ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Кваліфікаційна робота

Модель автоматизації збору даних з вебресурсів на основі генеративного штучного інтелекту

Виконав: здобувач групи СПм-23-3
Копайло Ярослав Русланович
Керівник: доц. Філімончук Т.В.

ПРОБЛЕМИ ПРЕДМЕТНОЇ ГАЛУЗІ

- недосконаліть традиційних підходів до вебскрапінгу;
- ручне налаштування та складність адаптації до змін вебсторінок;
- низька точність та продуктивність традиційних скраперів;
- високі витрати часу на підтримку та налаштування;
- обмеженість масштабування традиційних моделей.

ПОСТАНОВКА ЗАДАЧІ

Розробити модель автоматизації збору вебданих, яка забезпечує високу точність та швидкість збору даних, автоматично адаптується до змін у структурі вебресурсів та мінімізує ручне налаштування.

3

ФУНКЦІОНАЛЬНІ ВИМОГИ ДО МОДЕЛІ

- автоматичне генерування XPath-запитів;
- висока адаптивність до змін DOM-структури;
- інтеграція з генеративними мовними моделями;
- автоматична корекція правил збору даних;
- гнучкість модульної архітектури.

4

НЕФУНКЦІОНАЛЬНІ ВИМОГИ ДО МОДЕЛІ

- висока продуктивність (зниження часу налаштувань);
- надійність та стабільність роботи;
- простота в підтримці та експлуатації;
- масштабованість при обробці великих обсягів даних;
- мінімальні витрати обчислювальних ресурсів.

5

ІСНУЮЧІ АРХІТЕКТУРНІ МОДЕЛІ

Базова модель: ручні налаштування XPath:

DCM, DPSM, AnM, SCI;

Модель Template-based: регулярні вирази та шаблони:

DCM, DPSM, AnM, SCI, TMM;

Модель DeepDOM: машинне навчання для аналізу DOM:

DCM, DPSM, AnM, SCI, DLM;

6

БАЗОВІ КОМПОНЕНТИ МОДЕЛІ

$$M = \{ DCM, DPSM, AnM, SCI \}$$

DCM (data collection module) – модуль збору даних;
DPSM (data processing and storage module) – модуль обробки та зберігання даних;
AnM (analytical module) – модуль аналітики;
SCI (system configuration interface) – інтерфейс налаштування системи.

7

МОДИФІКОВАНА МОДЕЛЬ

$$M = \{ CE, AM, RM, DCM, DPSM, AnM, SCI \},$$

CE (control element) – керуючий елемент;
AM (analysis module) – модуль аналізу сайту;
RM (rules module) – модуль генерації правил скрапера.

$$CE = \{ MPS, CM, MS, LS, MS, CS \},$$

MPS (message processing service) – сервіс обробки повідомлень;
CM (control mechanism) – механізм управління;
MS (mechanism of synchronization) – механізм синхронізації компонентів;
LS (login service) – сервіс логування;
MS (monitoring service) – сервіс моніторингу системи;
CS (configuration service) – сервіс конфігурації.

8

МОДИФІКОВАНА МОДЕЛЬ

M = { CE, AM, RM, DCM, DPSM, AnM, SCI },

CE (control element) – керуючий елемент;
AM (analysis module) – модуль аналізу сайту;
RM (rules module) – модуль генерації правил скрапера.

AM = { SC, FM, ClS, RTM },

SC (site crawler) – механізм обходу сайту;
FM (filtration mechanism) – механізм фільтрації унікальних сторінок;
ClS (classification service) – сервіс класифікації сторінок;
RTM (results transfer mechanism) – механізм передавання результатів.

9

МОДИФІКОВАНА МОДЕЛЬ

M = { CE, AM, RM, DCM, DPSM, AnM, SCI },

CE (control element) – керуючий елемент;
AM (analysis module) – модуль аналізу сайту;
RM (rules module) – модуль генерації правил скрапера.

RM = { RCM, GAM, RVS, RURS },

RCM (rule creation mechanism) – механізм створення правил;
GAM (generation analysis mechanism) – механізм аналізу генерації;
RVS (rule validation service) – сервіс валідації правил;
RURS (rule update and reset service) – сервіс оновлення та переналаштування правил.

10

МОДИФІКОВАНА МОДЕЛЬ

M = { CE, AM, RM, DCM, DPSM, AnM, SCI }

DCM = { REM, MM, DGS, DTM },

REM (rule enforcement mechanism) – механізм виконання правил;
MM (multithreading mechanism) – механізм багатопотокової обробки;
DGS (data gathering service) – сервіс збирання даних;
DTM (data transmission mechanism) – механізм передавання даних.

DPSM = { DClM, NM, TS, SCM, AdM, CM, IS },

11

DClM (data cleansing mechanism) – механізм очищення даних;
NM (normalization mechanism) – механізм нормалізації форматів;
TS (transfer service) – сервіс передавання даних;
SCM (schema creation mechanism) – механізм створення схеми;
AdM (adaptation mechanism) – механізм адаптації структури;
CM (control mechanism) – механізм контролю даних;
IS (integration service) – сервіс інтеграції даних.

11

МОДИФІКОВАНА МОДЕЛЬ

M = { CE, AM, RM, DCM, DPSM, AnM, SCI }

AnM = { DQM, SM, DVS, GS },

DQM (data query mechanism) – механізм запиту даних;
SM (script mechanism) – механізм створення скриптів;
DVS (data visualization service) – сервіс візуалізації даних;
GS (generation service) – сервіс генерації PDF.

SCI = { PCM, CM, RGM },

12

PCM (processing control mechanism) – механізм управління обробкою;
CM (configuration mechanism) – механізм конфігурації обробників;
RGM (report generation mechanism) – механізм генерації звітів.

12

ПОРІВНЯННЯ ЗАПРОПОНОВАНОЇ МОДЕЛІ З ІСНУЮЧИМИ

- гнучкість та адаптивність (CE, RM, AM);
- масштабованість (DMC, DPSM);
- продуктивність та ефективність збору даних (RM, DMC, SCI);
- мінімізація ручних налаштувань (RM, AM);
- висока точність роботи (AnM).

13

ПОРІВНЯННЯ ІСНУЮЧИХ МОДЕЛЕЙ С ЗАПРОПОНОВАНОЮ

Підхід	Відсоток успіху (%)
LLM-based модель	89
Scrapy (ручні XPath)	46
DeepDOM	71
Template-based (регулярні вирази)	35

Відсоток успішних повторних зборів даних

14

ПОРІВНЯННЯ ІСНУЮЧИХ МОДЕЛЕЙ С ЗАПРОПОНОВАНОЮ

Підхід	Час на внесення змін (години)
LLM-based модель	0.9
Scrapy (ручні XPath)	6.2
DeepDOM	2.8
Template-based (регулярні вирази)	1.9

Середні витрати часу на оновлення налаштувань моделей

15

ВИСНОВКИ

- детально проаналізовано проблеми традиційних підходів;
- сформовано вимоги та розроблено модель на основі LLM;
- підтверджено ефективність запропонованої моделі;
- надано практичні рекомендації щодо застосування.

16