WEB CRAWLER FOR LEWIS HAMILTON

Sree Harsha Koyi

SXK230025

Creating a knowledge base from web content involves several key steps: crawling web pages, scraping and storing their content, cleaning and preprocessing the text, extracting important terms using TF-IDF, and finally, organizing this information into a structured format. Below is an overview of how the knowledge base was created using Python, along with insights into each phase of the process.

**Web Crawling**

The process began with a list of start URLs related to the Formula 1 driver Lewis Hamilton. A web crawler was developed to explore these URLs to a specified depth (2 levels deep in this case) while adhering to a limit of 25 URLs to prevent overloading the server and to manage computational resources effectively. The crawler used the requests library to fetch page contents and BeautifulSoup for parsing HTML and extracting links to new pages. It carefully managed a collection of URLs to visit and ensured that no URL was visited more than once.

**Scraping and Storing Content**

Once a set of URLs was collected, the script proceeded to scrape the textual content from each page. This step also involved the requests library for HTTP requests and BeautifulSoup for parsing the HTML content to extract text. The scraped text was stored in individual text files, one per URL, to facilitate the subsequent processing steps.

**Cleaning and Preprocessing Text**

The raw text from each file was then cleaned and preprocessed to remove HTML tags, special characters, and multiple spaces. The text was converted to lowercase to maintain consistency and avoid duplication of terms based on case differences. Although the initial plan included removing stop words and tokenizing the text, the decision was made to keep sentences intact to preserve context for the knowledge base.

**Extracting Important Terms Using TF-IDF**

The Term Frequency-Inverse Document Frequency (TF-IDF) technique was employed to identify the most relevant terms across the collected documents. This step was crucial for focusing the knowledge base on the most informative and distinguishing terms related to Lewis Hamilton and Formula 1. The TfidfVectorizer from sklearn.feature\_extraction.text was used to convert the cleaned text into a matrix of TF-IDF features, and the top terms were selected based on their scores.

**Building the Knowledge Base**

With the important terms identified, the next step was to organize the cleaned text into a structured knowledge base. This involved parsing each document into sentences and associating them with relevant terms. The result was a dictionary where each key was an important term, and the value was a list of sentences from the corpus that contained the term. This structure facilitates efficient search and retrieval of information related to specific topics or terms.

**Serialization for Persistence**

To ensure that the knowledge base could be easily accessed and used in the future without needing to repeat the entire process, it was serialized into a file using Python's pickle module. This step allows the knowledge base to be saved to disk and later loaded back into memory as needed.

**Search Functionality**

Finally, a function was implemented to search the knowledge base for facts related to a specific term. This function deserializes the knowledge base from the file and returns the list of sentences associated with the given search term, providing a convenient way to access the compiled information.

My Important Terms for Lewis Hamilton:

important\_terms = ['formula', 'hamilton', 'one', 'driver', 'f1', 'season', 'mercedes', 'championship','drivers' 'season','prix','circuits' ,'racefans', 'sport']

Screenshots of Knowledge Base:

A screen shot of a computer screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

