Progress Report on Automated Data Gathering and Analysis Project

Date: 11/13/2023

1.3 Team Members & Roles

- Bo Hu (bo12@illinois.edu)
- Bo Tian (botian3@illinois.edu)
- Weikun Wu (weikunw2@illinois.edu)
- Yangliang Lu (yl164@illinois.edu)

Submitted by: AutoDash

Phase 1 Accomplishments (2 - 4 weeks)

1. **GitHub Repository and Directory Structure**
    - **Responsibility**: Yanglian
    - **Status**: Completed
    - **Details**:Initialized the GitHub repository. Designed an efficient directory structure for organizing project files, ensuring smooth collaboration and version control..
2. **User Input Prompt Creation**
    - **Responsibility**: Botian
    - **Status**: Completed
    - **Details**: Implemented a user-friendly prompt in the interface that allows users to input specific keywords for data retrieval.
3. **Search Engine API Integration for Document Scraping**
    - **Responsibility**: Weikun
    - **Status**: Completed
    - **Details**: Created a function to interface with a search engine API. This function successfully retrieves top relevant documents based on user-given keywords.
4. **Database Selection and Utility Function Creation**
    - **Responsibility**: Yanglian
    - **Status**: Completed
    - **Details**: Choose an appropriate database for storing scraped documents. Developed utility functions to facilitate smooth data handling and retrieval.
5. **Document Preprocessing Function**
    - **Responsibility**: BoHu
    - **Status**: Completed
    - **Details**: Established a comprehensive preprocessing pipeline for the scraped documents, including tokenization, stemming, and lemmatization. The processed data is stored back in the database for efficient retrieval.
6. **Dashboard Design and Data Visualization**

    - **Responsibility**: BoTian
    - **Status**: Completed
    - **Details**: Designed the initial layout of the interactive dashboard. Created mock data to test and visualize how input data is presented.

Adjustments to the Project Plan

After consideration and team discussions, we decide to drop the following tasks as they are no longer deemed necessary for our project's objectives:

- **Dropping Network Graph and PageRank Analysis**: Initially tasked to Yanglian, this involved creating a network graph and using PageRank or authority-hub analysis for document scoring. We concluded that this component is not critical for the current scope of our project.
- **Dropping Pointwise Mutual Information (PMI) for Query Expansion**: Also under Yanglian's purview, this task aimed at implementing PMI for more precise user query results. After re-evaluating our goals, we've decided this feature is not essential at this stage.