Computer and Information Systems Project

Kyle Sinclair

3538597

COMP 495

06 Jan 2024

Project Proposal Draft

# Introduction

In today's fast-paced world, where information is abundant and readily available, the challenge often lies in the ability to distill this wealth of data into concise, digestible formats. This is particularly true in the realm of academic research, where time constraints and the sheer volume of publications can make it challenging for individuals to stay informed about recent studies in their fields of interest. The need for a solution that provides brief overviews of academic research is evident to serve those who wish to keep up with new findings without wading through extensive papers that may only marginally touch upon their specific interests.

Addressing this need, this project leverages the Entrez Direct API to systematically access PubMed. The core objective is to automate the daily collection of study abstracts, which are often dense and packed with specialized jargon. By applying TensorFlow alongside sophisticated natural language processing (NLP) techniques, the project aims to transform these exhaustive abstracts into more accessible summaries. This endeavor is not just about simplification but also about maintaining the essence and integrity of the original research, making it comprehensible to a wider audience without compromising its scientific accuracy.

The culmination of this process will see these transformed abstracts housed in a database, forming the backbone of a user-friendly website. This online platform is envisioned not merely as a repository but as an interactive gateway, enabling users to easily explore and engage with the latest research findings in a format that respects their time and cognitive load. The project, therefore, stands as a bridge between the expansive world of academic research and the practical needs of everyday information consumption.

# Project Objectives

The objective of this project is to apply a transformative process, employing TensorFlow and advanced natural language processing (NLP) techniques to parse and simplify these abstracts. This is not a mere exercise in reduction but a careful crafting to ensure that the summaries remain true to the original research while being comprehensible to non-experts. The culmination of this effort is the development of a user-friendly website, envisioned as more than just a static repository. It aims to be an interactive platform where users can engage with content, easily navigate through different topics, and interact with the material in a meaningful way. The project thus stands as a testament to the power of technology in enhancing the accessibility and understanding of complex scientific knowledge for a broad audience.

# Methods/Technical Strategy

The methodology underpinning this project is anchored in the seamless integration of sophisticated technology and user-focused design. Central to the project's strategy is the automated collection of study abstracts from PubMed, facilitated through the Entrez Direct API. This automated process ensures a consistent influx of current academic research, providing a rich source of data for further transformation. The core of this transformation is the application of TensorFlow and advanced Natural Language Processing (NLP) techniques. The aim is to ensure that the essence and accuracy of the original research are retained, and to make these insights accessible to a broader audience through a digestible sized amount of information. This dual focus on simplification and integrity is crucial in bridging the gap between specialized academic research and general public understanding.

Complementing the data transformation process is the development of a user-friendly website, serving as the primary platform for accessing these restructured abstracts. The website design will emphasize intuitive navigation, clear information presentation, and interactive features such as topic searches and user feedback mechanisms. This approach aims to create an engaging and educational online experience, facilitating easy exploration of the latest scientific findings. Throughout the project, a strong commitment to ethical and legal compliance in data handling is maintained, ensuring the project aligns with all relevant standards and regulations. The combination of Python for data processing, SQL for database management, and modern web development technologies for the user interface encapsulates the project's technical strategy. This blend of advanced AI processing and user-centric design is poised to create a novel resource that enhances public understanding and accessibility of complex scientific research.

# Required Materials

For the successful execution of the project, a range of essential materials and resources will be utilized, encompassing software tools, database systems, and development technologies.

**Raspberry Pi:** This compact and powerful computing device will be used to automate the daily collection of PubMed articles. Its reliability and low power consumption make it ideal for running continuous tasks like cron jobs.

**Entrez Direct API:** This API will be the primary tool for accessing PubMed's database, allowing for the automated retrieval of academic research abstracts.

**Python Programming Language:** Python, known for its versatility and wide-ranging libraries, will be the main programming language for data processing and transformation tasks.

**TensorFlow Library:** As a leading machine learning framework, TensorFlow will be utilized for applying Natural Language Processing (NLP) techniques to analyze and simplify the collected abstracts.

**SQL Database:** A robust SQL-based database system will be employed for storing, managing, and organizing the transformed data. This choice ensures efficient data handling and retrieval capabilities.

**Django Web Framework:** Django will be used to develop the website through which the content is delivered. Its compatibility with Python and its robust framework makes it suitable for creating a dynamic and interactive user interface.

**Git and GitHub:** These tools will be integral for version control and project collaboration. Git allows for efficient tracking of changes, while GitHub provides a platform for code storage and team collaboration.

**Online Tutorials and Resources:** A range of online educational materials, including tutorials, forums, and documentation, will be accessed to support the project's development, especially in areas like advanced Python programming, TensorFlow, Django development, and database management.

# Project Milestones

**Initial Planning and Setup (10 hours):** This foundational phase involves preparing the Raspberry Pi for its role in automating data collection, setting up and integrating the Entrez Direct API for accessing PubMed articles, and installing necessary software and libraries including Python and TensorFlow. Simultaneously, a project repository on GitHub for version control and collaboration will be established. This stage is crucial for laying the groundwork for all subsequent activities in the project.

**Data Collection and Preprocessing (30 hours):** The next step involves developing and testing a cron job script on the Raspberry Pi for the daily retrieval of study abstracts from PubMed. Ensuring consistent and reliable data collection is a key focus. Once data collection is in place, the attention turns to processing these abstracts. Python scripts, enhanced with TensorFlow and NLP techniques, will be programmed and refined to parse and simplify the abstracts. The objective is to maintain the accuracy and clarity of the information while making it more accessible.

**Database Implementation and Web Interface Development (48 hours):** Following the processing of data, a dual focus on database and web development begins. The project moves into setting up and configuring an SQL database for efficient and organized data storage and management. In parallel, the website's development using the Django framework gets underway. This includes designing and programming a user-friendly website with navigational and interactive features and conducting usability testing for any necessary refinements.

**Final Integration, Testing, and Launch (16 hours):** The subsequent phase involves the integration of all developed components: the automated data collection system, the data processing scripts, the database, and the website. This stage is critical for ensuring that each component functions cohesively as part of the larger system. Comprehensive testing is conducted to verify the functionality and reliability of the entire setup. Following successful testing, the website will be officially launched.

**Ongoing Launch, Monitoring, and Review (Continuous, with an initial dedicated review and documentation time of 15 hours):** Post-launch, the project enters a phase of continuous monitoring and maintenance, focusing on system performance and content updates. An initial review and documentation period is allocated to assess the project outcomes against its goals and to compile a comprehensive record of the project's journey, challenges faced, solutions implemented, and key learnings. This documentation serves as a vital resource for future reference and continuous improvement.

# Project Deliverables

1. **Automated Data Collection System:** A fully functional system on the Raspberry Pi for automated data collection, utilizing the Entrez Direct API to retrieve study abstracts from PubMed. This system will include a robust cron job script ensuring daily, reliable collection of data.
2. **Processed Data Set:** A comprehensive data set of processed and simplified abstracts. This dataset will be the result of applying TensorFlow and NLP techniques to the raw abstracts, transforming them into more accessible formats while retaining the core scientific information.
3. **SQL Database:** A well-structured SQL database designed to efficiently store, manage, and retrieve processed abstracts. This database will be integral to the overall functionality of the project, ensuring that data is organized and easily accessible for web interface integration.
4. **User-Friendly Website:** A fully developed website, using the Django framework, which presents simplified abstracts in a user-friendly and interactive manner. The website will feature intuitive navigation, search functionality, and a clean, engaging design, making it easy for users to explore and understand the latest scientific research.
5. **Integration and Testing Documentation:** Comprehensive documentation detailing the integration and testing process of all project components. This documentation will include descriptions of system functionalities, testing methodologies, and the outcomes of various testing stages, ensuring a transparent and replicable process.
6. **Final Project Report:** A detailed final report encapsulating the entire project from conception through to completion. This report will cover methodologies used, challenges encountered, solutions implemented, and the final outcomes. It will serve as a record of the project’s journey and its contributions to enhancing the accessibility of academic research.

# Conclusion

In conclusion, this project represents a significant stride towards bridging the gap between complex academic research and public accessibility. By leveraging advanced technologies such as TensorFlow and NLP, alongside the practical utility of the Raspberry Pi and Django, it aims to transform the world of scientific research. The project's deliverables are all geared towards making the latest research findings accessible. This endeavour serves a practical purpose in empowering professionals, students, and the curious public with streamlined, digestible information. The anticipated impact of this project extends beyond its immediate outcomes, setting a precedent for the innovative use of technology in educational and informational dissemination, and contributing meaningfully to the ongoing dialogue about accessibility in the realm of research.