ИРИТ-РТФ УрФУ

Название курса: «Анализ больших данных в информационной безопасности».

Осенний семестр 2023

Преподаватель: Чернышов Юрий Юрьевич, к.ф.-м.н., доцент кафедры «Информационные технологии и системы управления» ИРИТ-РТФ УрФУ.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Название | Краткое содержание | Вопросы к экзамену |
| 1.1 | Лекция "Введение" | Задачи анализа больших данных. Источники данных. Роли в проекте. Инструменты (python, библиотеки). Примеры проектов. | Вопрос 1: Какие задачи можно решать с помощью данных? Приведите примеры таких задач и опишите данные, которые нужны для решения? Приведите примеры задач, связанных с аналитикой данных, из ИБ.  Вопрос 2: Приведите примеры источников данных, пригодных для создания интеллектуальных алгоритмов. Что может являться помехой для эффективного использования данных? Какие используются форматы данных? Что в ИБ может являться источником данных для анализа. |
| 1.2 | Практика "Программирование на Python" | Основы Python   * История создания, применение Python * PEP, PEP8 * Импорт модулей * Типы данных, структуры, динамическая типизация * Условные выражения, операторы и циклы | Задача 1: Написать python скрипт для генерации python словаря с использованием подхода "dictionary comprehension", где ключом будут целые числа от 1 до 15 с шагом 2, а значениями - случайные числа от 0 до 100. Создать переменную, в которой будет храниться ссылка на этот словарь. Обратиться к несуществующему элементу созданного словаря так, чтобы не было вызвано исключение, а вернулось строка "unknown key"  Задача 2: Напишите функцию, которая вычисляет n-е число Фибоначчи. |
| 1.3 | Практика "Программирование на Python" | Основы Python   * Функции * Обработка исключений * ООП | Задача 3: Создайте функцию, которая принимает на вход любое количество позиционных и именованных аргументов и возвращает кортеж из списка позиционных элементов в обратном порядке, а также произвольный именованный элемент. Если именованных элементов нет, то возвращается пустой словарь.  Задача 4: Создайте класс Person, который описывает человека. Добавьте соответствующие методы класса, которые описывают действия, совершаемые человеком, а также свойства класса, которые характеризуют человека. Создайте несколько разных объектов этого класса. |
| 2.1 | Лекция "Анализ данных систем мониторинга" | Применение анализа данных в системах мониторинга. Виды аналитики (описательная, объясняющая, предиктивная, предписывающая). Обнаружение аномалий. Примеры применения в кибербезопасности. | Вопрос 3: В чем состоит различие между описательной, предиктивной и предписывающей аналитикой. Приведите примеры применения в задачах ИБ этих типов аналитики. Какой вид аналитики характеризуется большим вмешательством человека, а какой меньшим, и почему.  Вопрос 4: Что такое аномалия, какие виды аномалий бывают, какие способы их обнаружения? Приведите примеры аномалий в задачах ИБ. |
| 2.2 | Практика "Анализ данных" | Основы анализа данных   * Анализ данных в Python * Pandas * SciPy * NumPy * Matplotlib * Seaborn * Plotly | Задача 5: Создайте pandas датафрейм с именем df (размер df.shape = (100, 5)) из python словаря, в котором ключами являются названия колонок (названия выберите сами), а значениями - случайные целочисленные значения. Выведите в консоль часть датафрейма df с первыми двумя колонками и строками с 5 по 10.  Задача 6: Загрузите библиотеку matplotlib. Создайте несколько произвольных графиков различного типа. Нарисуйте диаграмму рассеивания (scatter plot) и гистограмму для значений df по произвольной колонке. Диаграмма должна быть с указанием заголовка и легендой. |
| 2.3 | Практика "Подготовка данных и обучение модели" | Основы подготовки данных для ML   * Жизненный цикл моделей ML * Этапы подготовки данных * Scikit-Learn * Очистка данных * Предобработка данных * Трансформация данных * Feature Engineering * Пример обучения модели регрессии и классификации | Задача 7: Для чего нужно разделение выборки на тренировочную и тестовую в машинном обучении? Загрузите набор данных "Ирисы" sklearn.datasets.load\_iris и разделите его с помощь функции sklearn на тренировочную и тестовую выборку.  Задача 8: Какие подходы для обработки пустых (пропущенных) значений вы можете назвать? Создайте датасет pandas с пропущенными значениями. Примените несколько методов заполнения пропущенных значений. В наборе данных Kaggle Titanic в поле "Возраст" (Age) много пропущенных значений. Предложите эффективные методы заполнения этих пропущенных значений. |
| 3.1 | Лекции "Методы хранения и анализа информации" | Типы БД  SQL:   * выборки * манипулирование данными * работа с таблицами | Вопрос 5: Какие бывают и чем отличаются типы БД? Какие бывают СУБД?  Вопрос 6: Что такое SQL? Опишите порядок и этапы исполнения SQL-запроса. Приведите пример SQL запроса с использованием агрегирующих функций. |
| 3.2 | Практика | Получение данных из XLS  Работа с SQLite из Python | Задача 9: напишите python скрипт, который создаёт в SQLite связанные таблицы "Teachers", "Subjects" (один учитель может вести несколько предметов, один предмет может быть рассказан только одним учителем), заполняет их данными.  Задача 10: напишите python скрипт, который получает из SQLite предметы "Subjects" и группирует (GROUP BY) их по преподавателям. |
| 3.3 | Практика | Поиск злоумышленников в массиве данных оператора связи (расследование) | Задача 11: Опишите алгоритм работы полного соединения (FULL JOIN). Создайте небольшую базу данных из двух таблиц и придемонстрируйте различные типы JOIN-ов (LEFT, RIGHT, FULL).  Задача 12: Что такое и для чего используется обобщённое табличное выражение (оператор WITH)? Напишите пример запроса с помощью SQLite. |
| 4.1 | Лекция "Машинное обучение" | Основы нейронных сетей. | Вопрос 7: Что такое машинное обучение? Что такое признаки, параметры, гиперпараметры, целевая переменная, метрики? Что такое переобучение, как его обнаружить и как с ним бороться?  Вопрос 8: Что такое полносвязная нейронная сеть? Каким образом происходит обучение нейронной сети? Зачем нужны функции активации? Приведите примеры различных функций активации. |
| 4.2 | Практика | Классификация (DNN TF) | Данные: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data>  Задача 13: Обучить и оценить модель машинного обучения решающая задачу обнаружения мошенничества (бинарная классификация) с помощью логистической регрессии  Задача 14: Обучить и оценить модель машинного обучения решающая задачу обнаружения мошенничества (бинарная классификация) с помощью полносвязной нейронной сети (архитектура произвольная). |
| 4.3 | Практика | Кластеризация. | <https://www.kaggle.com/code/utkarshsaxenadn/shop-customer-clustering/notebook>  Задача 15: Обучить  и оценить модель кластеризации клиентов магазина методом K-Means  Задача 16: Обучить  и оценить модель кластеризации клиентов магазина методом DBSCAN |
| 5.1 | Лекция "NLP для ИБ - 1" | NLP. токенизация, очистка, векторизация, классификаторы.  Типовые задачи ИБ с применением NLP: детекция скама, OSINT, анализ скриптов настройки и конфигурационных файлов. | Вопрос 9: Перечислите типовые задачи, решаемые при работе с текстом на естественном языке. В каких задачах ИБ может понадобится NLP, приведите примеры данных для этих задач? Перечислите и кратко охарактеризуйте методы векторизации текстов. Какие библиотеки можно использовать для этого? Из каких стадий может состоять пайплайн предварительной обработки текста. Какие библиотеки можно использовать для реализации этого пайплайна?  Вопрос 10: Опишите работу модели Word2Vec, её возможные архитектуры, применение, основные преимущества и недостатки. Как работает модель FastText и в чем её отличия от модели Word2Vec? Применение модели, преимущества и недостатки. |
| 5.2 | Практика | Классификация текстов (построение векторных представлений слов (bow, Tf-Ifd) и классификация с помощью алгоритмов машинного обучения) | Задача 17: Определение тематики текста. Загрузить набор данных о тематике статей sklearn.datasets.fetch\_20newsgroups. Выполнить векторизацию текстов, например, с помощью TfidfVectorizer. С использованием произвольной модели классификации построить классификатор тем текстов.  Задача 18: Определение тональности отзыва к фильмам. Загрузить набор данных tensorflow.keras.datasets.imdb. Векторизовать тексты отзывов к фильмам, привести к одинаковой длине. Построить полносвязную нейронную сеть, которая определяет позитивный и негативный отзыв по тексту. |
| 5.3 | Практика | Кластеризация текстов (построение эмбеддингов с помощью Fasttext и кластеризация с помощью алгоритмов машинного обучения) | Задача 19: с использованием эмбеддингов Fasttext построить эмбеддинги текстов в наборе данных sklearn.datasets.fetch\_20newsgroups, провести кластеризацию эмбеддингов текстов и сравнить результаты кластеризации с разметкой (известными классами статей). |
| 6.1 | Лекция "NLP для ИБ - 2" | CNN, RNN.  Attention, transformer.  LLM. | Вопрос 11: Что такое сверточные нейронные сети, ядро, карта признаков? В чем заключаются операции свертки и пулинга? Применение сверточных сетей в обработке естественного языка.  Вопрос 12: Что такое рекуррентные нейронные сети? Принцип работы рекуррентных сетей. Основные архитектуры рекуррентных сетей. Их достоинства и недостатки. Применение рекуррентных сетей в обработке естественного языка.  Вопрос 13:  Архитектура модели Transformer и её основные компоненты. Механизм внимания. Что такое BERT и GPT? Какие большие языковые модели вы знаете? Приведите пример промпта к большой языковой модели, решающей задачу из области ИБ. |
| 6.2 | Практика | Классификация текстов с помощью моделей глубокого обучения | Задача 20:  Определение тональности отзыва к фильмам. Загрузить набор данных tensorflow.keras.datasets.imdb. Векторизовать тексты отзывов к фильмам, привести к одинаковой длине. Построить рекуррентную нейронную сеть (например, LSTM), которая определяет позитивный и негативный отзыв по тексту. |
| 7.1 | Лекция "Компьютерное зрение в задачах ИБ" | Задачи CV: классификация, детекция, сегментация, трекинг. Инструменты. CNN. Типовые архитектуры (UNet, Yolo). Задачи биометрии. Метрики CV. | Вопрос 14: Типовые задачи компьютерного зрения и архитектуры их решающие. Примеры применение компьютерного зрения в задачах ИБ. |
| 7.2 | Практика | Классификация изображений с помощью глубоких сверточных сетей. Задачи биометрии. | Задача 21:  Классификация изображений с использованием сверточных нейронных сетей. Загрузить набор данных tensorflow.keras.datasets import cifar10. Построить сверточную сетку, которая определяет класс, к которому относится изображение. |
| 7.3 | Практика | Сегментация изображений | Задача 22: Классификация изображений на датасете <https://github.com/ari-dasci/OD-WeaponDetection/tree/master/Pistol%20classification> |