Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Максимов Алексей Борисович
Должность: директор департамента по образовательной политике
Дата подписания: 10.12.2025 15:55:51
Уникальный программный ключ:
8db180d1a3f02ac9e60521a5672742735c18b1d6

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение
высшего образования
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Факультет информационных технологий

УТВЕРЖДАЮ
Декан факультета
«Информационные технологии»
_____/Д.Г.Демидов/
«16» мая 2022

Рабочая программа дисциплины

«Обработка естественного языка и машинный анализ текстов»

Направление подготовки
09.03.01 Информатика и вычислительная техника

Профиль
«Системная и программная инженерия»

Квалификация
Бакалавр

Формы обучения
очная

Москва, 2022 г.

Разработчик(и):

старший преподаватель кафедры «Информационная безопасность»

/А.Ю. Гневшев/

Согласовано:
И.о. заведующего кафедрой «Информационная безопасность»,

/А.Ю. Гневшев/

Руководитель образовательной программы,

/А.Ю. Гневшев/

Заведующий кафедрой «Инфокогнитивные технологии»,

доцент, к.т.н. /Е.А.Пухова/

1 Цели, задачи и планируемые результаты обучения по дисциплине

Цель изучения дисциплины – изучение методов и алгоритмов анализа и обработки текстов на естественном языке с помощью компьютерных технологий.

Обучение по дисциплине «Обработка естественного языка и машинный анализ текстов» направлено на формирование у обучающихся следующих компетенций:

Код и наименование компетенций	Индикаторы достижения компетенции
ОПК-8. Способен разрабатывать алгоритмы и программы, пригодные для практического применения	ИОПК-8.1. Знает основные технологии создания и внедрения информационных систем, стандарты управления жизненным циклом информационной системы. ИОПК-8.2. Умеет проектировать блок-схемы алгоритмов, оценивать производительность алгоритмов и затраты памяти на работу алгоритма, разрабатывать программы на основе спроектированного алгоритма и проводить отладку программы, применять методы системного анализа и математического моделирования при разработке и эксплуатации ИС, проводить структурный анализ, функциональный анализ, объектно-ориентированный анализ иерархии классов. осуществлять организационное обеспечение выполнения работ на всех стадиях и в процессах жизненного цикла информационной системы. ИОПК-8.3. Владеет навыками разработки программ, построения блок-схем алгоритмов и оценки производительности алгоритмов, работы с унифицированным языком визуального моделирования, составления плановой и отчетной документации по управлению проектами создания информационных систем на стадиях жизненного цикла.

В процессе освоения образовательной программы данные компетенции, в том числе их отдельные компоненты, формируются поэтапно в ходе освоения обучающимися дисциплин (модулей), практик в соответствии с учебным планом и календарным графиком учебного процесса.

2 Место дисциплины в структуре образовательной программы

Дисциплина относится к базовой части основной профессиональной образовательной программы бакалавриата.

Дисциплина взаимосвязана логически и содержательно-методически со следующими дисциплинами и практиками ОПОП:

- Разработка технических текстов и документации.

3 Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 4 зачетных единицы, т.е. 144 академических часа (из них 72 часов – самостоятельная работа студентов и 72 часа – аудиторные занятия).

Разделы дисциплины изучаются на четвертом курсе в седьмом семестре, форма промежуточной аттестации – экзамен, курсовой проект.

3.1 Виды учебной работы и трудоемкость (по формам обучения)

3.1.1 Очная форма обучения

№ п/п	Вид учебной работы	Количество часов	Семестры	
			Семестр	Количество недель
1	**Аудиторные занятия**	**72**	7	18
	В том числе:			
1.1	Лекции	18		
1.2	Семинарские/практические занятия	18		
1.3	Лабораторные занятия	36		
2	**Самостоятельная работа**	**72**	7	18
3	**Курсовой проект**	**72**	7	18
4	**Промежуточная аттестация**		7	
	Экзамен			
	Итого:	**216**		

3.2 Тематический план изучения дисциплины (по формам обучения)

3.2.1 Очная форма обучения

№ п/п	Разделы/темы дисциплины	Трудоемкость, час					
		Всего	Аудиторная работа				Самостоятельная работа
			Лекции	Семинарские/ практические занятия	Лабораторные занятия	Практическая подготовка	
1	Введение	3	1	-	-	-	2
2	Модели и методы представления и организации знаний	18	4	-	6	-	8
3	Количественная спецификация естественно-языковых систем	7	4	-	-	-	3
4	Логико-статистические методы обработки естественного языка	17	3	-	6	-	3
5	Синтагматическая модель машинного анализа текста	17	3	-	6	-	8
6	Парадигматическая модель машинного анализа текста	11	3	-	-	-	8
7	Современные вопросы обработки естественного языка	16	-	2	6	-	8

8	Работа с табличными данными. Задачи классификации и кластеризации	12	-	4	-	-	8
9	Работа со структурированными данными при помощи нейронных сетей	18	-	4	6	-	8
10	Технологии text-to-image (T2I) генерации изображений на основе текста	12	-	4	-	-	8
11	Вайб-кодинг: программирование через диалог с AI	18	-	4	6	-	8
Итого		**144**	**18**	**18**	**36**	**-**	**72**

3.3 Содержание дисциплины

Тема 1. Введение в дисциплину «Обработка естественного языка и машинный анализ текстов»

Система языковедческих дисциплин и их роль в изучении языка. Основные направления языкознания: внутреннее устройство языка, его историческое развитие и функционирование в обществе. Прикладная лингвистика в вопросах исследования практических задач работы с языком, включая компьютерную и инженерную лингвистику. Междисциплинарный характер современной обработки естественного языка.

Тема 2. Модели и методы представления и организации знаний

Способы формализации знаний человека в интеллектуальных системах. Методы автоформализации знаний, лексикографическое, тезаурусное и формально-языковое описание, а также их роль в создании текстов, словарей и тезаурусов. Значение когнитивных структур и инструментов представления знаний для разработки информационных технологий обработки текста.

Тема 3. Количественная спецификация естественно-языковых систем

Основы лингвистической статистики и применение математических методов к анализу естественно-языковых текстов. Законы распределения языковых единиц, в том числе модель «ранг–частота», законы Ципфа и Мандельброта. Роль количественного анализа для выявления закономерностей функционирования языка и построения моделей текстов.

Тема 4. Логико-статистические методы обработки естественного языка

Дистрибутивно-статистические методы выявления ассоциативных связей между словами. Статистические и дистрибутивные подходы, анализ совместной встречаемости лексических единиц, а также формулы для вычисления коэффициентов связи. Отмечается применение критерия χ^2, точного теста Фишера и других статистических методов для определения степени семантической близости слов.

Тема 5. Синтагматическая модель машинного анализа текста

Формальное описание синтагм – конструктивные единицы текста (символ, слово, предложение, текст). Построение синтагматических конструктивов: словников, словоуказателей и корпусов текстов. Формирование основы для дальнейшего морфологического, синтаксического и семантического анализа.

Тема 6. Парадигматическая модель машинного анализа текста

Словарно-тезаурусное описание как основа парадигматической модели текста. Задачи преобразования текста в словарь-тезаурус и обратного порождения текста. Построение словарей, тезаурусов и конкордансов. Извлечение знаний из текстов.

Тема 7. Современные вопросы обработки естественного языка

Что такое NLP. Основные задачи NLP. Эволюция методов NLP: правила и шаблоны, статистические модели, нейронные сети, трансформеры. Этапы работы NLP: сбор данных предобработка данных (очистка данных, токенизация, стемминг, лемматизация, разметка данных) и создание датасета. Выбор представления данных

Тема 8. Работа с табличными данными. Задачи классификации и кластеризации

Классификация текстов. Постановка задачи классификации. Примеры задач классификации. Методы задач классификации: логистическая регрессия, решающие деревья, ансамбли моделей. Метрики качества задач классификации. Кластеризация данных. Парадигма обучения с учителем против парадигмы обучения без учителя. Восстановление распределения данных. Постановка задачи кластеризации.

Тема 9. Работа со структурированными данными при помощи нейронных сетей

Модели векторного представления слова и текста: мешок слов, TF-DF. Обучение представлений для слов. Модели, основанные на частоте появления слов. Модели, основанные на прогнозировании (word2vec, glove). Модель Transformer. Обзор моделей на основе трансформеров. BERT. GPT. Различия моделей GPT и BERT. Способы оценки моделей

Тема 10. Технологии text-to-image (T2I) генерации изображений на основе текста

Введение в Text-to-Image генерацию. Эволюция моделей Text-to-Image генерации: GAN, CLIP, DALL-E, диффузионные модели, Latent Diffusion и Stable Diffusion, SD 3, FLUX и интеграция с LLM. Основные архитектуры и принципы работы. Управляемость. Персонализация. Ключевые метрики оценки качества.

Тема 11. Вайб-кодинг: программирование через диалог с AI

Определение и основные концепции. Принципы и рабочий процесс Vibe Coding. Технология вайб-кодинг. Преимущества vibe coding. Недостатки и вызовы vibe coding. Применение Vibe Coding.

3.4 Тематика лабораторных занятий

1. Частотный анализ текстов
2. Сравнение языковых моделей на задаче реферирования
3. Автоматизация аннотации и индексирования с помощью больших языковых моделей
4. Создание интеллектуального агента с динамическим определением задачи на основе пользовательского ввода
5. Сравнение отечественных и зарубежных языковых моделей в машинном переводе
6. Исследование возможностей агентных систем для решения задач разного типа

3.5 Курсовой проект по дисциплине

Курсовое проектирование по данной дисциплине учебным планом запланировано, тема курсовой работы: «Анализ текста информационного ресурса».

Целью курсовой работы является приобретение умений исследования естественно-языкового описания предметной области и машинного анализа текстов. Выполнение курсового проекта студентом предусматривает следующее:

- изучение информационных теоретических материалов: по теории частот слов, формальным языкам информационных технологий, структуре и составу словарной компоненты баз лингвистических знаний, тезаурусам, лингвистическим процессорам и др.;

- ознакомление с исследовательскими практиками: исследование частотных характеристик фрагмента естественно-языкового описания предметной области, исследование словаря определений, индексирование фрагмента текста, построение графа дефиниций и др.
- приобретение навыков анализа и изучение методов и приемов исследования ЕЯ описания предметной области (ПО) в процессе выполнения заданий курсовой работы;
- приобретение знаний и навыков интегрированного использования программного обеспечения АСОИУ (текстовых процессоров, СУБД, электронных таблиц, специального программного обеспечения и других программных средств) для проведения анализа и исследования ЕЯ описания ПО;
- приобретение знаний и навыков по оформлению результатов анализа и исследования ПО при оформлении курсовой работы.

4 Учебно-методическое и информационное обеспечение

4.1 Основная литература

1. Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — Москва : Издательство Юрайт, 2023. — 85 с. — (Высшее образование). — ISBN 978-5-534-15561-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/520544

2. Воронов, М. В. Системы искусственного интеллекта : учебник и практикум для вузов / М. В. Воронов, В. И. Пименов, И. А. Небаев. — Москва : Издательство Юрайт, 2023. — 256 с. — (Высшее образование). — ISBN 978-5-534-14916-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/519916

3. Иванов, В. М. Интеллектуальные системы : учебное пособие для вузов / В. М. Иванов ; под научной редакцией А. Н. Сесекина. — Москва : Издательство Юрайт, 2022. — 91 с. — (Высшее образование). — ISBN 978-5-534-00551-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/492094

4.2 Дополнительная литература

1. Бессмертный, И. А. Системы искусственного интеллекта : учебное пособие для вузов / И. А. Бессмертный. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2023. — 157 с. — (Высшее образование). — ISBN 978-5-534-07467-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/512657

2. Анализ данных : учебник для вузов / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2023. — 490 с. — (Высшее образование). — ISBN 978-5-534-00616-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/511020

4.3 Электронные образовательные ресурсы

1. ЭОР Курс Машинное обучение и методы анализа данных https://online.mospolytech.ru/course/view.php?id=14701

4.4 Лицензионное и свободно распространяемое программное обеспечение

1. Microsoft Windows.
2. Веб-браузер, Chrome.
3. LibreOffice

5 Материально-техническое обеспечение

5.1 Требования к оборудованию и помещению для занятий

Практические занятия (семинары) и самостоятельная работа студентов должна проводиться в специализированных аудиториях с комплектом мультимедийного оборудования и/или доской для записей материалов. Число рабочих мест в аудитории должно быть достаточным для обеспечения индивидуальной работы студентов.

5.2 Требования к программному обеспечению

Для проведения практических занятий (семинаров) специального программного обеспечения для освоения дисциплины не требуется.

6 Методические рекомендации

6.1 Методические рекомендации для преподавателя по организации обучения

1. При подготовке к занятиям следует предварительно проработать материал занятия, предусмотрев его подачу точно в отведенное для этого время занятия. Следует подготовить необходимые материалы – теоретические сведения, задачи и др. При проведении занятия следует контролировать подачу материала и решение заданий с учетом учебного времени, отведенного для занятия.

2. При проверке работ и отчетов следует учитывать не только правильность выполнения заданий, но и оптимальность выбранных методов решения, правильность выполнения всех его шагов.

6.2 Методические указания для обучающихся по освоению дисциплины

Изучение дисциплины осуществляется в строгом соответствии с целевой установкой в тесной взаимосвязи учебным планом. Основой теоретической подготовки студентов являются аудиторные занятия, семинары и практики.

В процессе самостоятельной работы студенты закрепляют и углубляют знания, полученные во время аудиторных занятий, дорабатывают конспекты и записи, готовятся к промежуточной аттестации, а также самостоятельно изучают отдельные темы учебной программы.

На занятиях студентов, в том числе предполагающих практическую деятельность, осуществляется закрепление полученных, в том числе и в процессе самостоятельной работы, знаний. Особое внимание обращается на развитие умений и навыков установления связи положений теории с профессиональной деятельностью будущего специалиста.

Самостоятельная работа осуществляется индивидуально. Контроль самостоятельной работы организуется в двух формах:

- самоконтроль и самооценка студента;
- контроль со стороны преподавателей (текущий и промежуточный).

Текущий контроль осуществляется на аудиторных занятиях, промежуточный контроль осуществляется на зачете в письменной (устной) форме.

Критериями оценки результатов самостоятельной работы студента являются:

- уровень освоения студентом учебного материала;
- умения студента использовать теоретические знания при выполнении практических задач;

- сформированность компетенций;
- оформление материала в соответствии с требованиями.

7 Фонд оценочных средств

7.1 Методы контроля и оценивания результатов обучения

В процессе обучения используются следующие оценочные формы самостоятельной работы студентов, оценочные средства текущего контроля успеваемости и промежуточных аттестаций:

- Курсовой проект;
- экзамен.

7.2 Шкала и критерии оценивания результатов обучения

Показателем оценивания компетенций на различных этапах их формирования является достижение обучающимися планируемых результатов обучения по дисциплине.

Показатель	**Критерии оценивания**			
	2	**3**	**4**	**5**
ОПК-8. Способен разрабатывать алгоритмы и программы, пригодные для практического применения				
ИОПК-8.1. Знает основные технологии создания и внедрения информационных систем, стандарты управления жизненным циклом информационной системы. ИОПК-8.2. Умеет проектировать блок-схемы алгоритмов, оценивать производительность алгоритмов и затраты памяти на работу алгоритма, разрабатывать программы на основе спроектированного алгоритма и проводить отладку программы,	Обучающийся демонстрирует полное отсутствие или недостаточное соответствие материалу дисциплины знаний, указанных в индикаторах компетенций дисциплины.	Обучающийся демонстрирует неполное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины. Допускаются значительные ошибки, проявляется недостаточность знаний, по ряду показателей, обучающийся испытывает значительные затруднения при оперировании знаниями при их переносе на новые ситуации.	Обучающийся демонстрирует частичное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины. Но допускаются незначительные ошибки, неточности, затруднения при аналитических операциях.	Обучающийся демонстрирует полное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины. Свободно оперирует приобретенными знаниями.

применять методы системного анализа и математического моделирования при разработке и эксплуатации ИС, проводить структурный анализ, функциональный анализ, объектно-ориентированный анализ иерархии классов. осуществлять организационное обеспечение выполнения работ на всех стадиях и в процессах жизненного цикла информационной системы. ИОПК-8.3. Владеет навыками разработки программ, построения блок-схем алгоритмов и оценки производительности алгоритмов, работы с унифицированным языком визуального моделирования, составления плановой и отчетной документации по управлению проектами создания информационных систем на стадиях жизненного цикла.				

Шкала оценивания результатов промежуточной аттестации и её описание:
Форма промежуточной аттестации: экзамен.

Промежуточная аттестация обучающихся в форме экзамена проводится по результатам выполнения всех видов учебной работы, предусмотренных учебным планом по данной дисциплине (модулю), при этом учитываются результаты текущего контроля успеваемости в течение семестра. Оценка степени достижения обучающимися планируемых результатов обучения по дисциплине (модулю) проводится преподавателем, ведущим занятия по дисциплине (модулю) методом экспертной оценки. По итогам промежуточной аттестации по дисциплине (модулю) выставляется оценка «отлично», «хорошо», «удовлетворительно» или «неудовлетворительно».

К промежуточной аттестации допускаются только студенты, выполнившие все виды учебной работы, предусмотренные рабочей программой по дисциплине.

Шкала оценивания	Описание
Отлично	Среднее значение для всех формируемых на момент проведения аттестации уровней компетенций – 5. Выполнены все виды учебной работы, предусмотренные учебным планом. Студент демонстрирует соответствие знаний, умений, навыков приведенным в таблицах показателей, оперирует приобретенными знаниями, умениями, навыками, применяет их в ситуациях повышенной сложности. При этом могут быть допущены незначительные ошибки, неточности, затруднения при аналитических операциях, переносе знаний и умений на новые, нестандартные ситуации.
Хорошо	Среднее значение для всех формируемых на момент проведения аттестации уровней компетенций – 4. Выполнены все виды учебной работы, предусмотренные учебным планом. Студент демонстрирует неполное, правильное соответствие знаний, умений, навыков приведенным в таблицах показателей, либо если при этом были допущены 2-3 несущественные ошибки.
Удовлетворительно	Среднее значение для всех формируемых на момент проведения аттестации уровней компетенций – 3. Выполнены все виды учебной работы, предусмотренные учебным планом. Студент демонстрирует соответствие знаний, в котором освещена основная, наиболее важная часть материала, но при этом допущена одна значительная ошибка или неточность.
Неудовлетворительно	Не достигнуто пороговое значение хотя бы для одного уровня формируемых на момент проведения аттестации компетенций. Не выполнен один или более видов учебной работы, предусмотренных учебным планом. Студент демонстрирует неполное соответствие знаний, умений, навыков приведенным в таблицах показателей, допускаются значительные ошибки, проявляется отсутствие знаний, умений, навыков по ряду показателей, студент испытывает значительные затруднения при оперировании знаниями и умениями при их переносе на новые ситуации.

7.3 Оценочные средства

7.3.1 Экзаменационные вопросы

1. Что такое обработка естественного языка?
2. Какие основные задачи стоят перед машинным анализом текстов?
3. Что включает в себя распознавание языка?

4. В чем суть классификации текстов?
5. Как осуществляется кластеризация текстовых данных?
6. Что значит извлечение информации из текста?
7. Что понимается под рекомендательной системой на основе анализа текста?
8. Какие задачи решает автоматический перевод и аннотирование текста?
9. Что входит в информационный поиск с использованием методов обработки естественного языка?
10. Что представляют собой чат-боты и системы разговорных интерфейсов?
11. В чем заключается обработка текста на изображении?
12. Для чего нужны системы анализа и генерации текста на основе искусственного интеллекта?
13. Как оценить качество системы обработки и анализа текста?
14. Каковы этические аспекты применения искусственного интеллекта в анализе текста?
15. Какие существуют подходы к обработке естественного языка?
16. Какие алгоритмы машинного обучения применяются в обработке естественного языка и машинном анализе текстов?
17. Какие инструменты и библиотеки используются для обработки естественного языка на языке Python?
18. Что представляет собой классификация текстов с использованием машинного обучения?
19. Как работает кластеризация текста с использованием алгоритмов машинного обучения?
20. В каких сферах применяются системы обработки естественного языка и машинного анализа текста?
21. Какие функции выполняют рекомендательные системы на основе анализа текстовых данных?
22. Как происходит автоматический перевод текста с использованием нейронных сетей?
23. Какие подходы используются для создания систем генерации текста на основе машинного обучения?
24. Что означает понятие “обработка текста на изображении” в контексте обработки естественного языка?
25. В чём заключается работа чат-бота с использованием обработки естественного языка?
26. Как можно оценить качество работы системы обработки и анализа текстовой информации?
27. Каким образом происходит извлечение информации из текстовой информации с помощью машинного обучения?
28. В каких случаях применяются системы анализа и генерации текстов?
29. Какие технологии используются для создания чат-ботов с использованием обработки естественного языка?
30. Какие принципы и задачи лежат в основе машинного анализа текстов?

7.3.2 Типовые практические задания

1. Разработка алгоритма машинного обучения для классификации текстов по жанрам (например, роман, детектив, научная фантастика).
2. Создание системы автоматического аннотирования научных статей на основе анализа их заголовков и ключевых слов.
3. Реализация системы распознавания языка с использованием метода n-грамм и машинного обучения для определения языка текста.

4. Разработка рекомендательной системы для книг на основе анализа пользовательских предпочтений и отзывов других читателей.
5. Создание чат-бота для поддержки пользователей на основе обработки естественного языка с использованием правил и машинного обучения.
6. Разработка системы машинного анализа текстов для поиска информации в больших текстовых коллекциях.
7. Создание системы машинного перевода текста на основе использования нейросетей и больших языковых моделей.
8. Реализация системы генерации текстов на заданную тему с использованием шаблонов и машинного обучения (например, генерация описаний товаров для интернет-магазина).
9. Разработка программы для анализа тональности текстов с использованием методов машинного обучения и обработки естественного языка (например, определение позитивного, негативного или нейтрального тона текста).
10. Создание системы анализа и классификации текстов на основе алгоритмов машинного обучения и лингвистических методов.

7.3.3 Типовой экзаменационный билет

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(МОСКОВСКИЙ ПОЛИТЕХ)

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №1
по дисциплине
«Обработка естественного языка и машинный анализ текстов»
направление подготовки 09.03.01 Информатика и вычислительная техника

ВОПРОСЫ:

1. Какие принципы и задачи лежат в основе машинного анализа текстов?
2. Как происходит автоматический перевод текста с использованием нейронных сетей?
3. Что значит извлечение информации из текста?

Утверждено: _____ / _____ / «__» _____ 20__г.