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)

2026 г.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Дисциплина «Введение в обработку естественного языка» 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 «Introduction to Natural Language Processing» 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 «Introduction to Natural Language Processing».

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 профессиональных задач		Artificial Neural Networks (Deep Learning)**; Искусственные нейронные сети (Глубокое обучение)**; Artificial Neural Networks (Reinforcement Learning)**; Undergraduate practice / Преддипломная практика;

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

** - элективные дисциплины /практики

4. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Общая трудоемкость дисциплины «Введение в обработку естественного языка» 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
Контактная работа, ак.ч	*34*		*34*
Лекции (**ЛК**)	17		17
Лабораторные работы (**ЛР**)	17		17
Практические/семинарские занятия (**СЗ**)	0		0
Самостоятельная работа обучающихся, ак.ч.	*83*		*83*
Контроль (экзамен/зачет с оценкой), ак.ч.	*27*		*27*
Общая трудоемкость дисциплины ак.ч.	**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	Наименование темы		Содержание темы	Вид учебной работы*
Раздел 1	Введение	1.1	Определение, особенности, задачи, приложения, методы.	Определение обработки естественного языка как междисциплинарной области на стыке лингвистики, информатики и искусственного интеллекта. Особенности обработки естественного языка: неоднозначность, зависимость от контекста, наличие исключений. Основные задачи: токенизация, морфологический анализ, синтаксический разбор, анализ тональности, информационный поиск, машинный перевод, извлечение информации, автоматическое реферирование. Прикладные области: поисковые системы, голосовые ассистенты, анализ отзывов, автоматический перевод, чат-боты. Методы: правила и словари, статистические методы, машинное обучение.	ЛК, ЛР
		1.2	Краткая история мировой и советской/российской ОЕЯ. Сожержание и "идеология" курса. Ресурсы.	Этапы развития мировой обработки естественного языка: первые работы по машинному переводу, период использования правил и словарей, переход к вероятностным и статистическим методам, эпоха машинного обучения. Вклад советской и российской школы компьютерной лингвистики. Содержание и идеология курса: последовательное изучение уровней языка и методов машинной обработки. Ресурсы для изучения: учебные пособия, открытые корпуса текстов.	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.	Стемпинг, лемматизация, частеречная разметка. Алгоритм Портера для английского языка. Задачи морфологии: изучение строения слов, их форм и словоизменения. Назначение машинной морфологии: приведение слов к нормальной форме, определение грамматических характеристик. Стемпинг как упрощённое отсечение окончаний. Лемматизация как приведение к словарной форме с учётом контекста. Частеречная разметка как определение части речи слова в предложении. Алгоритм Портера для английского языка: последовательное удаление окончаний по правилам.	ЛК, ЛР
		2.2	Лемматизация для русского языка: mystem, pymorphy2, 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.	Словарь Зализняка как источник данных для лемматизации. Программные инструменты: mystem, pymorphy2, АОТ. Построение гипотез для незнакомых слов. Задача частеречной разметки. Скрытые марковские модели для частеречной разметки. Источники данных для обучения моделей.	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.	Морфология в поиске. Частотные свойства терминов в коллекции. Модель векторного пространства. Метрики качества поиска: точность, полнота, F1. Применение морфологии в информационном поиске для учёта словоформ. Законы Хипса и Ципфа, описывающие распределение частот терминов. Модель векторного пространства для представления документов и запросов в виде векторов.	ЛК, ЛР

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы*
		3.2	Взешивание 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.	Другие задачи обработки естественного языка в поиске. Автоматическое реферирование. Постановка задачи, области применения, примеры. Типы рефератов. Методы реферирования текстовых документов. Постобработка рефератов. Оценка качества реферирования. Подход tf.idf для оценки важности термина в документе относительно всей коллекции. Другие задачи: извлечение ключевых слов, классификация документов. Автоматическое реферирование как создание краткого изложения исходного текста. Типы рефератов: индикативные, информативные, обзорные. Методы реферирования: извлечение предложений, абстрактное реферирование.	ЛК, ЛР
Раздел 4	Языковые модели (Language Models)	4.1	Области применения, ограничения. N-граммы, оценка вероятностей, доступные большие коллеакции n-грамм. Оценка языковых моделей, перплексия.	N-граммы, оценка вероятностей, доступные большие коллекции n-грамм. Оценка языковых моделей, перспективы. Области применения языковых моделей: распознавание речи, машинный перевод, проверка орфографии, генерация текста. Ограничения: трудность учёта длинных зависимостей, требования к объёму данных. N-граммы как последовательности из n элементов. N-граммы как последовательности из n элементов. Оценка вероятностей на основе частот в корпусе. Доступные большие коллекции n-грамм. Перспективы развития языковых моделей: нейросетевые модели, большие языковые модели.	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.	Технические приёмы при работе с языковыми моделями масштаба Веба, языковые модели с памятью. Сглаживание Лапласа, Гуда-Тьюринга, Кнезера-Нея. Сглаживание как способ оценки вероятности для отсутствующих в корпусе последовательностей. Откат как переход к более коротким n-граммам при недостатке данных. Интерполяция как комбинирование вероятностей из n-грамм разной длины. Языковые модели с памятью для учёта длинных зависимостей.	ЛК, ЛР
Раздел 5	Синтаксические анализ (syntactic parsing)	5.1	Два формализма описания синтаксиса: система составляющих (constituency) и дерево зависимостей (dependency). Контекстносвободные	Система составляющих: иерархическое объединение слов во фразы и предложения. Дерево зависимостей: грамматические связи между словами. Контекстно-свободные грамматики для описания структуры составляющих. Ограничения контекстно-свободных грамматик: неспособность учесть лексические и контекстные предпочтения. Вероятностные контекстно-свободные грамматики с оценками вероятностей правил. Алгоритм вероятностного разбора для поиска наиболее	ЛК, ЛР

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы*
			грамматики (КСГ): возможности и ограничения. Вероятностные КСГ. Алгоритм вероятностного синтаксического разбора.	вероятного дерева.	
		5.2	Оценка результатов разбора. Лексикализация ВКСГ.	Лексикализация вероятностных контекстно-свободных грамматик. Метрики оценки синтаксического анализатора. Лексикализация как учёт конкретных лексических единиц в правилах грамматики.	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 (СМП, STM).	Именованные сущности, отношения. Основные подходы. Оценка. Машинный перевод: краткая история, сложности, подходы, приложения. Особенности извлечения информации: работа с неструктурированным текстом, выделение фактов. Источники данных: новостные ленты, научные публикации, социальные сети. Именованные сущности: имена людей, названия организаций, географические названия, даты. Отношения между сущностями. Основные подходы к извлечению: основанные на правилах и основанные на машинном обучении. Оценка качества извлечения информации. Машинный перевод: краткая история развития. Сложности перевода: многозначность, порядок слов, идиомы. Подходы к машинному переводу. Прикладные области.	ЛК, ЛР
		6.2	Источники данных для статистического машинного перевода. Выравнивание параллельного корпуса. IBM Models 1, 2. Фразовый СМП. Оценка систем МП: ручная, автоматическая (BLUE)	Выравнивание параллельного корпуса. Фразовый статистический машинный перевод. Оценка систем машинного перевода: ручная и автоматическая. Источники данных: параллельные корпуса текстов на двух и более языках. Выравнивание как установление соответствия между предложениями исходного и целевого языка. Фразовый статистический машинный перевод. Оценка систем машинного перевода: ручная оценка и автоматическая метрика.	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 связт оценок. Использование словарей тонально окрашенных слов. Автоматическое пополнение/создание словарей.	Сложности анализа тональности. Использование словарей тонально окрашенных слов. Автоматическое пополнение словарей. Анализ тональности как определение эмоциональной окраски текста. Источники данных: отзывы, сообщения в социальных сетях, новостные комментарии. Постановки задачи: бинарная классификация позитив или негатив, трёхклассовая с учётом нейтральной оценки, шкала оценок. Подход на основе классификации с обучением на размеченных примерах. Сложности: разнообразие форм выражения эмоций, ирония и сарказм, порядок и связь оценок. Использование словарей тонально окрашенных слов с положительными и отрицательными значениями. Автоматическое пополнение словарей из корпусов текстов.	ЛК, ЛР
		7.2	Обработка отрицаний.	Фразы в сравнении с отдельными словами. Аспекты и атрибуты: автоматическое	ЛК, ЛР

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы*
			Фразы vs. отдельные слова. Аспекты/атрибуты: автоматическое выделение и списки, составленные вручную.	выделение и ручное составление списков. Обработка отрицаний как инвертирование тональности при наличии слов отрицания. Анализ фраз в сравнении с анализом отдельных слов для учёта контекста. Выделение аспектов и атрибутов объекта оценки. Автоматическое выделение аспектов из текстов и составление списков вручную.	
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.	Логика высказываний, семантический веб, базы знаний, онтологии предметных областей, тезаурусы. Лексическая семантика: омонимы, многозначные слова, синонимы, антонимы, гипонимы и гиперонимы. Различные подходы к семантике. Логика высказываний для формализации смысла. Семантический веб как проект по приданию данным машиночитаемого смысла. Базы знаний как структурированные хранилища фактов. Онтологии предметных областей как формальные описания понятий и их связей. Тезаурусы как словари семантических отношений. Лексическая семантика изучает значения слов. Омонимы совпадают по написанию, но различны по смыслу. Многозначные слова имеют несколько связанных значений. Синонимы близки по значению. Антонимы противоположны по значению. Гипонимы и гиперонимы связаны отношением вид-род.	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	Положительная поэтесная взаимная информация. Вычисления близости на основе структуры словосочетаний. Векторное представление слов. Дистрибутивная семантика основана на принципе: слова близкие по значению встречаются в схожих контекстах. Семантическая близость вычисляется на основе анализа больших корпусов. Положительная поэтесная взаимная информация как мера связи между словом и его контекстом. Вычисление близости на основе структуры словосочетаний. Векторное представление слов в многомерном пространстве признаков.	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 ____ 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/
   - Наукометрическая база данных Lens.org https://www.lens.org

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

1. Курс лекций по дисциплине «Introduction to Natural Language Processing».

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

РАЗРАБОТЧИКИ

Доцент	Салтыкова О.А.
Должность	Фамилия И.О

РУКОВОДИТЕЛЬ ОП ВО

Профессор	Разумный Ю.Н.
Должность	Фамилия И.О