

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор высшей школы
программной инженерии
А.В. Малеев**

	Рабочая программа дисциплины (модуля)
по дисциплине:	Обработка естественного языка
по направлению:	Программная инженерия
профиль подготовки:	Разработка программно-информационных систем высшая школа программной инженерии высшая школа программной инженерии
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 18 час.

Подготовка к экзамену: 30 час.

Всего часов: 108, всего зач. ед.: 3

Программу составил: В.А. Малых, канд. техн. наук, доцент

Программа обсуждена на заседании высшей школы программной инженерии 19.03.2025

Аннотация

Этот курс посвящен основам разработки программного обеспечения. Правильный дизайн - важная часть любого проекта. Этот курс охватывает основы языка программирования Python, основные концепции и языковые конструкции. Наряду с этим этот курс предоставляет инструменты для использования языка программирования Python в сложных проектах. Вы получите представление о правильном дизайне кода, поддержании кодовой базы и интеграции ваших приложений с другими.

1. Цели и задачи

Цель дисциплины

- Научитесь писать эффективный и читаемый код.
- Изучите передовой опыт разработки программного обеспечения.
- Получите необходимый опыт работы с Python.
- Привыкайте к тестированию и документированию кода.
- Будьте готовы внедрить методы машинного обучения и глубокого обучения.

Задачи дисциплины

- Разработка программного обеспечения
- Python
- Тестирование
- Работа в разных средах.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
ОПК-5 Способен устанавливать программное и аппаратное обеспечение для информационных и автоматизированных систем	ОПК-5.2 Способен оценивать концепции и атрибуты качества программного обеспечения (надежности, безопасности, удобства использования), в том числе роль людей, процессов, методов, инструментов и технологий обеспечения качества
ОПК-6 Способен разрабатывать алгоритмы и программы, пригодные для практического использования, применять основы информатики и программирования к проектированию, конструированию и тестированию программных продуктов	ОПК-6.2 Умеет применять языки программирования для решения прикладных задач
	ОПК-6.3 Знает методы тестирования программного кода на ошибки и способен проводить тестирование на различных уровнях (модульное, интеграционное, системное)
	ОПК-6.4 Имеет навыки программирования и тестирования программных продуктов
ПК-3 Способен проектировать, разрабатывать, интегрировать, проверять на работоспособность программное обеспечение	ПК-3.1 Различает синтаксис языков программирования, особенности программирования на этих языках, стандартные библиотеки языков программирования
	ПК-3.2 Умеет выбирать языки программирования для написания программного кода с учетом технического задания
	ПК-3.3 Умеет излагать основные принципы построения и виды архитектуры программного обеспечения, методы и средства проектирования программного обеспечения, методологии разработки программного обеспечения и технологии программирования
ПК-6 Способен разрабатывать и внедрять	ПК-6.1 Знает, как создавать стандарты и методологии разработки программного обеспечения в организации

стандарты и процессы разработки,
производить их мониторинг и обновления

ПК-6.3 Владеет навыками мониторинга и обновления
стандартов с учетом изменяющихся требований и
технологий

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- представление о строении, функционировании зрительного анализатора;
- представление о психофизиологических и информационных моделях бинокулярного зрения;
- принципы функционирования видеоинтерфейса применительно к системам VR/AR.

уметь:

- принципы функционирования и методологию разработки распределенных систем применительно к задачам создания систем VR/AR;
- строение и принципы функционирования существующих и перспективных графических API.

владеть:

- методологией разработки ПМО всех звеньев систем VR/AR (включая графическое ядро, подсистемы управления виртуальной средой, видеоинтерфейс и др.);
- объектно-ориентированной методологией проектирования и разработки программного кода для всего спектра задач создания систем VR/AR.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение в обработку текстов	2	2		1
2	Методы сбора и хранения данных	2	2		1
3	Частотный анализ текстов	2	2		1
4	Морфологический анализ и разрешение неоднозначности	2	2		1
5	Синтаксический анализ. Универсальные зависимости	2	2		1
6	Выделение ключевых слов и словосочетаний	2	2		1
7	Векторная модель текста и слова, методы снижения размерности	2	2		1
8	Классификация текстов	2	2		1
9	Языковые модели	2	2		1
10	Классификация последовательностей	2	2		1
11	Суммаризация текстов, вопросно-ответные системы	2	2		1
12	Исправление опечаток	2	2		1
13	Обработка речи, речевые технологии	2	2		2
14	Информационный поиск	2	2		2
15	Мультимодальная обработка текстов	2	2		2
Итого часов		30	30		18
Подготовка к экзамену		30 час.			

Общая трудоёмкость	108 час., 3 зач.ед.
--------------------	---------------------

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 6 (Весенний)

1. Введение в обработку текстов

Основные задачи обработки и анализа текстов. Актуальность обработки и анализа текстов. Краткий исторический экскурс по обработке и анализу текстов. Обзор существующих систем обработки и анализа текстов. Классификация систем обработки и анализа текстов.

2. Методы сбора и хранения данных

Форматы данных, способы хранения, принципы работы интернета. Краулинг. Regexp. Unicode.

3. Частотный анализ текстов

Модель мешка слов. Закон Ципфа. Закон Хипса. Векторное представление текстов. Релевантность в векторной модели. Расширения модели мешка слов. Реализация модели мешка слов в библиотеках Gensim и NLTK.

4. Морфологический анализ и разрешение неоднозначности

Задача морфологического анализа. Типы языков. Алгоритмы морфологического разбора. Морфологическая разметка. Омонимия и неоднозначность. Алгоритм разрешения омонимии. Скрытые Марковские модели. Декодирование в скрытых Марковских моделях.

5. Синтаксический анализ. Универсальные зависимости

Задача синтаксического разбора предложений. Модель составляющих. Вероятностные контекстно-свободные грамматики. Модель зависимостей. Универсальные зависимости. Парсинг зависимостей. Архитектура SyntaxNet.

6. Выделение ключевых слов и словосочетаний

Лексический анализ. Словари и тезаурусы. Поиск синонимов. Частотные методы выделения ключевых слов и словосочетаний. Метрики совместной встречаемости. Выделение ключевых словосочетаний по морфологическим шаблонам. Выделение ключевых словосочетаний по синтаксическим шаблонам. Алгоритмы RAKE и TextRank. Программные средства для выделения ключевых слов: NLTK, Томига-парсер.

7. Векторная модель текста и слова, методы снижения размерности

Векторная модель документа, векторная модель слова. Поиск похожих текстов. Косинусная мера близости. Методы снижения размерности в векторной модели документа: сингулярное разложение, латентный семантический анализ. Связь с моделями скрытых тем. Латентное размещение Дирихле (LDA). Параметры модели. Выбор числа скрытых тем. Расширения модели LDA. Дистрибутивная семантика, векторная модель слова. Построение матрицы PPMI. Поиск близких слов по значению. Снижение размерности и факторизация матрицы PPMI. Эмбединги: word2vec, GloVe, AdaGram. Обучение моделей word2vec. Отрицательное сэмплирование.

8. Классификация текстов

Задачи классификации текстов и предложений по теме, тональности и жанру. Метод наивного Байеса, метод максимальной энтропии. Сверточные нейронные сети. Архитектура FastText.

9. Языковые модели

Счетные языковые модели. Проблема нулевых вероятностей. Преобразование Лапласа, преобразование Гуд-Тьюринга. Вероятностные нейронные языковые модели. Генерация текстов. Рекуррентные нейронные сети.

10. Классификация последовательностей

Задача классификации последовательностей. Частеречная разметка, определение семантических ролей, извлечение именованных сущностей. IOB разметка, IOBES разметка. Условные случайные поля.

11. Суммаризация текстов, вопросно-ответные системы

Абстрактивная и генеративная суммаризация текстов. Алгоритм TextRank. Вопросно-ответные системы. Архитектура энкодера-декодера для вопросно-ответных систем и чат-ботов.

12. Исправление опечаток

Модель зашумленного канала. Исправление опечаток по правилам. Редакционное расстояние.

13. Обработка речи, речевые технологии

Распознавание речи. Генерация речи.

14. Информационный поиск

Понятие релевантности. Использование векторной модели в задаче поиска. Косинусная мера релевантности. Использование языковой модели в задаче поиска. Обучение ранжированию. A|B - тестирование.

15. Мультимодальная обработка текстов

Связь обработки текстов с обработкой изображений. Генерация изображения по тексту. Поиск изображения по описанию.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Стандартная аудитория

6. Перечень рекомендуемой литературы

Основная литература

Фонд библиотеки МФТИ:

Маркус, Г. Искусственный интеллект: Перезагрузка. Как создать машинный разум, которому действительно можно доверять : практическое руководство / Г. Маркус, Э. Дэвис. - Москва : Альпина ПРО, 2021. - 300 с. - ISBN 978-5-907394-93-3. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/1905852>

Литература:

Маннинг, К. Д. Введение в информационный поиск = Introduction to Information Retrieval / К. Д. Маннинг, П. Рагхаван, Х Шютце ; перевод с английского Д. А. Ключина ; под редакцией П. И. Браславского [и др.] URL: <https://www.labirint.ru/books/736133/>

Дополнительная литература

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

<http://dm.fizteh.ru/>

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

Мультимедийные технологии можно использовать на лекциях и практических занятиях, в том числе на презентациях.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой.

Самостоятельная работа включает в себя:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к практическим занятиям, выполнение нескольких индивидуальных домашних заданий.

Промежуточный контроль знаний проводится в виде письменных опросов по теории. Кроме этого в ходе освоения курса студент должен выполнить проект, содержащий несколько взаимосвязанных заданий с их последующей защитой.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Программная инженерия
профиль подготовки:	Разработка программно-информационных систем высшая школа программной инженерии высшая школа программной инженерии
курс:	<u>3</u>
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Экзамен

Разработчик: В.А. Малых, канд. техн. наук, доцент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
ОПК-5 Способен устанавливать программное и аппаратное обеспечение для информационных и автоматизированных систем	ОПК-5.2 Способен оценивать концепции и атрибуты качества программного обеспечения (надежности, безопасности, удобства использования), в том числе роль людей, процессов, методов, инструментов и технологий обеспечения качества
ОПК-6 Способен разрабатывать алгоритмы и программы, пригодные для практического использования, применять основы информатики и программирования к проектированию, конструированию и тестированию программных продуктов	ОПК-6.2 Умеет применять языки программирования для решения прикладных задач
	ОПК-6.3 Знает методы тестирования программного кода на ошибки и способен проводить тестирование на различных уровнях (модульное, интеграционное, системное)
	ОПК-6.4 Имеет навыки программирования и тестирования программных продуктов
ПК-3 Способен проектировать, разрабатывать, интегрировать, проверять на работоспособность программное обеспечение	ПК-3.1 Различает синтаксис языков программирования, особенности программирования на этих языках, стандартные библиотеки языков программирования
	ПК-3.2 Умеет выбирать языки программирования для написания программного кода с учетом технического задания
	ПК-3.3 Умеет излагать основные принципы построения и виды архитектуры программного обеспечения, методы и средства проектирования программного обеспечения, методологии разработки программного обеспечения и технологии программирования
ПК-6 Способен разрабатывать и внедрять стандарты и процессы разработки, производить их мониторинг и обновления	ПК-6.1 Знает, как создавать стандарты и методологии разработки программного обеспечения в организации
	ПК-6.3 Владеет навыками мониторинга и обновления стандартов с учетом изменяющихся требований и технологий

2. Показатели оценивания компетенций

В результате изучения дисциплины «Обработка естественного языка» обучающийся должен:

знать:

- представление о строении, функционировании зрительного анализатора;
- представление о психофизиологических и информационных моделях бинокулярного зрения;
- принципы функционирования видеоинтерфейса применительно к системам VR/AR.

уметь:

- принципы функционирования и методологию разработки распределенных систем применительно к задачам создания систем VR/AR;
- строение и принципы функционирования существующих и перспективных графических API.

владеть:

- методологией разработки ПМО всех звеньев систем VR/AR (включая графическое ядро, подсистемы управления виртуальной средой, видеоинтерфейс и др.);
- объектно-ориентированной методологией проектирования и разработки программного кода для всего спектра задач создания систем VR/AR.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

В чем разница между изменяемыми и неизменяемыми объектами в Python? Каковы преимущества использования изменяемых или неизменяемых типов? Что такое «вызов по назначению»?

2. Что такое закрытие в Python? Когда мы сможем это использовать?
3. Что такое декоратор в Python? Как реализовать собственный декоратор? (Декоратор с параметрами)
4. Что такое диспетчер контекста в Python? использование
5. Генераторы и итераторы в питоне? Реализация, использование
6. Наследование классов, мпо, миксины
7. GIL
8. Многопроцессорность и многопоточность, в чем разница
9. Как мы можем сделать частные атрибуты в классе?
10. Интерфейс класса для создания хешируемых объектов.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Докажите, что если m , n - два взаимно простых целых числа разной четности, то числа $m^2 - n^2$ и $2mn$ также взаимно просты.
2. Напишите и докажите общую формулу для количества различных представлений данного целого числа n в виде суммы двух квадратов. (Представители, которые не получены друг от друга путем изменения знаков и порядка слов, считаются разными.)
3. На основе полученной формулы выведите нижнюю границу максимального числа равных расстояний между заданными n точками на плоскости, используя правильную прямоугольную решетку.
4. Постройте правильный пятиугольник с помощью циркуля и линейки.
5. Постройте правильный 15-угольник, используя циркуль и линейку.
6. Вам дается один сегмент. Требуется построить с помощью циркуля и линейки отрезок длины x , удовлетворяющий уравнению
7. Основываясь на предыдущем задании, докажите, что правильный семиугольник нельзя построить с помощью циркуля и линейки.
8. Докажите, что трисекция угла невозможна.
9. Опишите все возможные комбинации количества черных и белых шаров в урне для голосования, чтобы при случайном вылове двух шаров в выборке без возврата, вероятность вылова двух белых шаров составляла точно 0,5.
10. Рассмотрим соотношение сторон a , b , c треугольника, в котором треугольник с вершинами в основании биссектрис равнобедренный. Предполагая, что стороны, сходящиеся на стороне с большого треугольника, равны, сведем это соотношение к следующему
11. Далее мы рассматриваем куб, определяемый первым из трех уравнений (отказ от требования, чтобы a , b , c были сторонами треугольника). Покажите, что полученный куб неразложим, то есть определяющий его многочлен не учитывается.
12. В дополнение к этому, покажите, что наш куб неособен, то есть нет ни одной точки на его проективизации, в которой каждое направление касалось бы (или того же самого, в котором все три первые частных производных многочлена, определяющего его, вырождают).

Примеры экзаменационных билетов

Билет №1

1. Напишите и докажите общую формулу для количества различных представлений данного целого числа n в виде суммы двух квадратов.
2. Докажите, что трисекция угла невозможна.

Билет №2

1. Рассмотрим соотношение сторон a , b , c треугольника, в котором треугольник с вершинами в основании биссектрис равнобедренный.

2. Опишите все виды комбинаций чисел черных и белых шаров в урне для голосования, чтобы, если два шара случайно выловлены в выборке и не вернулись, вероятность вылова двух белых шаров была точно 0,5.

Критерии оценивания

Оценка «отлично (10)» выставляется студенту, проявившему всестороннее, систематическое и глубокое знание материала образовательной программы, самостоятельно выполнившего все задания, предусмотренные программой, глубоко изучившему основную и дополнительную литературу, рекомендованную программой. , активно работает в классе и понимает основные научные концепции по изучаемой дисциплине, проявил творческий подход и научный подход в понимании и представлении материала образовательной программы, ответ на который характеризуется использованием богатых и адекватных терминов, а также последовательным и логичным изложением материала;

Оценка «отлично (9)» выставляется студенту, который продемонстрировал всестороннее систематическое знание материала образовательной программы, самостоятельно выполнил все задачи, предусмотренные программой, глубоко усвоил основную литературу и знаком с рекомендуемой дополнительной литературой. по программе, активно проработал на занятиях, показал системность знаний по дисциплине, достаточную для дальнейшего изучения, а также умение самостоятельно расширять ее, ответ которой отличается точностью используемых терминов, а изложение материала в нем последовательное и логичное;

Оценка «отлично (8)» выставляется студенту, который проявил полное знание материала образовательной программы, не допускает существенных неточностей в своем ответе, самостоятельно выполнил все задания, предусмотренные программой, изучил основную литературу, рекомендованную учебной программой. программа, активно проработанная на занятиях, показала системность его знаний по дисциплине, достаточных для дальнейшего изучения, а также способность самостоятельно их усиливать;

Оценка «хорошо (7)» выставляется студенту, который проявил достаточно полное знание материала образовательной программы, не допускает существенных неточностей в ответе, самостоятельно выполнил все задания, предусмотренные программой, изучил основную рекомендованную литературу по программе, активно работал на занятиях, проявил системность своих знаний по дисциплине, достаточных для дальнейшего изучения, а также способность самостоятельно их усиливать;

Оценка «хорошо (6)» выставляется студенту, который проявил достаточно полное знание материала образовательной программы, не допускает существенных неточностей в своем ответе, самостоятельно выполнил основные задачи, предусмотренные программой, изучил основную литературу. рекомендован программой, показал систематичность своих знаний по дисциплине, достаточную для дальнейшего изучения;

Оценка «хорошо (5)» дается студенту, продемонстрировавшему знание материала основной образовательной программы в объеме, необходимом для дальнейшего обучения и будущей работы по профессии, который, не проявляя достаточной активности на уроках, тем не менее самостоятельно выполнял, овладел основными задачами, предусмотренными программой, освоил основную литературу, рекомендованную программой, допустил ошибки в их выполнении и ответе во время тестирования, но имеет необходимые знания для исправления этих ошибок самостоятельно;

Оценка «удовлетворительно (4)» дается студенту, обнаружившему знание материала основной образовательной программы в объеме, необходимом для дальнейшего обучения и будущей работы по профессии, который, не проявляя достаточной активности на уроках, тем не менее самостоятельно выполнял, выполнил основные задачи, предусмотренные программой, изучил основную литературу, но допустил ошибки в их выполнении и в своем ответе во время теста, но имеет необходимые знания для исправления этих ошибок под руководством преподавателя;

Оценка «удовлетворительно (3)» выставляется обучающемуся, проявившему знание материала основной образовательной программы в объеме, необходимом для дальнейшего обучения и будущей работы по профессии, не проявившего активности на занятиях, самостоятельно выполнившего основные задания, предусмотренные законодательством. программа, но допускающая ошибки в их выполнении и в ответе при тестировании, но обладающая необходимыми знаниями для устранения под руководством преподавателя наиболее существенных ошибок;

Оценка «неудовлетворительно (2)» выставляется студенту, который показал пробелы в знаниях или недостаток знаний по значительной части материала основной образовательной программы, не выполнил самостоятельно основные задания, требуемые программой, допустил принципиальные ошибки в выполнении предусмотренных программой задач, не имеющего возможности продолжить учебу или начать профессиональную деятельность без дополнительной подготовки по данной дисциплине;

Оценка «неудовлетворительно (1)» ставится студенту при отсутствии ответа (отказ от ответа) или когда представленный ответ не соответствует сути вопросов, содержащихся в задании.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Во время дифференцированного зачета студенту разрешается использовать программу дисциплины.