в области систем искусственного интеллекта и робототехнических систем;; ПК-1.2 Умеет формулировать цель и задачи научных исследований в профессиональной области;; ПК-1.3 Владеет приемами для формулировки цели и задач научных исследований, умеет выбирать методы и средства решения задач профессиональной деятельности;

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

Дисциплина «Введение в обработку естественного языка» относится к обязательной части блока 1 «Дисциплины (модули)» образовательной программы высшего образования.

В рамках образовательной программы высшего образования обучающиеся также осваивают другие дисциплины и/или практики, способствующие достижению запланированных результатов освоения дисциплины «Введение в обработку естественного языка».

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 формулировать цели, задачи научных исследований в области управления аэрокосмическими системами, выбирать методы и средства решения		<i>Artificial Neural Networks (Deep Learning)**;</i> <i>Искусственные нейронные сети (Глубокое обучение)**;</i> <i>Artificial Neural Networks (Reinforcement Learning)**;</i> <i>Undergraduate practice /</i> <i>Преддипломная практика;</i>

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
	профессиональных задач		

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

** - элективные дисциплины /практики

4. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Общая трудоемкость дисциплины «Introduction to Natural Language Processing» составляет «4» 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)
			2
<i>Контактная работа, ак.ч.</i>	34		34
Лекции (ЛК)	17		17
Лабораторные работы (ЛР)	17		17
Практические/семинарские занятия (СЗ)	0		0
<i>Самостоятельная работа обучающихся, ак.ч.</i>	83		83
<i>Контроль (экзамен/зачет с оценкой), ак.ч.</i>	27		27
Общая трудоемкость дисциплины	ак.ч.	144	144
	зач.ед.	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	Введение	1.1	Варианты названия дисциплины, определение, особенности, задачи, приложения, методы.	ЛК, ЛР
		1.2	Краткая история мировой и советской/российской ОЕЯ. Соержание и "идеология" курса. Ресурсы.	ЛК, ЛР
Раздел 2	Морфологический анализ	2.1	Чем занимается морфология, для чего нужна машинная морфология. Стемминг, лемматизация, частеречная (ЧР) разметка. Алгоритм Портера для английского.	ЛК, ЛР
		2.2	Лемматизация для русского языка: mystem, rymorphy2, АОТ. Словарь Зализняка как источник данных для лемматизации. Как строить гипотезы для незнакомых слов. Задача ЧР-разметки. Скрытые марковские модели для ЧР-разметки. Источники данных. Алгоритм Витерби.	ЛК, ЛР
Раздел 3	Классический сценарий информационного поиска, особенности задач информационного поиска	3.1	Базовые метрики качества поиска: точность, полнота, F1. Морфология в поиске. Частотные свойства терминов в коллекции: законы Хипса и Ципфа. Модель векторного пространства.	ЛК, ЛР
		3.2	Взвешивание терминов: подход tf.idf. Другие задачи ОЕЯ в поиске. Автоматическое реферирование. Постановка задачи, области применения, примеры. Типы рефератов. Методы реферирования текстовых документов. Постобработка рефератов. Оценка, инициативы по оценке методов автоматического реферирования. Вебсниппеты (рефераты с учетом запроса): особенности задачи, методы. Оценка вебсниппетов.	ЛК, ЛР
Раздел 4	Языковые модели (Language Models)	4.1	Области применения, ограничения. N-граммы, оценка вероятностей, доступные большие коллекции n-грамм. Оценка языковых моделей, перплексия.	ЛК, ЛР
		4.2	Сглаживание (smoothing), откат (backoff) и интерполяция (interpolation). Технические приемы при работе с языковыми моделями масштаба Веба, языковые модели с "памятью". Сглаживнаие Лапласа, Гуда-Тьюринга, Кнезера-Нея.	ЛК, ЛР
Раздел 5	Синтаксический анализ (syntactic parsing)	5.1	Два формализма описания синтаксиса: система составляющих (constituency) и дерево зависимостей (dependency). Контекстносвободные грамматики (КСГ): возможности и ограничения. Вероятностные КСГ. Алгоритм вероятностного синтаксического разбора.	ЛК, ЛР
		5.2	Оценка результатов разбора. Лексикализация ВКСГ.	ЛК, ЛР
Раздел 6	Извлечение информации из текстовых документов	6.1	Особенности задачи, источники данных. Именованные сущности (NEs), отношения. Основные подходы. Оценка. Машинный перевод: краткая история, сложности, подходы, приложения. Перевод, основанный на правилах (RBMT) и статистический машинный перевод	ЛК, ЛР

Номер раздела	Наименование раздела дисциплины	Содержание раздела (темы)		Вид учебной работы*
			(СМП, STM).	
		6.2	Источники данных для статистического машинного перевода. Выравнивание параллельного корпуса. IBM Models 1, 2. Фразовый СМП. Оценка систем МП: ручная, автоматическая (BLUE)	ЛК, ЛР
Раздел 7	Анализ тональности (sentiment analysis): источники данных и различные постановки задачи, примеры приложений	7.1	Подход на основе классификации. Сложности анализа тональности: разнообразие форм выражения, ирония и сарказм, порядок и связь оценок. Использование словарей тонально окрашенных слов. Автоматическое пополнение/создание словарей.	ЛК, ЛР
		7.2	Обработка отрицаний. Фразы vs. отдельные слова. Аспекты/атрибуты: автоматическое выделение и списки, составленные вручную.	ЛК, ЛР
Раздел 8	Семантика, разные подходы и определения	8.1	Семантика, разные подходы и определения: логика высказываний, семантический веб, базы знаний, онтологии предметных областей, тезаурусы. Лексическая семантика: омонимы, многозначные слова, синонимы, антонимы, гипонимы/гиперонимы. Семантические словари -- тезаурусы. WordNet: синсеты и отношения. Методы определения семантической близости слов на основе тезауруса.	ЛК, ЛР
		8.2	Дистрибутивная семантика: семантическая близость на основе корпуса. Положительная поточечная взаимная информация (PPMI). Вычисления близости на основе структуры словосочетаний. Векторное представление семантики слов на основе на основе нейронных сетей	ЛК, 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)
Лекционная	Аудитория для проведения занятий лекционного типа, оснащенная комплектом специализированной мебели; доской (экраном) и техническими средствами мультимедиа презентаций.	
Компьютерный класс	Компьютерный класс для проведения занятий, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная персональными компьютерами (в количестве 14 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. Kang Y. et al. Natural language processing (NLP) in management research: A literature review //Journal of Management Analytics. – 2020. – Т. 7. – №. 2. – С. 139-172.
2. Vajjala S. et al. Practical natural language processing: a comprehensive guide to building real-world NLP systems. – O'Reilly Media, 2020.

Дополнительная литература:

1. Cambria E., White B. Jumping NLP curves: A review of natural language processing research //IEEE Computational intelligence magazine. – 2014. – Т. 9. – №. 2. – С. 48-57.
2. Mihalcea R., Liu H., Lieberman H. NLP (natural language processing) for NLP (natural language programming) //International Conference on intelligent text processing and computational linguistics. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2006. – С. 319-330.

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 «Введение в обработку естественного языка».

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

РАЗРАБОТЧИК:

Доцент

Должность, БУП

Подпись

Салтыкова Ольга

Александровна

Фамилия И.О.

РУКОВОДИТЕЛЬ БУП:

Должность БУП

Подпись

Фамилия И.О.

РУКОВОДИТЕЛЬ ОП ВО:

Заведующий кафедрой

Должность, БУП

Подпись

Разумный Юрий

Николаевич

Фамилия И.О.