МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Погорєлов Олексій Валентинович

УДК 004.4

СТВОРЕННЯ ІНСТРУМЕНТУ ДЛЯ АВТОМАТИЗОВАНОГО ПАРСИНГУ САЙТІВ НА ОСНОВІ ТЕХНОЛОГІЇ NODE.JS

Галузь знань 12 «Інформаційні технології» за спеціальністю
122 «Комп’ютерні науки»
122 – БКР.А – 401з. 217401з04

Автореферат
бакалаврської кваліфікаційної роботи на здобуття освітньої кваліфікації
«бакалавр комп’ютерних наук»

Миколаїв – 2021

Бакалаврська кваліфікаційна робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: викладач кафедри інтелектуальних інформаційних систем **Таранов Микита Олександрович**

Рецензент: канд. техн. наук, старший викладач кафедри інженерії програмного забезпечення **Дворецький Михайло Леонідович,**

Захист відбудеться «24» червня 2021 р. о $9^{\underline{00}}$ год. на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

З бакалаврською кваліфікаційною роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «16» червня 2021 р.

Секретар
екзаменаційної комісії,
викладач кафедри ІІС М. О. Таранов

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Сучасні темпи розвитку інтернет-індустрії просто неймовірні. Все більше і більше користувачів переходить на інтернет-платформи, створює онлайн магазини, онлайн блоги, каталоги, портфоліо, форуми. Всі ці ресурси переповнені інформацією, яка є дуже коштовним ресурсом для багатьох компаній, котрі й займаються її збором, класифікацією та обробкою.

Однак збір цієї інформації надзвичайно складний і трудомісткий процес. Один зі способів її збирання звісно вручну. Але на жаль так необхідну кількість інформації зібрати неможливо, тому що навіть обробка невеликого об’єму займає безліч часу.

Саме з ціллю вирішення подібних задач, прийнято рішення з розробки власного інструменту, а саме: «Створення інструменту для автоматизованого парсингу сайтів на основі технології node.js»

Актуальність теми обумовлена тим, що зростає потреба у вивченні, аналізі, оптимізації та створенні технологій для ефективного збору та аналізу даних з вебресурсів.

Метою бакалаврської кваліфікаційної роботи є розробка зручного та функціонального чат-боту для автоматизованого збору даних про публікації та користувачів у соціальній мережі Reddit на основі модуля для парсингу сайтів Nightmare.

Практичне призначення полягає у створенні унікального застосунку для збору даних з унікальною архітектурою за всіма специфікаціями.

Пояснювальна записка до бакалаврської кваліфікаційної роботи складається із вступу, вступу, 3 розділів, висновків. Загальний обсяг роботи складає: 78 сторінок, 26 рисунків, 7 таблиць та 21 посилань на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі кваліфікаційної бакалаврської роботи описано проблему, обґрунтовано актуальність обраної теми, сформовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У першому розділі проведений докладний опис предметної сфери, а саме, що таке автоматизований парсинг сайтів, його популярність на сьогоднішній день. Розглянуто подібні технології та визначено чому саме використання парсингу є кращим рішенням. Також розглянуто принципи роботи, функції та сфера використання парсингу. А саме використовується у сфері програмування (читання URL-адрес, читання HTML-коду, аналіз XML-розмітки і так далі) та сфері маркетингу (аналітика, порівняння цін, збір контенту, вибірки даних). Також розглянуто аналоги і приклади сервісів, що використовуються.

У другому розділі розглянуто та обрано інструменти для створення системи.

Інструмент для програмування обрано платформу Node.js. Розглянуто можливості її використання, функціонал переваги та недоліки.

Також проведено огляд та аналіз бібліотек для автоматизованого парсингу, для платформи Node.js. Розглянуто такі бібліотеки, як Simplecrawler, Cheerio, Osmosis, Puppeteer, Apify SDK, Nightmare. Серед них обрано бібліотеку Nightmare, за її функціонал, легкість та просто у роботі.

Для збору і зберігання даних обрано базу даних. Проведено порівняльний аналіз SQL та NoSQL баз даних. Серед них обрано нереляційну базу даних MongoDB.

В результаті для створенням інформаційної системи використано такі програмні засоби як: модуль Nightmare, для платформи програмування Node.js, базу даних для збереження та обробки даних обрано нереляційну MongoDB, інтерфейс для взаємодії з користувачем – чат-бот у месенджері Telegram — за допомогою бібліотеки node-telegram-bot-api.

У третьому розділі описано послідовно виконання всіх етапів розробки продукту за специфікацією, що вказана у першому розділі дипломної роботи.

Для розробки було пройдено етапи зі створення: парсера для збирання даних, чат-бота для взаємодії з парсером, розгортання інформаційної системи на локальному сервері, інтерфейс роботи з базою даних.

Початок розробки програмного продукту повинен починатися з планування та створення його архітектури. Важливо подбати про всі ймовірні випадки в його використанні, щоб система не зламалась, а також подбати про майбутнє масштабування системи.

У першому випадку користувач вводить запит, по якому буде відбуватися пошук даних у Reddit. Результатом його роботи буде записування всіх знайдених і оброблених публікацій з Reddit до бази даних. Далі користувач матиме отримати повідомлення та ознайомитись з першими п'ятьма результатами. Також в повідомленні надаватимуться навігаційні кнопки для відкриття детальної інформації до кожної з публікацій і кнопками для перегляду наступних або попередніх п’ятьох результатів.

Робота іншого сценарію можлива в тому випадку, коли користувач робив запити раніше, і вже має доступ до повідомлення з навігаційними кнопками.

Реалізація парсера. Наступним кроком при створенні парсеру, слід визначити схему даних. Дана схема необхідна для подальшої роботи з інформаційною системою. Адже вона містить інформацію про дані, що слід вилучити зі сторінки.

Під час створення парсеру головною задачею є обхід сторінок типової структури із постійним вилученням подібних даних.

З її допомогою можливо швидко зібрати дані з веб-ресурсу, що мають визначену та не складну структуру.

Реалізація чат-бота. Із інтерфейсом Телеграму, бот буде виглядати наступним чином. Зліва стандартна панель із переліком активних чатів

Телеграму, справа більша область чату з ботом, повідомлення користувача та бота виглядатимуть стандартно для всіх діалогів у даному месенджері - як закруглені прямокутники із невеликою стрілкою в куті. Повідомлення, що відправляє бот у відповідь на запит виглядатимуть певним чином: саме повідомлення міститиме нумерований список з п’яти заголовків статей, текст яких обрізаний до кінця рядка, під текстом розташована клавіатура з можливістю обрати та переглянути кожну публікацію зі списку детальніше та дві кнопки для перемотування списку результатів назад та вперед.

Чат-бот буде відповідає таким критеріям:

- під час його першого запуску, користувачеві надсилається інструкція з використання, котра описує можливості чат-бота та вказання щодо створення запитів;
- кожне наступне повідомлення користувача буде зчитане уже як запит з пошуку та буде надсилатися до сервера, де розташований парсер;
- виконавши роботу з пошуку даних, зберігатиме їх в базі даних;
- бот створюватиме вибірки з отриманих результатів та надсилатиме їх назад користувачеві. Разом з результатами буде надсилатися інтерактивна навігація у вигляді кнопок для перегляду результатів пошуку по сторінках, та кнопка збереження результатів пошуку у вигляді файлу на поточний пристрій;
- також бот матиме в наявності панель з обмеженим доступом спеціально для адміністратора. Дана панель надаватиме доступ до активних задач котрі здійснюють пошук, детальну інформацію про них (час запуску, текст запиту, тривалість роботи тощо). За допомогою інтерактивних кнопок адміністратору надаватиметься можливість керувати павуками, наприклад якщо його робота триватиме надто довго і можливі зібрані дані займатимуть багато місця.

У розділі з охорони праці були викладені вимоги щодо створення безпечних і здорових умов праці на робочому місці або у виробничому приміщенні та забезпечення безпеки людини у надзвичайних ситуаціях.

В ході опрацювання спеціальної частини з охорони праці визначено, які вимоги повинні бути встановлені на підприємстві для безпечної роботи робітників.

Серед опрацьованих матеріалів вирішено питання, щодо умов праці з урахуванням мікроклімату та оснащення робочого місця, техніки безпеки при експлуатації різної оргтехніки, приладів та технологій, забезпечення електробезпеки. Також розібрано вимоги гігієни праці та виробничої санітарії. Останнім розглянутим питанням став такий важливий критерій, як пожежна безпека для комп’ютерних приміщень. Проведено розрахунки освітленості для приміщення, в якому проходила робота над створенням програмного продукту.

ЗАГАЛЬНІ ВИСНОВКИ

Дана кваліфікаційна робота була направлена на розробку системи автоматизованого парсингу сайтів, з можливістю аналізу та обробки зібраних даних, зі зручним, простим і зрозумілим інтерфейсом. Робота даної системи відображена на взаємодії із соціальною мережею Reddit.

Проблема збору та обробки великої кількості інформації є доволі комплексною та складною задачею, з якою довелося зіткнутися в даній роботі. Над її вирішення працює велика кількість світових компаній, шляхом розроблення і вдосконалення власних методів та сервісів збору та обробки даних з вебресурсів і не тільки.

Шляхом ознайомлення з різними сервісами та методами їх роботи у сфері парсингу, розв’язано задачі, щодо власної системи, а саме: визначено засоби для реалізації, створено програму автоматизованого збору даних з інтернет-ресурсу Reddit, організовано реалізацію роботи з базою даних, реалізовано створення інтерфейс для роботи з інструментом парсингу.

Після проведеного аналізу здобутої інформації сформовано стек технологій та специфікації майбутнього програмного продукту.

Для роботи над створенням інформаційної системи використано такі програмні засоби як: модуль Nightmare, для платформи програмування Node.js, базу даних для збереження та обробки даних обрано нереляційну MongoDB, інтерфейс для взаємодії з користувачем – чат-бот у месенджері Telegram — за допомогою бібліотеки node-telegram-bot-api.

Подальшими напрямками для роботи можуть бути питання, що стосуються:

- вдосконалення створеної системи (рефакторинг коду, оптимізація);
- розширення та додавання нових функцій до бота та до парсера;
- повернення користувачеві в інтерпретованому вигляді як повідомлення чи файл.

- додавання функціоналу для роботи з іншими базами даних, можливості різноформатного експорту результатів.

Розроблений програмний продукт показав прийнятні результати під час тестування.

В ході опрацювання спеціальної частини з охорони праці визначено, які вимоги повинні бути встановлені на підприємстві для безпечної роботи робітників.

Серед опрацьованих матеріалів вирішено питання, щодо умов праці з урахуванням мікроклімату та оснащення робочого місця, техніки безпеки при експлуатації різної оргтехніки, приладів та технологій, забезпечення електробезпеки. Також розібрано вимоги гігієни праці та виробничої санітарії. Останнім розглянутим питанням став такий важливий критерій, як пожежна безпека для комп’ютерних приміщень. Проведено розрахунки освітленості для приміщення, в якому проходила робота над створенням програмного продукту.

АНОТАЦІЯ

Погорєлова Олексія Валентиновича. Тема: «Створення інструменту для автоматизованого парсингу сайтів на основі технології node.js ». – На правах рукопису.

Бакалаврська кваліфікаційна робота на здобуття освітньої кваліфікації «бакалавр з комп'ютерних наук» в галузі знань 12 «Інформаційні технології» за спеціальністю 122 «Комп'ютерні науки».

Чорноморський національний університет імені Петра Могили, Миколаїв.

Об'єктом дослідження даної роботи є функціональна система парсингу сайтів з можливістю збереження та обробки зібраних даних.

Предметом дослідження є системи збору інформації у мережі інтернет та системи керування збереженням та аналізом зібраної інформації.

Метою даної роботи є дослідження наявних методів для збирання та зберігання великих об'ємів даних, пошук способів розробки програмного забезпечення для реалізації обраних методів. Результатом є створена система парсингу сайту Reddit, а також чат-бот для взаємодії системи з користувачем.

Методи дослідження ґрунтуються на сфері аналізу документів у форматі XML та методах збору, обробки та аналізу даних.

Дипломна робота має 3 розділи: аналіз предметної сфери, об'єкту та предмету дослідження. Постановка задачі: розв'язання задачі автоматизованого збору слабоструктурованих даних; створення парсера з використанням модуля Nightmare та чат-бота для месенджера телеграм.

В першому розділі проводиться огляд та аналіз сфери автоматизованого парсингу сайтів, пошук існуючих і функціональних аналогів, котрі дадуть можливість краще зрозуміти сферу парсинг, для розроблення власної системи.

В другому розділі проводиться підбір системного та технічного забезпечення, проводиться порівняльний аналіз існуючих варіантів для вибору кращого інструменту реалізації.

У третьому розділі описується процес розробки та створення системи парсингу та чат-бота, проводиться аналіз роботи системи.

Бакалаврська кваліфікаційна робота містить сторінок – 78, рисунків – 26, таблиць – 4, додатків – 3, джерел – 21.

Ключові слова: веб-скрапінг, пошук описів, чат-бот, соціальна мережа.

ABSTRACT

Pohorielov Oleksii. Topic: "Creating a tool for automated site parsing based on node.js technology". – On the rights of the manuscript.

Bachelor's qualification work for the educational qualification "Bachelor of Computer Science" in the field of knowledge 12 "Information Technology" in the specialty 122 "Computer Science".

Petro Mohyla Black Sea National University, Mykolaiv.

The object of study of this work is a functional system of parsing sites with the ability to store and process the collected data.

The subject of research of this thesis is the field of site parsing.

The purpose of this work is to study the existing methods for collecting and storing large amounts of data, finding ways to develop software for the implementation of selected methods. The result is a created parsing system for the Reddit site, as well as a chatbot for the system to interact with the user.

The research methods are based on the field of document analysis in XML format and methods of data collection, processing and analysis.

Thesis has 3 sections: analysis of the subject area, object and subject of research. Problem statement: solving the problem of automated collection of poorly structured data; creating a parser using the Nightmare module and a chat bot for telegram messenger.

The first section reviews and analyzes the field of automated parsing of sites, the search for existing and functional analogues that will give a better understanding of the field of parsing, to develop your own system.

The second section selects the system and hardware, comparative analysis of existing options for choosing the best implementation tool.

The third section describes the process of developing and creating a parsing system and chatbot, analyzes the system.

Keywords: web scraping, description search, chatbot, social network.