МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра материаловедения, технологии
и управления качеством

ПРИМЕНЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ПРИ СБОРЕ И АНАЛИЗЕ ДАННЫХ ДЛЯ КОНТРОЛЯ КАЧЕСТВА НА ПРИМЕРЕ САЙТА ОБРАЗОВАТЕЛЬНОЙ ОРГАНИЗАЦИИ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента магистратуры 3 курса 3201 группы
направления 27.04.02 «Управление качеством»
профиль «Менеджмент качества в инженерной и образовательной деятельности»
института физики

Каштанова Вячеслава Алексеевича

Научный руководитель, доцент, к.ф.-м.н.		А. В. Козловский
должность, уч. степень, уч. звание	подпись, дата	инициалы, фамилия
Зав. кафедрой, д.ф.-м.н., профессор		С.Б. Вениг
должность, уч. степень, уч. звание	подпись, дата	инициалы, фамилия

Саратов 2024

Введение. В современных реалиях у каждой организации есть возможность создать свой собственный сайт, на котором она стремится рассказать о себе, своей продукции и достижениях. При этом при возрастающей роли информатизации во всех процессах, веб-сайты стали выходить за рамки «визитной карточки», обретая более важную роль в информационном пространстве организации. Так, например, сеть пиццерий «Додо Пицца» размещает на своих информационных ресурсах всю информацию, как об организации рабочих процессов, так и всю аналитику.

К сайтам государственных организаций предъявляются определенные требования, упрощающие их мониторинг со стороны надзорных органов. При этом требования к предоставляемой информации растут от года к году, как со стороны пользователей, так и со стороны государства. Для образовательных учреждений существует ряд документов, регулирующих содержание сайтов и работу по их обновлению. Для проверки качества информации вручную сотрудникам требуется очень много времени, особенно если проверяемый массив данных содержит десятки, сотни тысяч элементов. В связи с этим растет актуальность автоматизации процесса проверки информации, содержащейся на сайте, на соответствие всем требованиям нормативной документации. Автоматизировать сбор и анализ данных, размещаемых на сайтах организаций, в том числе образовательных, можно с помощью языков программирования, таких как Python.

Таким образом, была определена цель выпускной квалификационной работы: изучение возможностей языков программирования для автоматизированного сбора данных сайта образовательной организации, а также проведение исследовательского анализа этих данных. Для достижения цели необходимо решить следующие задачи.

1. Изучить требования рособрнадзора и образовательных стандартов к сайту образовательной организации.
2. Освоить основы языка программирования Python.

3. Провести обзор литературы на тему «Парсинг веб-сайтов с помощью Python».

4. Написать программу для сбора данных из раздела «Сведения об образовательной организации» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».

5. Провести предобработку и исследовательский анализ данных из раздела «Сведения об образовательной организации» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».

6. Написать программу для сбора данных из раздела «Ранжированные списки поступающих» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».

7. Провести предобработку и исследовательский анализ данных из раздела «Ранжированные списки поступающих» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».

8. Сделать выводы об эффективности автоматизированного сбора данных сайтов образовательных организаций.

Выпускная квалификационная работа занимает 53 страницы, имеет 35 рисунков и 2 приложения.

Обзор составлен по 20 информационным источникам.

Во введение рассматривается актуальность работы, устанавливается цель и выдвигаются задачи для достижения поставленной цели.

Первый раздел представляет собой теоретическую основу, необходимую для выполнения задач.

Во втором разделе работы представлена практическая часть с кодом программ, написание которых необходимо для достижения цели работы, а так же аналитика полученных данных.

Основное содержание работы

В теоретической части работы сформулированы основные требования нормативной документации к сайтам образовательных организаций. Продемонстрированы основные преимущества языка программирования

Python. Разобрана структура сайта СГУ, в частности раздел «Сведения об образовательной организации. Руководство. Педагогический (научно-педагогический) состав».

Согласно Приказу Роскомнадзора от 14.08.2020 № 831 «Об утверждении Требований к структуре официального сайта образовательной организации в информационно-телекоммуникационной сети «Интернет» и формату представления информации», требования приказа определяют структуру официального сайта образовательной организации в информационно-телекоммуникационной сети Интернет, а также формат представления образовательной организацией информации, обязательной к размещению на Сайте образовательной организации. В пункте 3.6 содержатся требования к подразделу «Руководство. Педагогический (научно-педагогический) состав».

Для выполнения данной работы применяется язык программирования Python.

Основными преимуществами Python являются.

- Скорость выполнения программ написанных на Python очень высока. Это связанно с тем, что основные библиотеки Python написаны на C++ и выполнение задач занимает меньше времени, чем на других языках высокого уровня.
- В связи с этим можно писать свои собственные модули для Python на C или C++.
- В стандартных библиотеках Python вы можете найти средства для работы с электронной почтой, протоколами Интернета, FTP, HTTP, базами данных, и пр.
- Скрипты, написанные при помощи Python выполняются на большинстве современных ОС. Такая переносимость обеспечивает Python применение в самых различных областях.
- Python подходит для любых решений в области программирования, будь то офисные программы, вэб-приложения, GUI-приложения и т.д.

- Над разработкой Python трудились тысячи энтузиастов со всего мира.

Для выполнения задач разобрана структура сайта СГУ. Определены разделы для парсинга информации с сайта.

Определены библиотеки для написания рабочего кода. Requests – это модуль Python, который вы можете использовать для отправки всех видов HTTP-запросов. Pandas предоставляет богатый набор функций для DataFrame. Например, выравнивание данных, статистика данных, нарезка, группировка, объединение данных и т.д. NumPy – это низкоуровневая структура данных, которая поддерживает многомерные массивы и широкий спектр математических операций с массивами. Pandas имеет интерфейс более высокого уровня. Он также обеспечивает оптимизированное согласование табличных данных и мощную функциональность временных рядов. Seaborn – библиотека для создания статистических графиков на Python. Она построена на основе matplotlib и тесно интегрируется со структурами данных pandas. Matplotlib – библиотека на языке программирования Python для визуализации данных двумерной и трёхмерной графикой. Получаемые изображения могут быть использованы в качестве иллюстраций.

Далее последовательно разобран процесс получения HTML-кода страницы сайта, сформирован DataFrame с необходимыми для аналитики данными.

На основании полученных таблиц приведена аналитика с построением графиков для общего стажа и стажа работы по специальности для всего преподавательского состава. Согласно данным в размещенной на сайте информации у некоторых преподавателей стаж работы по специальности выше, чем общий стаж, что говорит об ошибке при наполнении сайта. На диаграмме разброса наглядно показаны эти ошибки.

Написана Программа для сбора данных из раздела «Ранжированные списки поступающих» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского». Для выполнения этой задачи была применена библиотека

Selenium. Selenium – это не одна единственная библиотека, а целый комплекс библиотек и инструментов, совместимых не только с языком Python, но и с Java и JavaScript, C Sharp, Ruby, а также с Kotlin. Тем не менее, будет рассмотрен только вариант на Python. После формирования всей необходимой информации была выполнена предобработка и исследовательский анализ данных из раздела «Ранжированные списки поступающих» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского». Составлена сводная таблица для всех структурных подразделений университета для копий и оригиналов документов, поданных по результатам приемной кампании 2023 года. Представлена визуализация в виде графиков.

Приведены графики количества абитуриентов по дням для каждого подразделения, демонстрирующие динамику заявок.

Заключение. В рамках выполнения данной работы.

1. Изучены требования рособрнадзора и образовательных стандартов к сайту образовательной организации.
2. Освоены основы языка программирования Python.
3. Проведен обзор литературы на тему «Парсинг веб-сайтов с помощью Python».
4. Написана программа для сбора данных из раздела «Сведения об образовательной организации» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».
5. Выполнены предобработка и исследовательский анализ данных из раздела «Сведения об образовательной организации» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».
6. Написана программа для сбора данных из раздела «Ранжированные списки поступающих» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».
7. Проведены предобработка и исследовательский анализ данных из раздела «Ранжированные списки поступающих» веб-сайта ФГБОУ ВО «СГУ имени Н.Г. Чернышевского».

8. Сделаны выводы об эффективности автоматизированного сбора данных сайтов образовательных организаций.

Для уменьшения ошибок при заполнении данных, повышения контроля и упрощения сбора данных необходимо:

- интегрировать сайт СГУ с другими информационными ресурсами университета, такими как IpsilonUni, ЕИП СГУ, или порталы «Личный кабинет сотрудника» и «Портал самообслуживания»;
- руководителям структурных подразделений сообщить о работе и пользе этих порталов.

Список использованных источников

1. Парсинг сайтов: что это, для чего нужен, классификация парсеров [Электронный ресурс] // Блог Яндекс Практикума [Электронный ресурс] : [сайт]. – URL: https://practicum.yandex.ru/blog/chto-takoe-parsing-sayta (дата обращения: 22.12.2023). – Загл. с экрана – Яз. Рус.

2. Парсинг: что это такое и как его применять [Электронный ресурс] // Skillbox Media [Электронный ресурс] : [сайт]. – URL: https://skillbox.ru/media/marketing/chto-takoe-parsing-i-chto-o-nyem-obyazatelno-nuzhno-znat-marketologu/ (дата обращения: 22.12.2023). – Загл. с экрана – Яз. Рус.

3. Приказ Федеральной службы по надзору в сфере образования и науки №831 от 14.08.2020 «Об утверждении Требований к структуре официального сайта образовательной организации в информационно-телекоммуникационной сети «Интернет» и формату представления информации» (ред. от 12.01.2022) [Электронный ресурс] // Электронный фонд правовых и нормативно-технических документов [Электронный ресурс] : [сайт]. – URL: https://docs.cntd.ru/document/565780511 (дата обращения: 02.12.2023). – Загл. с экрана. – Яз. Рус.

4. Федеральный закон №273-ФЗ от 29.12.2012 «Об образовании» // Собрание законодательства Российской Федерации. – 2012. – № 53, (ч. I). – ст. 7598.

5. Парсинг сайтов: как с точки зрения закона выглядит один из самых полезных ИТ- инструментов по миру (и в России)? [Электронный ресурс] // Хабр [Электронный ресурс] : [сайт]. – URL: https://habr.com/ru/post/340302/ (дата обращения: 22.12.2023). – Загл. с экрана. – Яз. Рус.

6. Самоучитель Python [Электронный ресурс] // Python 3 для начинающих [Электронный ресурс] : [сайт]. – URL: https://pythonworld.ru/samouchitel-python (дата обращения: 22.12.2023). – Загл. с экрана. – Яз. Рус.

7. Парсинг-руководство для новичков [Электронный ресурс] // Parsing Cloud [Электронный ресурс] : [сайт]. – URL: https://parsing-cloud.ru/articles/parsing-rukovodstvo-dlya-nachinauschih (дата обращения: 22.12.2023). – Загл. с экрана. – Яз. рус.

8. Мэтиз, Э. Программирование игр, визуализация данных, веб-приложений / Э. Мэтиз. – СПб. : Издательский дом «Питер», 2020. – 295 с.

9. Доусон, М. Программируем на Python / М. Доусон. – СПб. : Издательский дом «Питер», 2020. – 380 с.

10. Поляков, К. Ю. Программирование. Часть 2 / К. Ю. Поляков. – М. : Издательство «Просвещение», 2023. – 176 с

11. Сайт ФГБОУ ВО «СГУ имени Н.Г. Чернышевского» [Электронный ресурс] // sgu.ru [Электронный ресурс] : [сайт]. – URL: https://www.sgu.ru/ (дата обращения: 22.12.2023). – Загл. с экрана – Яз. Рус.

12. Requests [Электронный ресурс] // Python Package Index - PyPI [Электронный ресурс] : [сайт]. – URL: https://pypi.org/project/requests/ (дата обращения: 22.12.2023). – Загл. с экрана – Яз. Рус

13. Лутц, М. Изучаем Python/ М. Лутц. – СПб. : Издательский дом «Питер», 2019. – 252 с.

14. Pandas – Python Data Analysis Library [Электронный ресурс] // Pandas [Электронный ресурс] : [сайт]. – URL: https://pandas.pydata.org/ (дата обращения: 22.12.2023). – Загл. с экрана – Яз. рус

15. Seaborn: Statistical Data Visualization [Электронный ресурс] // Seaborn [Электронный ресурс] : [сайт]. – URL: https://seaborn.pydata.org/ (дата обращения: 22.12.2023). – Загл. с экрана – Яз. рус

16. Matplotlib: Visualization with Python [Электронный ресурс] // Matplotlib [Электронный ресурс] : [сайт]. – URL: https://matplotlib.org/ (дата обращения: 22.12.2023). – Загл. с экрана – Яз. рус

17. Beautiful Soup Documentation [Электронный ресурс] // Beautiful Soup [Электронный ресурс] : [сайт]. – URL: https://beautiful-soup-4.readthedocs.io/en/latest/# (дата обращения: 22.12.2023). – Загл. с экрана – Яз. Рус

18. The Selenium Browser Automation Project [Электронный ресурс] // Selenium [Электронный ресурс] : [сайт]. – URL:https://www.selenium.dev/documentation/ (дата обращения: 22.12.2023). – Загл. с экрана – Яз. Рус.

19. Применение парсинга [Электронный ресурс] // vc.ru [Электронный ресурс] : [сайт]. – URL: https://vc.ru/u/554100-dmitriy-eliseev/236548-10 (дата обращения: 22.12.2023). – Загл. с экрана – Яз. Рус.

20. Максим Кульгин (makasin4ik) [Электронный ресурс]// Habr [Электронный ресурс]: [блог]. – URL: https://habr.com/ru/post/340302/ (дата обращения: 22.12.2023). –Загл. с экрана. – Яз. Рус.