Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Центр сопровождения образовательных инициативных проектов

УТВЕРЖДЕНО:

Руководитель сетевой ОПОП
В.В. Кашпур

Рабочая программа дисциплины

Поиск и сбор аналитических данных

по направлению подготовки

09.04.03 Прикладная информатика

Направленность (профиль) подготовки:
«Дата-аналитика для бизнеса»

Форма обучения
Очная

Квалификация
Магистр

Год приема
2023

Томск – 2023

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-3 Способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;

ПК-1 Способность управлять получением, хранением, передачей, обработкой больших данных.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-3.1 Осуществляет сбор, обработку и анализ научно-технической информации, необходимой для решения профессиональных задач;

ИПК-1.1 Осуществляет мониторинг и оценку производительности обработки данных в организации, разработку предложений по повышению производительности обработки больших данных;

ИПК-1.4 Разрабатывает предложения по развитию и совершенствованию системы получения, хранения, передачи, обработки больших данных.

2. Задачи освоения дисциплины

Раздел 1: Поиск данных

- научиться различать данные и информацию, разные типы и форматы данных;
- получить представление об основных инструментах для обработки разных форматов данных;
- научиться определять способы определения необходимых для решения задачи данных, формулировать требования к данным, различные источники данных;
- научиться расставлять приоритеты при выборе сбора данных;
- освоить способы хранения и описания набора данных (словарь данных);
- научиться писать техническое задание на сбор данных;
- научиться анализировать и оценивать качество данных;
- получить представление о способах обеспечения качества данных;
- освоить основы организации процесса разметки данных, основные принципы разметки данных, виды и инструменты разметки данных с учетом их специфики;
- научиться обеспечивать надежность разметки данных;
- получить представление об альтернативах разметки данных: слабое обучение и активное обучение.

Раздел 2: Сбор данных с использованием API

- освоить основы использования API;
- получить представление, что такое REST и SOAP API, как REST API и SOAP связаны с HTTP протоколом;
- освоить основы работы клиент-серверного общения, структуры запросов и ответов, получить представление о существующих методах запросов;
- научиться находить документацию к API, зарегистрировать свое приложение и получить токен доступа API, находить необходимые методы в документации;
- научиться выполнять запросы, получать код состояния ответа, обрабатывать ошибки запросов, передавать и просматривать заголовки сообщений, получать содержимое ответа;
- научиться выполнять запрос к API с помощью requests, преобразовывать объект Python в объект JSON;
- научиться создавать сессию для доступа к API;

- научиться создавать файл с CSV форматом и записывать в него данные с помощью Python.

Раздел 3: Парсинг данных с сайтов

- получить представление о том, что такое парсинг, зачем он нужен, что такое программа-парсер и как она работает;
- научиться работать с панелью инструментов разработчика;
- научиться с помощью requests создавать сессию и получать с ее помощью HTML код;
- научиться преобразовывать HTML в дерево объектов BeautifulSoup;
- научиться обрабатывать специальные коды HTML;
- научиться собирать данные со всех страниц;
- научиться собирать данные с помощью селекторов (BeautifulSoup);
- научиться устанавливать драйвера браузера для имитации работы браузера, эмулировать работу браузера, получать элементы исходного кода с помощью Selenium;
- научиться взаимодействовать с элементами интерфейса, использовать селекторы в Selenium;
- научиться обрабатывать полученные данные с помощью BS4 (Selenium + BeautifulSoup), собирать данные в CSV файл;
- научиться создавать свой проект Scrapy, использовать интерпретатор командной строки, задавать правила сбора данных, выполнять поиск по исходному коду, запускать “паука”.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к Блоку 1 «Дисциплины (модули)».
Дисциплина относится к обязательной части образовательной программы.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Семестр 1, зачет.
Семестр 2, зачет с оценкой.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения предшествующих дисциплин первого семестра: Основы системного мышления; Python для анализа данных; Чистка, обработка и исследовательский анализ данных.

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 5 з.е., 180 часов, из которых:
– лекционные занятия: 18 ч.;
– практические занятия: 46 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Раздел 1: Поиск данных

Тема 1. Данные.

Краткое содержание темы. Данные. Отличия данных от информации. Типы данных. Отличия типов данных. Большие данные. Основные форматы данных. Отличия форматов данных. Обзор инструментов для обработки разных форматов данных. Синтетические данные: типы, способы генерации.

Тема 2. Сбор данных

Краткое содержание темы. Способы определениях необходимых данных для решения задачи. Требования к данным. Типы данных с точки зрения их получения. Способы получения данных. Преимущества и недостатки разных способов получения данных. Источники данных. Открытые источники данных. Способы отбора необходимых для решения задачи данных. Расстановка приоритетов при выборе сбора данных. Покупка данных. Хранение и описание набора данных (словарь данных). Написание технического задания на сбор данных.

Тема 3. Качество данных

Краткое содержание темы. Качество данных. Аспекты качества данных. Проблемы с качеством данных. Обеспечение качества данных. Оценка качества данных. Предвзятость в данных. Качество данных открытых источников. Влияние качества данных на принятие бизнес-решений. Управление качеством данных.

Тема 4. Разметка данных

Краткое содержание темы. Разметка данных. Значимость разметки данных в процессе аналитического решения. Организация процесса разметки данных. Основные принципы разметки данных. Виды разметки данных в зависимости от задачи и типа данных. Инструменты разметки данных с учетом их специфики. Инструкции по разметке данных. Обеспечение надежности разметки данных. Альтернативы разметки данных: слабое обучение и активное обучение.

Раздел 2: Сбор данных с использованием API

Тема 1. Знакомство с понятием API

Краткое содержание темы. API. Web API. REST API и SOAP. Как REST API и SOAP связаны с HTTP протоколом. Какие форматы поддерживают REST и SOAP API.

Тема 2. Краткий обзор HTTP-протокола

Краткое содержание темы. HTTP-протокол. Задача HTTP-протокола. Участники обмена сообщениями. Клиент-серверное общение. Структура запросов и ответов. Методы запросов.

Тема 3. API социальной сети ВКонтакте

Краткое содержание темы. Где найти документацию к API ВКонтакте. OAuth 2.0. Access token. Как зарегистрировать свое приложение в ВКонтакте. Как получить токен доступа API ВКонтакте.

Тема 4. Библиотека Requests

Краткое содержание темы. Выполнение запроса. Получение кода состояния ответа. Обработка ошибок запросов. Передача и просмотр заголовков сообщений. Получение содержимого ответа.

Тема 5. Использование документации API ВКонтакте

Краткое содержание темы. Поиск методов в документации. Определение метода. Выполнение запроса из документации к API.

Тема 6. Сбор данных с помощью Requests

Краткое содержание темы. Выполнение запроса к API с помощью requests. Преобразование объекта Python в объект JSON.

Тема 7. Библиотека vk_api

Краткое содержание темы. Методы в vk_api для работы с API ВКонтакте. Сессия для доступа к API.

Тема 8. Запись данных в CSV формат

Краткое содержание темы. Формат файла CSV. Создание файла с CSV форматом. Запись данных с помощью Python.

Раздел 3: Парсинг данных с сайтов

Тема 1. Знакомство с парсингом

Краткое содержание темы. Парсинг, программа-парсер, краулинг, сложности при парсинге сайта.

Тема 2. Законность парсинга

Краткое содержание темы. Ограничения парсинга, прописанные в законодательстве РФ. Ограничения сети-интернет, прописанные в законодательстве РФ.

Тема 3. Панель инструментов разработчика.

Краткое содержание темы. Панель инструментов разработчика. Инструменты. XHR запросы.

Тема 4. Библиотека BeautifulSoup

Краткое содержание темы. Инструмент для парсинга XML и HTML. Создание сессии с помощью requests, получение HTML кода. Преобразование HTML в дерево объектов BeautifulSoup. lxml. Обработка специальных кодов HTML. Сбор данных со всех страниц.

Тема 5. Библиотека BeautifulSoup: Работа с селекторами и пагинацией на сайте

Краткое содержание темы. Сбор информации с помощью селекторов. Сбор данных со всех страниц сайта.

Тема 6. Библиотека Selenium: Драйверы и особенности работы в Google Colab

Краткое содержание темы. Использование нескольких инструментов парсинга. Установка драйверов браузера для имитации работы браузера. Как эмулировать работу браузера, получать элементы исходного кода с помощью Selenium.

Тема 7. Библиотека Selenium

Краткое содержание темы. Взаимодействие с элементами интерфейса. Использование селекторов в selenium.

Тема 8. Использование Selenium с BeautifulSoup.

Краткое содержание темы. Обработка полученных данных с помощью BS4. Сбор данных в CSV файл.

Тема 9. Фреймворк Scrapy

Краткое содержание темы. Создание проекта Scrapy. Использование интерпретатора командной строки Scrapy. Как задавать правила сбора данных, выполнять поиск по исходному коду с помощью Scrapy, запускать своего “паука”. Сравнение BS4, Selenium и Scrapy: Гибкость, Производительность и Экосистема.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости синхронных занятий, проведения тестов по пройденному материалу, выполнения практических домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

Оценочные материалы текущего контроля размещены на сайте ТГУ в разделе «Информация об образовательной программе» – https://www.tsu.ru/sveden/education/eduop/.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Для получения зачета в первом семестре необходимо выполнение итоговой практической работы по Разделу 2: Сбор данных с использованием API.

Пример итогового задания:

С помощью методов API ВКонтакте получите 1000 подписчиков группы "Лентач", отсортированных по дате регистрации.

Необходимо собрать следующие данные в CSV файл: пол, название город, семейное положение (ФИО партнера не указывать).

Оценка осуществляется по пятибалльной системе. Для получения **зачета** необходимо получить оценку не менее 3 баллов.

Для получения **зачета с оценкой во втором семестре** необходимо выполнение итоговой практической работы (по Разделу 1: Поиск аналитических данных и Разделу 3: Парсинг данных с сайтов).

Пример итоговой работы:

Задание 1

Напишите код, который выполнит ввод слова "Lenovo" в поисковую строку сайта 1000kem.ru (https://1000kem.ru/contacts/) и выполните поиск.

Задание 2

Соберите информацию с сайта nbcomputers.ru (https://www.nbcomputers.ru/catalog/noutbuki/) о ноутбуках данного интернет-магазина.
Данные, которые необходимы:

Название ноутбука

Цена ноутбука

Код товара

Результат необходимо записать в CSV файл.

Задание 3

Оцените качество собранных в результате выполнения **задания 2** данных.

Результаты **зачета с оценкой** определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценочные материалы для проведения промежуточной аттестации размещены на сайте ТГУ в разделе «Информация об образовательной программе» – https://www.tsu.ru/sveden/education/eduop/.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в LMS «Data-Diving».

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине (https://www.tsu.ru/sveden/education/eduop/).

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

1. Просто о больших данных : пер. с англ. / Джудит Гурвиц, Алан Ньюджент, Ферн Халпер, Марсия Кауфман ; Сбербанк. - Москва : Эксмо, 2015. - 393 с. - (Библиотека Сбербанка; т. 58:) . URL: http://sun.tsu.ru/limit/2016/000553356/000553356.pdf
2. Еременко К. Работа с данными в любой сфере: как выйти на новый уровень, используя аналитику : Практическое пособие. - Москва : ООО "Альпина Паблишер", 2019. - 303 с.. URL: https://znanium.com/catalog/document?id=352376. URL: https://znanium.com/cover/1078/1078503.jpg
3. Макшанов А. В. Большие данные. Big Data : учебник для вузов / Макшанов А. В.,Журавлев А. Е.,Тындыкарь Л. Н.. - Санкт-Петербург : Лань, 2022. - 188 с.. URL: https://e.lanbook.com/book/198599. URL: https://e.lanbook.com/img/cover/book/198599.jpg
4. Замятин А. В. Интеллектуальный анализ данных : учебное пособие : [для студентов университетов и вузов] / А. В. Замятин ; Нац. исслед. Том. гос. ун-т. - Томск : Издательский Дом Томского государственного университета, 2020. - 193 с.: ил., табл.. URL: http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000722107
5. Dimitris Kouzis – Loukas. Learning Scrapy. – Packt Publishing, 2016. – 270 с.
6. Митчелл Райан. Скрапинг веб-сайтов с помощью Python. – ДМК Пресс, 2016. – 280 с.
7. Gábor László Hajba. Website Scraping with Python. Using BeautifulSoup and Scrapy. – apress, 2018. – 244 с.
8. Heydt Michael. Python Web Scraping Cookbook. – Packt Publishing, 2018. – 339 с.

б) дополнительная литература:

1. – Дятлов, А.В. Анализ данных в социологии : учебник / А.В.Дятлов, Д.А.Гугуева ; Южный федеральный университет. - Ростов-на-Дону ; Таганрог : Издательство Южного федерального университета, 2018. - 226 с. - ISBN 978-5-9275-2690-1. - Текст : электронный. - URL: https://znanium.com/catalog/product/1039664
2. – Юре, Л. Анализ больших наборов данных / Л. Юре, Р. Ананд, Д. У. Джеффри ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2016. — 498 с. — ISBN 978-5-97060-190-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/93571
3. Зыков Р. В. Роман с Data Science : как монетизировать большие данные. - Санкт-Петербург [и др.] : Питер, 2022. - 318 с.
4. Аналитическая культура. От сбора данных до бизнес-результатов / Карл Андерсон; пер. с англ. Юлии Константиновой; [науч. ред. Руслан Салахиев]. — М.: Манн, Иванов и Фербер, 2017.
5. Миркин Б. Г. Введение в анализ данных : учебник и практикум для бакалавриата и магистратуры : [для студентов вузов, обучающихся по инженерно-техническим, естественно-научным и экономическим направлениям и специальностям] / Б. Г. Миркин ; "Высшая школа экономики" Национальный исследовательский университет. - Москва : Юрайт, 2015. - 173, [1] с.: ил. - (Авторский учебник)

в) ресурсы сети Интернет:

1. Физтех. Статистика. Введение в анализ данных, 2023. Введение в анализ данных: Сбор данных из открытых источников
URL: https://miptstats.github.io/courses/ad_fivt/data_parsing.html
2. Современный веб-скрапинг с BeautifulSoup и Selenium
URL:https://code.tutsplus.com/ru/tutorials/modern-web-scraping-with-beautifulsoup-and-selenium--cms-30486
3. Начало работы с Вебом
URL: https://developer.mozilla.org/ru/docs/Learn/Getting_started_with_the_web
4. Изучение веб-разработки URL: https://developer.mozilla.org/ru/docs/Learn

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

– ***Облачная среда для работы с кодом:***

Google Colab (ссылка на офиц.инструкцию)

Требования: наличие браузера, подключения к сети Интернет, учетной записи google

ИЛИ

– ***Стационарная среда для работы с кодом:***

Jupyter Lab в составе Anaconda (ссылка на скачивание среды Anaconda, инструкция по установке на Windows с официального сайта (на англ.яз., там же: инструкция для macOS и Linux), перевод инструкции на рус.яз.)

Требования: наличие браузера, подключение к сети Интернет

ОС: ***Windows*** не ниже 8 64-bit x86; ***macOS*** 10.13+ 64-bit x86 & M1; ***Linux***, включая Ubuntu, RedHat, CentOS 7+ 64-bit x86, 64-bit aarch64 (AWS Graviton2), 64-bit Power8/Power9, s390x (Linux on IBM Z & LinuxONE). (Для более старых версий операционных систем можно найти соответствующие установщики в архиве).

Место на диске: не менее 5 Гб для скачивания и установки (+ объем для установки всех необходимых библиотек)

Занимаемая оперативная память: от 3 Гб

ИЛИ

Стационарная программа для работы с кодом, используемая в лекциях:
Visual Studio Code (ссылка на скачивание, инструкция по установке на Windows с офиц сайта (на англ.яз., там же в оглавлении можно найти для macOS и Linux), инструкция по установке Python и расширений для использования в VSCode с офиц сайта (на англ.яз.)

Требования: наличие браузера, Интернета, установленного Python 3

ОС: ***Windows*** не ниже 8.0, 8.1 для 32-bit или 10, 11 для 64-bit; ***macOS*** 10.13+; ***Linux*** (Debian): Ubuntu Desktop 16.04, Debian 9 или Linux (Red Hat): Red Hat Enterprise Linux 7, CentOS 7, Fedora 34

Место на диске: не менее 500 Mb для скачивания и установки

Занимаемая оперативная память: от 1 Гб

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index
– ЭБС Лань – http://e.lanbook.com/
– ЭБС Консультант студента – http://www.studentlibrary.ru/
– Образовательная платформа Юрайт – https://urait.ru/
– ЭБС ZNANIUM.com – https://znanium.com/
– ЭБС IPRbooks – http://www.iprbookshop.ru/

14. Материально-техническое обеспечение

Занятия по учебной дисциплине проводятся с использованием дистанционных образовательных технологий. Каждый обучающийся обеспечен доступом к образовательной платформе https://edu.data-diving.ru/.

15. Информация о разработчиках

Басина Полина Александровна, аналитик, научно-исследовательская лаборатория прикладного анализа больших данных НИ ТГУ;

Абазовская Анастасия Александровна, разработчик, ООО «Академия Дата-дайвинг»; лаборант, научно-исследовательская лаборатория прикладного анализа больших данных НИ ТГУ