АННОТАЦИЯ ДИСЦИПЛИНЫ

«Модели и методы обработки естественного языка»

Дисциплина «Модели и методы обработки естественного языка» является частью программы магистратуры «Разработка программно-информационных систем» по направлению «09.04.04 Программная инженерия».

Цели и задачи дисциплины

Формирование комплекса знаний, умений и навыков в области применения моделей, методов и инструментов обработки естественного языка для разработки информационного, лингвистического, программного обеспечения интеллектуальных информационных систем..

Изучаемые объекты дисциплины

Естественный язык; цифровое представление естественного языка; модели естественного языка; методы обработки естественного языка; программные инструменты обработки естественного языка..

Объем и виды учебной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		4
1. Проведение учебных занятий (включая проведе-ние текущего контроля успеваемости) в форме: 1.1. Контактная аудиторная работа, из них:	72	72
- лекции (Л)	18	18
- лабораторные работы (ЛР)	24	24
- практические занятия, семинары и (или) другие виды занятий семинарского типа (ПЗ)	26	26
- контроль самостоятельной работы (КСР)	4	4
- контрольная работа		
1.2. Самостоятельная работа студентов (СРС)	72	72
2. Промежуточная аттестация		
Экзамен		
Дифференцированный зачет	9	9
Зачет		
Курсовой проект (КП)		
Курсовая работа (КР)		
Общая трудоемкость дисциплины	144	144

Краткое содержание дисциплины

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
4-й семестр				
Информационный поиск	2	4	4	10
Индексирование. Булевый поиск, ранжированный поиск. Оценка релевантности документа. Принципы работы поисковых движков. Квазиреферирование и автоматическое аннотирование документов. Основные стратегии квазиреферирования. Обзорное реферирование. Типы аннотаций.				
Введение в обработку естественного языка	4	2	0	6
Понятие естественного языка. Характеристики естественного языка. Уровни языка. Этапы анализа текста. Основные задачи обработки естественного языка. Языковые модели. Морфологический анализ и синтез. Стемминг, лемматизация, полный морфоанализ. Морфологические процессоры. Инструменты обработки естественного языка в экосистеме Python.				
Машинный перевод	2	2	4	8
Стратегии машинного перевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика. Современные подходы к машинному переводу. Многоязычный машинный перевод. Стратегии перевода специализированных текстов (деловых, технических)				
Классификация и кластеризация текстов	2	4	4	10
Интеллектуальный анализ данных: Data Mining и Text Mining. Особенности классификации и кластеризации текстов. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы.				
Векторные модели текста	2	2	4	8
Методы сбора текстовых данных. Предварительная обработка текста. Извлечение языковых данных. Работа с N-граммами. Лемматизация. Векторизация. Обзор подходов к векторизации. Метрики. Векторная модель документа. Использование векторного представления для анализа текстов. Работа в N-мерном пространстве.				

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
Корпусная лингвистика	2	2	2	8
Понятие корпуса. Соотношение корпуса и базы данных. Создание и применение корпусов. Обработка и преобразования корпуса текста: сегментация, лексемизация, промежуточный анализ корпуса. Разметка корпуса. Виды разметки и области их применения. Обзор существующих общедоступных корпусов.				
Методы машинного обучения в обработке естественного языка	2	4	4	12
Формальные методы определения автора текста. Статистические методы атрибуции. Авторский инвариант и лингвистические спектры. Применение методов кластеризации и классификации для установления авторства текстов. Методы обнаружения спама. Автоматический анализ тональности текстов и извлечение мнений из текстов. Искусственные нейронные сети как основа для лингвистических моделей. Большие лингвистические модели (LLM). Современные архитектуры ИНС для лингвистических задач: трансформеры, диффузионные модели. Применение лингвистических моделей в задачах text2image, text2speech, text2video. GPT, BERT, CLIP, Whisper.				
Семантический анализ текста	2	4	4	10
Способы представления смысла текста. Понятие денотата, сигнификата, референта. Треугольник Фреге. Семантический анализ текста на основе семантико-синтаксических моделей управления. Разметка частей речи. Выделение именованных сущностей. Извлечение информации и отношений из текста. Извлечение информации и знаний из текстов. Подход А.И. Новикова. Денотатный граф. Моделирование предметной области средствами денотатного графа. Модель «СМЫСЛ <-> ТЕКСТ» И.А. Мельчука. Ограничения существующих семантических моделей.				
ИТОГО по 4-му семестру	18	24	26	72
ИТОГО по дисциплине	18	24	26	72