МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Кір'якіді Анна Володимирівна

УДК 004.4

ІНФОРМАЦІЙНА СИСТЕМА ОБРОБКИ СЛАБОСТРУКТУРОВАНИХ ДАНИХ НА ОСНОВІ SCRAPY

Галузь знань 12 «Інформаційні технології» за спеціальністю
122 «Комп'ютерні науки та інформаційні технології»
122 - ДР.А - 403.21610306

Автореферат
дипломної роботи на здобуття освітньої кваліфікації
«бакалавр комп'ютерних наук та інформаційних технологій»

Миколаїв – 2020

Дипломна робота є рукопис.

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

Науковий керівник: д-р техн. наук, професор кафедри Інтелектуальних інформаційних систем Гожий О. П.

Рецензент: к.ф.-м.н., доцент, доцент кафедри ІІС, секція прикладної та вищої математики Воробйова А. І.

Захист відбудеться «25» червня 2020 р. о $9^{\underline{30}}$ 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 дипломною роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «18» червня 2020 р.

Секретар
екзаменаційної комісії,
викл. О. С. Скакодуб

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

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

Метою дипломної роботи є розробка зручного та функціонального чат-боту для автоматизованого збору даних про твіти та корисутвачів у соціальній мережі Твіттер на основі фреймворку для парсингу сайтів Scrapy.

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

Структура дипломної роботи. Пояснювальна записка до дипломної роботи складається із вступу, 4 розділів, висновків, додатків. Загальний обсяг роботи складає 79 сторінок, 7 рисунків та 14 посилань на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

ВСТУП

В наш час у світі дуже важливе значення в інформаційних технологіях має аналіз даних. Багато компаній займаються збиранням, класифікацією та обробкою великих об’ємів даних. І найбільше сховище даних у світі, звісно, Інтернет. Але для того, зібрати з нього дані достатньо важко. Перший спосіб — збирати інформацію вручну. Очевидно, що багато інформації таким способом не збереш і навіть збір невеликого об’єму займатиме багато часу. Другий спосіб — використання API. Наприклад, інформацію з - StackOverflow - відомого форуму для програмістів, можна знайти або за допомогою пошуку, або з використанням API. Даний метод дуже зручний, але є значний недолік — далеко не кожен ресурс в Інтернеті має даний інструмент. Тому залишається найскладніший, але водночас і найнадійніший

метод збору слабоструктурованих даних — веб-скрапінг або, як ще його називають, парсинг.

Інструменти парсингу дуже стрімко розвиваються і удосконалюються. Вони використовуються в великій кількості сфер. Наприклад, збір даних для дослідження ринку. Вилученні данні допоможуть прослідкувати за тенденціями у тому направленні, до якого прагне той чи інший бізнес. Наступний приклад це — створення рішень для офлайнового використання і зберігання великих об’ємів даних. Таким чином можна уникнути залежності від з’єднання з Інтернетом оскільки дані будуть доступні незалежно від можливості підключатися до Інтернету. Іще одна сфера використання парсингу сайтів — пошук співробітників, такий інструмент може мати багато фільтрів, за якими пошук людей буде набагато ефективнішим, ніж простий ручний пошук. Для тих, хто активно користується послугами онлайн-шопінгу існують свої зручні інструменти, які відстежують ціни на товари, шукають один і той самий товар у різних інтернет-магазинах.

Об'єктом дослідження є інструменти парсингу для веб-ресурсів, що існують зараз, способи зберігання та аналізу великих об’ємів даних.

Метою даної роботи є створення власного рішення для збору даних на основі наявних інструментів. Об’єктом для парсингу був обраний Twitter - соціальна мережа, складена з великої кількості мікроблогів. Даний ресурс є зручним для дослідження і вивчення роботи парсерів, тому що формат Twitter передбачає невеликі дописи, розміром не більше 280 символів. Такий формат блогів буде зручним для аналізу твітів (дописів). Для зручної роботи з парсером було вирішено створити чат-бота на базі месенджеру Телеграм, з якого буде можливо керувати збором даних.

На основі мети даної дипломної роботи було визначено наступні задачі: обраним інструментальним засобом реалізувати програмно інструмент для автоматизованого збору даних з Twitter, реалізувати роботу з базою даних, реалізувати чат-бота для роботи зі створеним інструментом веб-парсингу.

Для розв'язання цих задач було обрано наступні технології: фреймворк Scrapy написаний мовою програмування Python для розробки павука, який збирає дані з веб-сайту, для бази даних було обрано MongoDB, а для створення чат-боту у месенджері Telegram — бібліотеку python-telegram-bot. Проєкт було створено за допомогою середи розробки PyCharm, також було використано програмне забезпечення для роботи з базою даних MongoDB Compass.

Перший розділ

Сфера парсингу є дуже популярною та існує багато компаній, які роблять збір даних на замовлення. Деякі надають лише технології та інструменти для розробки власних рішень, інші - пропонують провести повноцінний збір та аналіз даних з сайтів конкурента для певного замовника. Такі послуги користуються успіхом, в основному, для бізнесу. Вихідні дані, після отриманого аналізу, є дуже цінною інформацією для замовника послуг і можуть сказати про можливі варіанти поліпшення ведення бізнесу, недоліки та сильні сторони конкурентів у сфері, аналіз якої робила компанія з парсингу веб-сайтів.

Сканування соціальних мереж - далеко не новинка у сфері аналізу даних.

У світі криптовалюти існує комплексний індекс - Fear&Greed - відображає настрій учасників ринку. Знаходження цього коефіцієнту значною мірою відбувається шляхом аналізу соціальних мереж, зокрема, у Твіттері.

Ця платформа буде зручна для аналізу, бо повідомлення на ній адаптовані до машинного аналізу внаслідок існування хештегів. Розробники слідкують за кількістю дописів із хештегом “біткоїн”, що були опубліковані, та дивляться, як швидко на них реагують користувачі. Якщо твітів багато і на них реагують швидко, то це свідчить про жадібність - жагу людей купувати криптовалюту. В такий аналіз також враховується і зміст повідомлення, і тому “біткоїн незабаром рухне” говорить про страх, а не жадібність. Це явно демонструє живий приклад аналізу твітів.

Також у даному розділі описані основні технології, які використовуються у даній сфері, такі як: Beautiful Soup, Selenium, Html5lib, Nightmare, Puppeteer, Pomp та Scrapy. Було обрано основну базу даних для проєкту: MongoDB.

Згідно з темою дипломної роботи було вирішено створити інформаційну систему за певними вимогами. Система повинна мати функціонал для зручного збору, зберігання та видачі зібраної інформації. Вона складатиметься з трьох частин: чат-боту, парсера та бази даних. У розділі описано повну специфікацію програмного продукту, який було створено.

Другий розділ

Попри видиму легкість даної задачі, вона є досить складною і потребує детального вивчення та планування реалізації.

Після визначення основної специфікації продукту необхідно чітко визначити стек технологій для його створення. Для вибору технологій важливо обрати мову програмування, а вже після того буде можливим обрати інші технології.

У розділі описано мову програмування Python, її історію та зв’язок з іншими мовами.

Для виконання даної роботи було вирішено використовувати мову програмування Python. Дана мова є інтерпретованою об'єктно-орієнтованою мовою програмування високого рівня зі строгою динамічною типізацією. При створенні мови Гвідо ван Россум - автор Python - взяв для нової мови дуже багато із вже наявних тоді мов програмування. Серед великого списку мов, які вплинули на Python, можна зазначити такі: C та C++ (взято деякі синтаксичні структури), ABC (відступи для групування операторів), Lisp (риси функціонального програмування, операції map, reduce, lambda, filter тощо), Java (модулі logging, unittest, threading), Fortran (зрізи масивів та інше). Python підтримує динамічну типізацію та велику

кількість типів даних серед яких цікавими є цілі числа довільної довжини та комплексні числа.

Далі у розділі описано технології Twisted та Scrapy. Twisted - це фреймворк для асинхронної роботи з мережею, написаний на Python. Його було створено для інтернет-додатків та має в собі модулі для різних цілей. Twisted є дуже потужною системою для асинхронної роботи з мережею. Він полегшує реалізацію користувальницьких мережевих додатків та підтримує безліч поширених мережевих протоколів, включаючи SMTP, POP3, IMAP, SSHv2 і DNS.

Scrapy - фреймворк для обходу веб ресурсів із подальшим вилученням інформації певної структури. Його архітектура розроблена для створення легко масштабованих систем для веб сканування та перевірки поведінки сайтів. Далі у розділі детально описана архітектура роботи фреймворку та його основні частини.

Третій розділ

В даному розділі описано послідовно виконання всіх етапів розробки продукту за специфікацією, що вказана у першому розділі дипломної роботи. Для розробки парсеру було пройдено етапи зі створення: “павуку” для збирання даних, чат-боту для взаємодії з “павуком”, розгортка “павука” на локальному сервері, інтерфейс роботи з базою даних.

Розробка кожного програмного продукту повинна починатися з побудови його архітектури. Дуже важливо подбати про всі процеси та можливі сценарії для коректної роботи системи. У розділа наведена схема роботи інформаційної системи.

Наступною частиною створення парсеру є визначення схеми даних (клас Items) і вона представлена у розділі.

Основна задача при створенні даного парсера є обхід сторінок типової структури із постійним вилученням схожих даних. Для подібних потреб у фреймворку Scrapy є заготовлені класи павуків. Один із таких має назву

CrawlSpider, що в перекладі буде “павук, що повзає”. За його допомогою можна здійснювати парсинг веб-ресурсів, що мають визначену та нескладну структуру веб-сторінок. Саме подібний до цього механізм використано при створені даного продукту.

Для реалізації даного проєкту для взаємодії з користувачем було обрано створити чат-бота у месенджері Телеграм. Цей вибір було обґрунтовано великою популярністю месенджера та простотою створення чат-ботів на даній платформі. Було створено схематичний макет інтерфейсу. У розділі наведено поетапне створення чат-боту.

Окрім написання основного коду, треба розгорнути отриману систему на локальному сервері. Всі кроки описані послідовно у відповідному підрозділі.

У розділі з охорони праці описано основні правила техніки безпеки на виробництві та описано захист від іонізуючого випромінювання.

ЗАГАЛЬНІ ВИСНОВКИ

Дана робота присвячена аналізу сфери парсингу даних, обробці слабоструктурованих даних та сфери чат-ботів. Розв'язання проблем, пов’язаних з зі збиранням та обробкою великих об’ємів даних є дуже комплексною та складною задачею. Такою діяльністю займається багато компаній по всьому світу, тому дана тема є досить актуальною.

Під час роботи над дипломною роботою було вивчено багато методів та інструментів для парсингу сайтів, зберігання великої кількості даних, програмування чат-ботів. Після ознайомлення з теоретичним матеріалом було зроблено аналіз вивченого та сформовано стек технологій та специфікації майбутнього програмного продукту.

В результаті виконання даної дипломної роботи було створено інформаційну автоматизовану систему для пошуку та зберігання дописів у мережі Твіттер на

основі фреймворку Scrapy. Кінцевий продукт відповідає всім визначеним у роботі специфікаціям. Було створено дизайн програмного інтерфейсу та програмна реалізація.

АНОТАЦІЯ

Кір'якіді А. В. Інформаційна система обробки слабоструктурованих даних на основі Scrapy. – На правах рукопису.

Дипломна робота на здобуття освітньої кваліфікації «бакалавр комп'ютерних наук та інформаційних технологій» в галузі знань 12 «Інформаційні технології» за спеціальністю 122 «Комп'ютерні науки та інформаційні технології»

Чорноморський національний університет імені Петра Могили, Миколаїв

У світі в наш час способи розв'язання задачі збирання та обробки великих об'ємів даних мають великий попит та є досить популярними для вивчення та розвитку. Одним з об'єктів для масового збирання та аналізу інформації стала соціальна мережа Твіттер. Таким чином, розробка та застосування системи для зручного пошуку та збирання інформації з веб ресурсу Твіттер нині є актуальною.

Метою даної роботи є дослідження наявних методів для збирання та зберігання великих об'ємів слабоструктурованих даних, пошук способів розробки програмного забезпечення для реалізації обраних методів.

Результатами роботи є створення системи, яка містить в собі реалізацію обраного методу збору даних, сховище для зберігання зібраних даних та має зручний інтерфейс для роботи.

Предметом дослідження даної дипломної роботи є сфери парсингу сайтів із використанням фреймворку Scrapy та зберігання й роботи з великими об'ємами даних з використанням нереляційних баз даних.

Методи дослідження ґрунтуються на сфері аналізу документів у форматі XML та методах аналізу слабоструктурованих даних.

Досягнення поставленої мети вимагає розв'язання наступних задач:

1. Пошук, порівняння та вивчення сучасних методів веб скрапінгу та

додаткових технологій;

2. Визначення параметрів для створення інформаційної системи;
3. Розробка оригінальної архітектури та реалізація комп’ютерної системи згідно з визначеними вимогами;

Дипломна робота містить наступні розділи: “Аналіз сфери автоматизованого парсингу даних. Постановка задачі”, “Технології для розв'язання задач зі збору та зберігання великих об’ємів слабоструктурованих даних”, “Створення парсера з використанням фреймворку Scrapy”.

Дипломна робота містить 79 сторінок, 7 рисунків, 4 таблиці,

3 додатки. Було використано 14 джерел.

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

ABSTRACT

Thesis topic: "Information system for processing poorly structured data based on Scrapy".

In today's world, ways to solve the problem of collecting and processing large amounts of data are in great demand and are quite popular for study and development. One of the objects for mass collection and analysis of information was the social network Twitter. Thus, the development and application of a system for convenient search and collection of information from the web resource Twitter is now relevant.

The purpose of this work is to study the existing methods for collecting and storing large amounts of poorly structured data, finding ways to develop software for the implementation of selected methods.

The results of the work are the creation of a system that includes the implementation of the selected method of data collection, storage for the collected data and has a user-friendly interface.

The subject of this thesis is the areas of site parsing using the Scrapy framework and storage and handling of large amounts of data using non-relational databases.

Research methods are based on the field of document analysis in XML format and methods of analysis of poorly structured data.

Achieving this goal requires solving the following tasks:

Search, comparison and study of modern methods of web scraping and additional technologies;

Defining parameters for creating an information system;

Development of original architecture and implementation of a computer system in accordance with certain requirements;

Testing of the implemented system.

Thesis contains the following sections: “Analysis of the field of automated data parsing. Problem statement”, “Technologies for solving problems of collecting and storing large amounts of poorly structured data”, “Creating a parser using the Scrapy framework ”.

A diploma project contains 79 pages, 7 figures, 4 tables,

3 applications. 14 sources were used.

Key words: *web scraping, message analysis, description search, chatbot, social network*.