**Online Course Recommendation System Using NLP Techniques**

**Research Question:** How do LSTM and T5-based NLP techniques compare in effectiveness for recommending online courses based on user input, and can reinforcement learning further enhance these recommendations?

**Objectives:**

1. Collect and preprocessing online course data, including course details, ratings, difficulty level etc.
2. Implement LSTM and T5-based NLP models for course recommendation using the collected data.
3. Evaluate the performance of LSTM and T5 models in recommending courses based on user input, considering metrics such as accuracy, relevance, and diversity.
4. Explore the integration of reinforcement learning techniques to further refine course recommendations and adapt to user feedback.
5. Compare the effectiveness of LSTM, T5, and reinforcement learning-enhanced models in recommending online courses.
6. Analyse and interpret the results to identify strengths, weaknesses, and potential areas for improvement in each approach.
7. Summarize the findings from the study, draw conclusions regarding the effectiveness of LSTM, T5, and reinforcement learning in course recommendation, and provide recommendations for future research and system enhancements based on the findings.

**Project Background:** Online education has witnessed exponential growth in recent years, with an ever-expanding array of courses available to learners worldwide. However, with this abundance of options comes the challenge of navigating through the extensive catalogue to find courses that align with individual interests, learning goals, and proficiency levels. Traditional recommendation systems have been employed to address this issue, leveraging techniques such as collaborative filtering and content-based filtering, overlooking the nuances of individual learning goals and preferences. In recent years, advancements in natural language processing (NLP) have opened up new avenues for enhancing recommendation systems by enabling deeper understanding of user queries and course content. Techniques such as Long Short-Term Memory (LSTM) networks and Transformer models, exemplified by the Text-To-Text Transfer Transformer (T5), have demonstrated remarkable capabilities in processing and generating natural language, making them well-suited for course recommendation tasks. For instance, Wang et al. (2019) showcased the effectiveness of LSTM-based models in recommendation systems, while Raffel et al. (2019) introduced the T5 model as a versatile framework for various NLP tasks. However, the application of these advanced NLP techniques specifically to online course recommendation remains relatively unexplored. This project aims to fill this gap by investigating and comparing the effectiveness of LSTM and T5-based recommendation systems in suggesting online courses tailored to individual user queries, with the additional exploration of reinforcement learning techniques to further enhance recommendation accuracy and personalization (Chen et al., 2020). By leveraging the power of NLP and reinforcement learning, this research seeks to contribute to the development of more intelligent and adaptive course recommendation systems, ultimately enhancing the online learning experience for users.

**Reference:**

**Dataset Overview:** The dataset for this project was sourced from Class Central, a popular online course aggregator that lists courses from various providers. The focus of the dataset is on courses related to programming, data science, and computer science, all of which are taught in English. This dataset includes a variety of attributes such as course ratings, certificates, institutions, course type, duration, pricing, and more. The aim is to leverage this rich dataset to build and evaluate a recommendation system that can suggest relevant courses to users based on their input queries.

**Data Collection Process:** The data was collected by scraping the Class Central website (<https://www.classcentral.com/subjects>) using Scrapy, a robust web crawling and scraping Python module. The scraping process focused exclusively on courses related to programming, data science, and computer science that are taught in English. This targeted approach ensured that the dataset remained relevant to the scope of the project and manageable in size.

**Summary of the Data:** The dataset comprises various features that provide comprehensive details about each course. These features include:

* Course Details: 'course\_name', 'description', 'course\_link', 'course\_provider', 'course\_subject', 'course\_type', 'course\_institution', and 'additional\_course\_detail'.
* Ratings and Reviews: 'course\_avg\_rating' and 'course\_num\_rating'.
* Course Characteristics: 'course\_level', 'course\_is\_classroom', 'course\_is\_university', 'duration', 'start\_date', 'is\_all\_time\_best', and 'teacher'.
* Certification and Pricing: 'course\_certificate' and 'pricing'.

This rich dataset facilitates an in-depth analysis of courses, allowing the recommendation system to make well-informed suggestions based on user input.

**Ethical Considerations:** In conducting this data collection, the following ethical considerations were observed:

* Compliance with Terms of Service: The scraping process was conducted in accordance with Class Central’s terms of service to ensure that no rules were violated during data collection.
* Privacy Protection: The dataset does not contain any personally identifiable information (PII) about the users or the course instructors, safeguarding privacy and anonymity.
* Transparency and Fair Use: The data is used solely for educational and research purposes, aiming to enhance the user experience by providing more personalized course recommendations.
* Attribution: Proper attribution to the source (Class Central) is maintained, acknowledging the origin of the dataset.

**Data Management Plan:** A robust data management plan was established to handle the collected data efficiently:

* Data Storage: The dataset is securely stored in a structured format, with backups maintained to prevent data loss.
* Data Cleaning: The data underwent a thorough cleaning process to remove duplicates, handle missing values, and standardize formats, ensuring high-quality input for the recommendation system.
* Access Control: Access to the dataset is restricted to authorized personnel involved in the project, with appropriate measures in place to prevent unauthorized access.
* Documentation: Detailed documentation is maintained, outlining the data collection process, cleaning procedures, and any transformations applied to the data. This ensures transparency and reproducibility of the research.