Project Report: URL Management System

Overview

The URL Management System is designed to facilitate the storage, retrieval, search, and recommendation of URLs leveraging NLP techniques. It uses various modules to achieve this. This a detailed report of the functioning of the URL Manager.

---

2. Components and Modules

a. Database Management (DatabaseHandler)

Purpose: Manages the MySQL database for storing URLs, their content, summaries, keywords, and save timestamps.

Key Functions:

save_url(url): Saves a new URL entry into the database.

fetch_content(url): Retrieves and saves the main content from a URL (subject to robots.txt restrictions).

keyword_summarize_text(url): Generates summaries and extracts keywords using NLP techniques.

search_urls(keyword): Searches for URLs containing a specified keyword in their keywords field.

b. URL Search and Retrieval (SearchHandler)

Purpose: Provides functionality to search for URLs based on keywords stored in the database.

Key Functions:

search_urls(keyword): Retrieves URLs that contain a specified keyword in their keyword field.

c. URL Recommendation System (Recommender)

Purpose: Generates recommendations for URLs based on keywords extracted from the last saved URL.

Key Functions:

recommend_urls(): Fetches URLs based on keyword matches and relevance

send_recommendation_email(receiver_email): Sends a daily email with recommended URLs to a specified email address.

d. Scheduler and User Interface (URLManager)

Purpose: Orchestrates the functionalities provided by DatabaseHandler, SearchHandler, and Recommender through a command-line interface and scheduling system.

Key Functions:

Command-Line Interface:

save_url_workflow(): Guides users through saving URLs, fetching content, and summarizing keywords.

search_url_workflow(): Allows users to search for URLs based on keywords.

recommend_urls(): Displays and optionally emails recommended URLs based on previous saves.

- Scheduling:

schedule_recommendations(): Sets up daily email recommendations using the schedule library.

run_scheduler(): Executes scheduled tasks in a separate thread to ensure non-blocking operation.

---

3. Implementation Details

- Technologies Used:

1. Python for application logic and scripting.
2. SQLAlchemy for database management and ORM.
3. requests and BeautifulSoup for web scraping and content extraction.
4. Natural Language Processing (NLP) libraries (nltk, spacy) for text normalization, keyword extraction, and summarization.
5. schedule library for task scheduling.
6. smtplib and email libraries for email functionality.

- Workflow:
    - Users interact through a command-line interface to save URLs, search for URLs by keywords, and receive recommendations.
    - Scheduled tasks run daily to send email recommendations to specified recipients.

---

5. Conclusion

The URL Management System provides a robust framework for managing, searching, and recommending URLs based on keyword analysis. With its modular architecture and extensible design, the system can be further enhanced to meet evolving user needs and technological advancements. By implementing the recommended enhancements, the system can offer a seamless user experience and maintain high performance under varying workloads.