INFORMATION RETRIEVAL PROJECT

PROJECT TITLE - **BookHarbor: A Literary Exploration Engine**

UNDER THE GUIDANCE OF - **Professor Houwei Cao**

TEAM MEMBERS:

Tanvi Dipan Patel - 1316221

Bahaduri Prachiti Jagdish - 1317686

Pankti Bhatt - 1287645

Pratiksha Gurudev Kande - 1309129

TABLE OF CONTENTS

ABSTRACT

The "BookHarbor: A Literary Exploration Engine" project aims to revolutionize literary exploration in the digital age by offering an intuitive tool for discovering books, authors, and genres tailored to users' interests and learning objectives. It addresses the challenge posed by the overwhelming abundance of digital literary content, providing a seamless interface for users to navigate and discover relevant materials.

Here are some of the key points of our project:

- **Purpose:** BookHarbor aims to simplify literary exploration by providing an intuitive tool for discovering books, authors, and genres aligned with users' preferences and learning goals.
- **Techniques:** Leveraging sophisticated web scraping, data processing methods, and user-centric interface design, BookHarbor utilizes Python, Flask, and JavaScript for its seamless functionality.
- **Data:** Collected from a variety of repositories, including Open Library and Archive.org, using advanced web scraping techniques to ensure a comprehensive database.
- **Benefits:** BookHarbor enhances user experiences by facilitating efficient exploration of literature, saving time through quick and personalized results, and providing a user-friendly interface.
- **Future Implications:** The project's roadmap includes continuous enhancements such as refining search algorithms to enhance accuracy, expanding the database to encompass an extensive collection of literary works, and incorporating user feedback for ongoing improvements.

INTRODUCTION

Defining the Problem

"BookHarbor: A Literary Exploration Engine" is our solution to a big problem: too many books, too little time! In today's digital world, a flood of books is available online, making it tough for readers to find the right ones. People often feel lost in this sea of options, struggling to discover books, authors, or genres that match what they're looking for. We aim to create an easy, user-friendly way for book lovers to find exactly what they want without getting overwhelmed.

Motivation

The reason we're making BookHarbor is simple: finding the perfect book shouldn't be like finding a needle in a haystack! Every day, more and more books are added online, which is awesome but also overwhelming. It's like looking for a diamond in a treasure chest. We're here to make this process smoother by building a tool that helps people quickly find books that they'll love and that fit their interests and learning goals.

Examples and Applications

Imagine you're in a massive digital library, but you don't know where to start. You're looking for something specific, but there are just too many options, and you're not sure which ones are right for you. That's what it's like for many book lovers online! BookHarbor is like having a super helpful librarian who knows exactly what you're looking for. It'll make it simple to find books that match your tastes and what you want to learn, all in one place.

REVIEW OF RELATED WORK

Existing Literature

Before we created BookHarbor, we checked out other projects and research papers about exploring books online. We found some cool stuff that helped us understand how people are exploring literature. There are blogs, websites, and even some tools similar to what we're building.

Intersection with Our Project

Our team worked collaboratively to conceptualize and shape BookHarbor. Through collective brainstorming and shared insights, we identified what users might need in a literary exploration tool. Our discussions and collaborative efforts allowed us to create a platform tailored to users' preferences and made exploring books online simpler and more intuitive.

Relevant Sources and Their Significance

We looked at different sources, like academic papers, blogs by book enthusiasts, and even other tools like BookFinder and Goodreads. These sources were super helpful in understanding what readers want when exploring books online. They showed us what features people love, what's missing, and what could be improved.

DESCRIPTION OF DATASETS

Sources and Collection Methods

For BookHarbor, our primary data sources encompassed renowned literary repositories like Open Library and Project Gutenberg. Leveraging web scraping techniques, we aimed to extract comprehensive book details such as titles, authors, genres, and additional informative links.

Challenges Encountered

Our data collection journey was riddled with challenges, particularly concerning access limitations. Numerous websites restricted our ability to extract information, significantly restricting our data sources. However, among the limited accessible platforms, we primarily extracted data from two prominent websites, Open Library and Project Gutenberg.

Data Annotation Procedures

The nature of our data extraction process limited the need for extensive annotation. However, we meticulously performed data validation checks to ensure the accuracy and consistency of the retrieved information from both Open Library and Project Gutenberg.

Representative Examples

Offering a glimpse into the diversity of our dataset, here are a few noteworthy examples:

- "To Kill a Mockingbird" by Harper Lee: Fiction, Classic - Open Library: To Kill a Mockingbird
- "Sapiens: A Brief History of Humankind" by Yuval Noah Harari: Non-Fiction, History - Open Library: Sapiens
- "The Hobbit" by J.R.R. Tolkien: Fantasy, Fiction - Project Gutenberg: The Hobbit

APPROACH AND METHODOLOGY

1. **Technical Methods Employed:**

- Web Scraping: Utilized for extracting comprehensive book details from various online repositories like Open Library and Project Gutenberg.
- Data Processing: Involved parsing and organizing the scraped data to ensure accuracy and coherence in the information presented.
- User Interface Design: Created an intuitive interface to facilitate efficient book searches and enhance user experience.

2. **Tools Utilized:**

- Python: Leveraged for its versatility and extensive libraries, serving as the primary scripting language for web scraping, data processing, and backend functionalities.
- Flask: Implemented as the web framework, enabling the development of a robust and scalable web application, ensuring seamless data presentation and interaction.

3. **Significance of Tools:**

- Python: Its extensive libraries facilitated data retrieval, processing, and backend tasks, ensuring a comprehensive collection of book information.
- Flask: Provided the foundation for constructing a user-friendly web interface, enabling efficient book searches and seamless user interaction.

4. **Original Contributions:**

- The integration of Python and Flask formed the backbone of this project, enabling efficient data extraction, processing, and user interaction. This approach ensures users have access to a diverse and well-organized collection of literary resources.

5. **Libraries Utilized:**

- requests: Crucial for making HTTP requests in our project, fetching web pages while performing web scraping from Project Gutenberg and Open Library.
- BeautifulSoup: Used for parsing HTML and extracting relevant data from web pages.
- urllib.parse: Specifically used for URL manipulation and parsing, assisting in constructing URL queries for search requests and joining URL parts.
- Flask: This micro web framework forms the backbone of our web application.
- sklearn: Modules like TfidfVectorizer and cosine_similarity are employed for text vectorization and calculating similarity metrics.

EXPERIMENTS AND RESULTS

1. Evaluation Methodology:

- **Metrics Used for Evaluation:** In our evaluation, we primarily employed the TF-IDF (Term Frequency-Inverse Document Frequency) cosine similarity metric. This metric measures the similarity between book titles obtained from different sources. We set a threshold of 0.8 for similarity, considering titles above this threshold as matches.
- **Baseline and Experiment Setup:** Our baseline involved comparing book titles from Project Gutenberg and Open Library using traditional search methods. The experiment consisted of BookHarbor's search capabilities, leveraging machine learning-based similarity algorithms to find corresponding book titles.
- **Conducting Experiments:** The experiments were conducted using a combination of actual user search queries and predefined book titles across various genres. We measured the matching accuracy and speed of BookHarbor against traditional search methods.

2. Presenting the Results:

- **Findings and Comparisons:** BookHarbor successfully identified a substantial number of unique book titles across diverse genres and authors, outperforming traditional search methods. We compared the accuracy and speed of BookHarbor's searches against the baseline, showcasing the efficiency and effectiveness of our tool.

3. Discussion on Significance:

- **Accuracy and Efficiency:** BookHarbor's accuracy in identifying similar book titles was notably higher compared to traditional search methods. Additionally, it displayed efficiency by swiftly presenting relevant titles, thereby enhancing the user experience.
- **Scalability:** We observed BookHarbor's potential scalability by conducting experiments with an increasing number of search queries. Despite the volume, the tool maintained its accuracy and speed, indicating scalability in handling larger datasets.
- **Encountered Issues:** While BookHarbor demonstrated remarkable accuracy and efficiency, there were occasional challenges related to book titles with very similar content or titles from different editions.

CONCLUSION AND FUTURE WORKS

Main Contributions and Key Findings

BookHarbor significantly simplifies literary exploration by providing an intuitive platform for discovering books across diverse genres and authors. Our experiments demonstrated the tool's efficiency in finding similar book titles, showcasing its potential in aiding users' exploration of literature.

Successful Aspects and Challenges Faced

The integration of Python and Flask allowed for a robust tool that efficiently scraped and processed data from prominent repositories. However, challenges were encountered in accessing certain platforms due to restrictions, limiting the breadth of data sources.

Considerations for Future Work

- **Enhanced Search Algorithms:** Improvements in search algorithms could refine BookHarbor's accuracy and broaden its range of recommendations.
- **Database Expansion:** Expanding the database to encompass a wider array of literary sources will offer users a more extensive collection of books and genres.
- **Integration of User Feedback:** Incorporating user feedback mechanisms will aid in continuously enhancing BookHarbor based on user preferences and suggestions.
- **Scalability and Accessibility:** Ensuring the tool's scalability and compatibility across various devices will enhance user accessibility and convenience

INDIVIDUAL CONTRIBUTIONS

ROLES AND RESPONSIBILITIES:

1. **Pankti:** Responsible for sourcing and organizing book data from various repositories. Ensured the collected data was structured, accurate, and extensive, forming the foundation of BookHarbor's database.
2. **Prachiti:** Collaborated closely with Pankti on the backend, focusing on data processing and refining. Played a crucial role in refining the scraped information, ensuring its quality and coherence.
3. **Tanvi:** Led the frontend development, crafting the interface to be user-friendly. Created an intuitive design for BookHarbor, enabling seamless book searches and user interaction.
4. **Pratiksha:** Partnered with Tanvi on the front end, contributing to the visual aesthetics and responsiveness of BookHarbor. Worked on the design elements to enhance the user experience.

REFERENCES

- Readwell: A Book Recommender Android Application with a Point-of-Sales System Madelein Villegas; Edric Castel Hao; Ira Clark Ungos; Ana Antoniette Illahi; John Anthony C. Jose 2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)
- Web-based personalized hybrid book recommendation system Salil Kanetkar; Akshay Nayak; Sridhar Swamy; Gresha Bhatia 2014 International Conference on Advances in Engineering & Technology Research (ICAETR - 2014)
- Data Analysis by Web Scraping using Python David Mathew Thomas; Sandeep Mathur 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)
- Course recommendation based on semantic similarity analysis Hualong Ma; Xiande Wang; Jianfeng Hou; Yunjun Lu 2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE)
- Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef Shilpa Chaudhari;R. Aparna;Vinay G Tekkur;G L. Pavan;Shreekanth R Karki 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)
- Text similarity detection method of power customer service work order based on TFIDF algorithm Lifeng Du; Changhua Hu 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE)