KDD Summer Internship Assignment 2024

Objective

Create a web application that summarizes PubMed articles. This project will test your knowledge and skills in Web Development and Data Science. The integration of Generative AI models and deployment are optional and can be done for extra credit.

Instructions

1. **Data Exploration and Preparation**
   - Load the PubMed Summarization dataset from Hugging Face.
   - Explore the dataset to understand its structure and contents.
   - Preprocess the dataset to clean and prepare the text for summarization.
2. **Web Application Development**
   - Develop a web application using Streamlit or Flask.
   - The application should allow users to input or upload PubMed articles.
   - Implement the summarization feature using any method (e.g., rule-based, extractive summarization) if not using Generative AI.
   - Display the original article and the summarized version on the web interface.
3. **Optional: Generative AI Model Integration**
   - For 4th semester students: Use API-based models like GPT-4 (OpenAI) or Gemini (Google) for text summarization. Utilize provided API keys to integrate these models.
   - For higher semester students: Use open-source models such as LLaMA-2 (Meta) or T5 (Hugging Face) for text summarization. Download and fine-tune these models on your local machine.
4. **Optional: Advanced Features (Bonus)**
   - Implement user authentication to save and manage summaries.
   - Add options for users to select the length or style of the summary (e.g., brief, detailed).
   - Provide an interactive visualization of the summary quality metrics.
   - Integrate video generation capabilities using Veo (Google) for creating visual summaries or explainers of the articles.
   - Include an image generation feature using Imagen 3 (Google) to create illustrative images based on the article summaries.
5. **Documentation**
   - Document the entire process, including data exploration, model selection, fine-tuning, and web application development.
   - Ensure the documentation includes instructions for running the application and any additional setup required.

Submission Requirements

- A GitHub repository containing the code for data preparation, model fine-tuning (if applicable), and web application.
- **Optional:** A deployed version of the web application (provide the URL).
- A detailed report in the GitHub repository README.

Evaluation Criteria

- **Data Exploration and Preparation (25%)**
  - Quality and thoroughness of data cleaning and preprocessing.
- **Web Application Development (35%)**
  - Functionality, usability, and design of the web application.
- **Optional: Generative AI Model Integration (30%)**
  - Selection and fine-tuning of the model.
  - Performance evaluation and metrics.
- **Optional: Advanced Features (10%)**
  - Implementation of additional features and their impact.
- **Documentation (30%)**
  - Clarity, organization, and completeness of the documentation.

References

- Latest models like Claude 2 and LLaMA-2 offer state-of-the-art performance in language generation tasks (Analytics Vidhya, EWeek).
- Google's Veo and Imagen 3 for advanced video and image generation capabilities (Google I/O 2024).

This assignment ensures students are exposed to the latest advancements in web development and data science, with optional challenges in Generative AI for those who wish to explore further.