DSCI-560 Assignment No. 4 - Part 1

Team Formation

Team name: HealthEdBot

Names & USC ID
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang(8386626967)

https://github.com/DSCI560

Minutes of the offline meeting on February 7

Participants: project team members

Conference topic: Initial Setup

Initial Setup

Tools / Libraries: Confirmed the use of requests/selenium for web scraping, BeautifulSoup4 for parsing HTML content, and MySQL for database management, leveraging prior experience and existing documentation.

Resource Utilization: Provided specific resources to guide the team on scraping Reddit, including tutorials on BeautifulSoup and Praw API, to ensure a comprehensive understanding of data extraction techniques.

Minutes of the online meeting on February 8

Participants: project team members

Conference topic: Data Collection / Storage

Explored potential data sources with a focus on public APIs and web scraping opportunities. Reddit was mentioned as a primary source due to its vast and diverse content.

Various data storage solutions were considered, including relational databases and cloud storage options, with MySQL being a preferred choice for its familiarity and scalability.

The need for a flexible data collection framework that could adapt to different data formats and volumes was discussed, though specific tools were not finalized.

Minutes of the offline meeting on February 9

Participants: project team members

Conference topic: Data Collection / Storage

Delved deeper into data collection methods, with a comparison between using APIs like Praw for Reddit and web scraping techniques for other sites.

Examined different data storage architectures, focusing on normalization, scalability, and data integrity. The discussion led to a consensus on designing a modular schema that could evolve with the project.

The conversation underscored the importance of handling API rate limits and efficient data retrieval strategies but concluded without setting a definitive approach.

Conclusion
Data Collection / Storage
Task Details: Outlined the process for creating efficient database schemas to store data collected from Reddit's /r/tech forum. Emphasized the importance of designing tables that facilitate easy data retrieval and analysis.
API Request Handling: Discussed strategies to circumvent API limits by implementing a robust request management system that could batch requests and handle timeouts gracefully.
Input Flexibility: Highlighted the requirement for the script to be versatile in handling variable input sizes, ensuring scalability for different data volumes.

Minutes of the online meeting on February 10
Participants: project team members
Conference topic: Data Preprocessing
Comprehensive Preprocessing: Detailed the preprocessing steps, including cleaning text data of HTML tags and special characters, anonymizing user data to uphold privacy standards, and converting data into analysis-ready formats.
Advanced Text Analysis: Addressed the challenge of processing image-embedded text by incorporating OCR technology, aiming to enrich the dataset with additional insights derived from images.
Delved deeper into data collection methods, with a comparison between using APIs like Praw for Reddit and web scraping techniques for other sites.
Examined different data storage architectures, focusing on normalization, scalability, and data integrity. The discussion led to a consensus on designing a modular schema that could evolve with the project.
The conversation underscored the importance of handling API rate limits and efficient data retrieval strategies but concluded without setting a definitive approach.

Minutes of the online meeting on February 11
Participants: project team members
Conference topic: Data Preprocessing
Introduced the concept of data preprocessing, highlighting the necessity to clean and structure raw data for analysis.
Mentioned various tools and libraries that could aid in data preprocessing, such as Pandas for data manipulation and NLTK for text processing, alongside requests/selenium and BeautifulSoup4 for web scraping.
The team acknowledged the need for a preprocessing strategy that includes data validation and normalization but decided to detail the exact methodologies in subsequent sessions.

Finally we recorded a video to show our work.

Team name: HealthEdBot
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang(8386626967)