**CSCE 5290: Natural Language Processing**

**Project Proposal**

**Group 22**

**11601537 - Bhavitha Budeti**

**11601575 - Divya Gopisetti**

**11682376 - Hari Krishna Sai Rachuri**

**11647442 - Kartheek Gurijala**

**Title: Content Plagiarism Checker and Question-Answer Generator**

**Abstract:**

The abstract discusses a tool that combines plagiarism detection and question answer (QA) generation serving the needs of academic, publishing and content creation fields. The plagiarism detection module uses algorithms to examine content in different formats producing detailed reports with references to original sources for easy verification. At the time the automatic QA generation component applied natural language processing techniques to analyze content extracting important concepts and generating a wide range of insightful questions along with coherent answers. By utilizing methodologies like NLP and machine learning this system is primarily implemented in Python. Features a user web interface for easy access and seamless integration, with databases. The goal is to provide a solution that meets all requirements within the timeframe.

**Motivation:**

The development of the Content Plagiarism Checker and Question Answer Generator system has an objective – to assist individuals and organizations in the publishing and content creation fields. This system aims to ensure that content remains original and trustworthy while also providing a way to generate questions and answers, from that content. It offers two functionalities; a plagiarism checker and an automatic question answer generator both designed to cater to the needs of users in these sectors. By utilizing natural language processing (NLP) algorithms, machine learning techniques and Python programming language – widely used in this domain – the system analyzes content identifies concepts. Generates relevant questions along with corresponding answers. Its applications are broad ranging; it can be employed for tasks like comparing assignments for plagiarism detection or conducting quizzes, assignments or technical events for students. The future plans for this system involve improving its plagiarism checking logic enhancing question generation capabilities ensuring scalability and performance optimization while incorporating development based on user feedback. The technologies employed during its development include NLTK (Natural Language Toolkit), Streamlit framework and scikit learn library. All showcasing how this platform holds promise in providing a dependable and user-friendly solution, for content analysis and creation.

**Significance:**

This project is highly valuable, in the realms of academia, publishing and content creation. Its primary significance lies in its ability to ensure that content is original and maintains integrity, which's crucial in publishing contexts. By providing an efficient solution for detecting plagiarism this project addresses the issue of content authenticity thereby upholding academic and publishing standards. Additionally, the feature that automatically generates questions and answers contributes to creating resources like quizzes, assignments and technical events for students.

The impact of this project can also be seen in its potential to streamline content creation processes enhance honesty and improve the quality of educational materials and published content. Furthermore, by utilizing technologies such as natural language processing (NLP) algorithms and machine learning techniques this project aligns itself with cutting edge approaches that offer solutions to longstanding challenges in analyzing and creating content.

It is crucial to achieve the objectives of this project as it addresses the need for originality and high-quality content, within publishing environments. By utilizing cutting edge technologies and innovative approaches the project seeks to develop a platform that's easy to use for identifying plagiarism and generating question answer pairs. The ultimate goal is to enhance integrity improve efficiency in creating content and enhance the quality of materials.

**Objectives:**

The main objectives of this project are to create a system that's reliable and easy to use offering two functionalities; detecting plagiarism and generating question answer pairs automatically. The goals include:

**1. Plagiarism Detection**: The system aims to identify instances of plagiarism, in types of content ensuring that the material is original and trustworthy. The success of this objective will be determined by the system’s ability to detect and report instances of plagiarism with a level of precision and dependability.

**2. Question Answer Generation**: The system intends to analyze content extract concepts and generate a wide range of relevant questions that cover various aspects of the content. Additionally, it aims to produce coherent answers aligned with those questions providing insights for users. The success of this objective will be measured by the systems capability to generate insightful questions well as answers based on the provided content.

**3. User Friendly Interface**: The project strives to offer a web-based platform with an interface that's easy for users to navigate allowing them to effortlessly upload files, for plagiarism checks and question answer generation. The achievement of this goal will depend on how the system can provide an user friendly experience making it easy for users to interact with its features.

**4. Incorporation of Cutting Edge Technologies**: By utilizing technologies, like natural language processing (NLP) algorithms, machine learning techniques and the popular programming language Python the system aims to utilize tools to accomplish its objectives. The success of this objective will be assessed based on the effective utilization of these technologies to enhance the accuracy and efficiency of plagiarism detection and question-answer generation.

The success of the project will depend on how the system achieves these goals. It needs to offer an efficient and easy, to use platform for detecting plagiarism and generating question answers. Moreover, we will also evaluate its potential, for improvement, scalability and performance to determine its success.

**Features:**

The aim of developing the Content Plagiarism Checker and Question Answer Generator is to create a platform that's dependable and user-friendly allowing for plagiarism detection and automatic question answer generation. The project will involve implementing natural language processing (NLP) algorithms to analyze content building a web-based interface using the Streamlit framework and integrating modules and libraries, like NLTK, scikit learn and Beautiful Soup for text processing, vectorization and web scraping.

What makes this project unique are the techniques employed such as text fingerprinting, semantic similarity analysis and machine learning based approaches. These techniques enhance the accuracy and reliability of plagiarism detection. Additionally, transformer-based models are utilized in question generation to improve the quality and relevance of generated questions. These elements ensure that both plagiarism detection and question answer generation achieve levels of precision and efficiency.

The project milestones include developing and integrating functionalities for both plagiarism checking and question answer generation. Another milestone involves creating a user web interface while also implementing technologies, for content analysis.

These important achievements play a role in the success of the project. They help ensure that the content is authentic and reliable while also enabling us to generate valuable questions and answers based on the provided information.

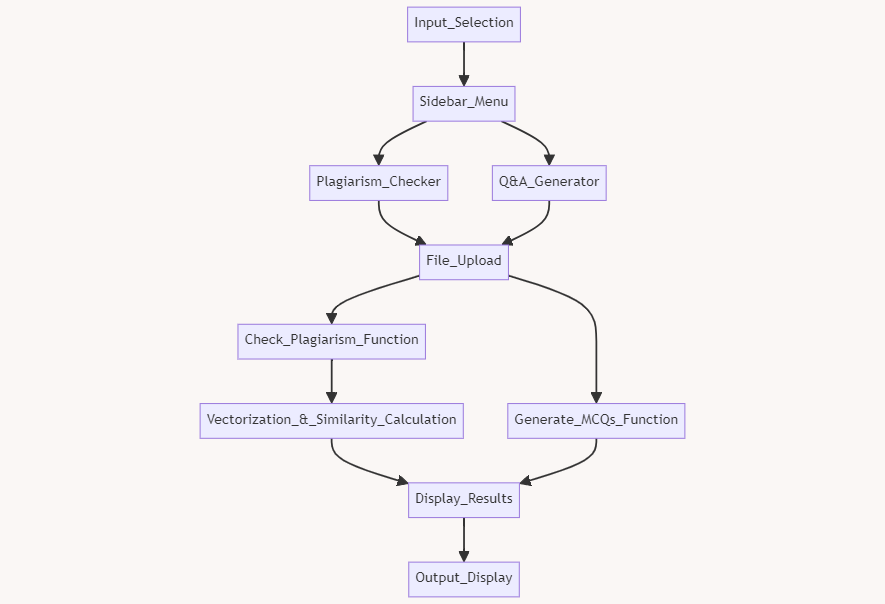
**Dataset:**

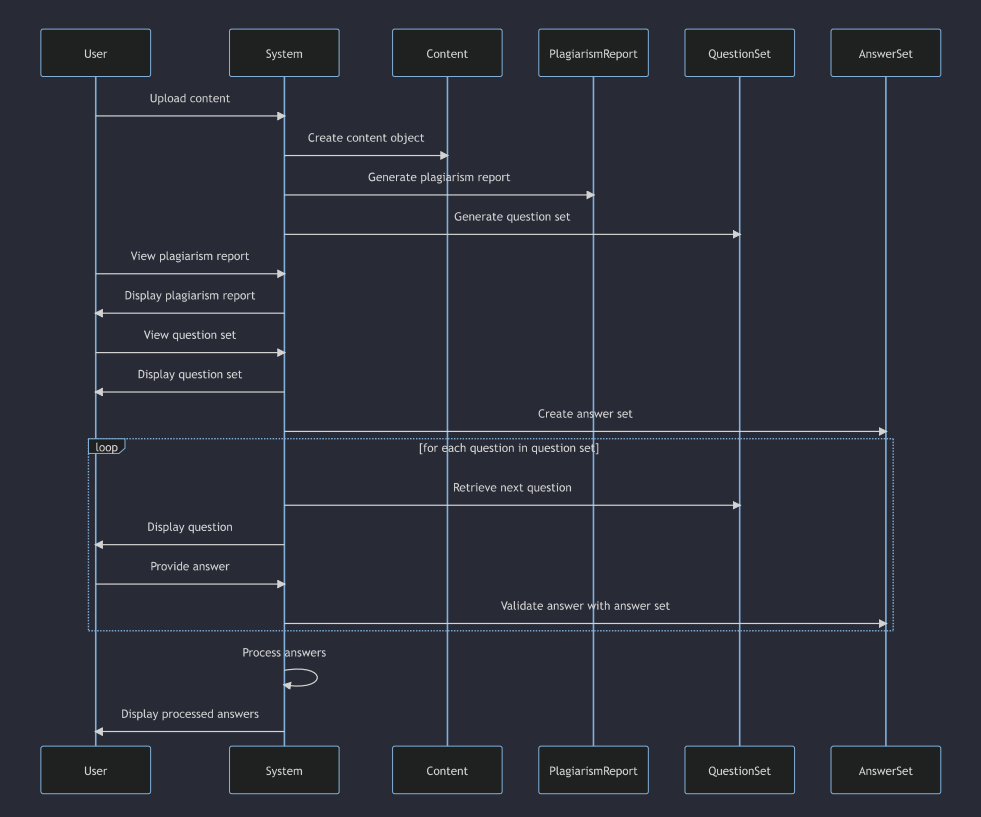
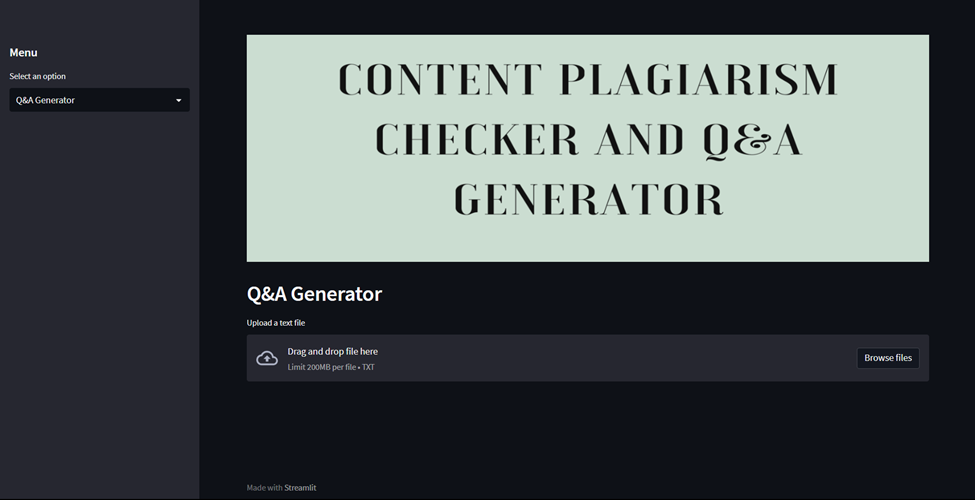
This project does not explicitly rely on a specific dataset but utilizes various input files and text input boxes within a Streamlit application to provide functionalities associated with the Content Plagiarism Checker and Question-Answer Generator system.

The input files encompass content from different domains and fields in English. These domains may include various topics such as literature, science, arts, politics, and more. These files serve the purpose of testing and demonstrating the system's functionalities. Preprocessing of the input files involves converting Microsoft Word (.docx) documents to text format using the docx2txt module, facilitating further analysis and processing. Additionally, web scraping is mentioned for extracting data from online resources, highlighting the dynamic nature of the content sources.

Within the Streamlit application, users can input text for plagiarism checking and question-answer generation. The application utilizes the content from these input sources to showcase the system's capabilities. The emphasis lies on analyzing and processing content from various sources rather than relying solely on a predefined dataset.

**Visualization:**





The workflow has been created to handle content created by users ensuring that it is original and of quality. It includes processes, for managing user generated content checking for plagiarism and facilitating the creation and validation of question answer sets within the system.

Additionally, we have developed the draft version of the home screen of this application, which is attached herewith. This home screen serves as the interface for users to interact with the system. They can upload their content. Utilize features, like plagiarism checking and question answer generation.