Our team has decided to develop a web application that takes a user’s input of a scientific research paper and output relevant papers to the user, while also summarizing the paper and visualize a timeline of how recent the paper and its references are. My part in the senior design project focuses on developing a data science feature that utilizes classical methods to determine the similarity and distance between documents. This project is particularly significant from my academic perspective as it integrates various concepts I have learned throughout my computer science education. By leveraging classical algorithms, I aim to create a solution that is both efficient and effective in analyzing textual data. This project not only allows me to apply theoretical knowledge but also provides an opportunity to contribute to the field of data science by enhancing document comparison techniques.

Throughout my college education at the University of Cincinnati, I have taken numerous courses that have prepared me for this project. Courses such as Data Structure have provided a strong foundation in data structures and algorithms, essential for understanding and implementing classical methods for document similarity. Additionally, CS 4033: AI Priciples and Applications introduced me to machine learning concepts, which are crucial for understanding the underlying principles of data analysis. Also courses such as STAT 3041 Data Science and Statistics, CS5154 Information Retrieval and CS 5152 Intelligent Data Analyis have equipped me with both technical skills, such as programming in Python and using libraries like NumPy and Pandas, to properly perform query and analysis as well as data mining, such as problem-solving and critical thinking. I plan to apply these skills to develop robust algorithms for my project.

My co-op experiences have also significantly contributed to my readiness for this project. As a Machine Learning Intern at Saigontourist Cable Television Company, I gained hands-on experience in performing exploratory data analysis and developing machine learning models using Python and TensorFlow. This experience honed my technical skills in data manipulation and model training, which are directly applicable to my senior design project. Additionally, my role as a Project Management Co-op at EMCOR Facilities Services taught me valuable non-technical skills, such as project planning and teamwork, which are essential for managing the project's development process. Since our application is a web application, my time at the Cincom Systems as a Full-Stack Developer would be invaluable for the team since I would apply my knowledge and help them properly manage the back-end as well as the front-end. These experiences have provided me with a comprehensive skill set that I intend to leverage in the successful execution of my project.

I am highly motivated to work on this project due to my passion for data science and its potential to solve real-world problems. The challenge of developing an efficient method for document similarity excites me, as it combines my interests in algorithms and data analysis. My preliminary approach involves researching existing classical methods, such as cosine similarity and Euclidean distance, and experimenting with their application to document data. I plan to iteratively design and test these methods to ensure accuracy and efficiency. I also expect to be able to hone my programming skills in Python and be able to analyze and visualize the data much more efficiently. My expected results include a functional tool that accurately measures document similarity, contributing to advancements in data analysis techniques.

This project would be a great stepping stone for me to break into the role of data scientist in the future. I anticipate that my project will result in a tool that not only measures document similarity effectively but also provides insights into document clustering and categorization. To evaluate my contributions, I will set specific milestones and criteria for success, such as the accuracy of similarity measurements and the efficiency of the algorithms. I will consider the project successful when it meets these criteria and can be applied to real-world datasets with reliable results. Regular feedback from peers and mentors will also guide my self-evaluation, ensuring that I maintain high standards throughout the project development process.