ASSIGNMENTS FOR DATA SCIENTIST ROLE

Note:

- ***Please solve any one of the Problem Statements***
- ***If you solve Problem Statement – 1, you are free to choose between either Component 1 or Component 2. Additionally, you can attempt both to display your skills.***
- ***If you solve Problem Statement – 2, you must solve the whole problem.***

Problem Statement-1:

The primary objective of this project is to design, develop, and deploy a deep learning solution that combines medical image segmentation and medical OCR with a strong emphasis on core model building and data processing. The project will consist of two main components:

Component 1: Medical Image Segmentation

Task: Develop a deep learning model to perform accurate and robust segmentation of medical images. This task may include:

- **Dataset Collection:** Identify and curate a dataset of medical images suitable for segmentation tasks. Consider using publicly available datasets such as:
    - **SIIM-ACR Pneumothorax Segmentation:**
        - Description: Chest X-ray images for pneumothorax segmentation.
        - Link: SIIM-ACR Pneumothorax Segmentation
    - **Data Science Bowl 2017 - Lung Cancer Detection:**
        - Description: CT scans of lungs for detecting lung cancer.
        - Link: Data Science Bowl 2017 - Lung Cancer Detection
- **Data Preprocessing:** Implement data preprocessing techniques to standardize, normalize, and augment the dataset. Techniques may include data resampling, image resizing, and contrast adjustment.
- **Model Architecture**: Design a deep learning architecture tailored for medical image segmentation. Popular architectures include U-Net, Deep Lab, and Mask R-CNN. Consider adapting these models for the specific medical imaging modality you are working with.
- **Training and Validation**: Train the model on the prepared dataset and validate its performance using appropriate evaluation metrics (e.g., Dice coefficient, IoU, F1-score).
- **Hyperparameter Tuning**: Optimize hyperparameters such as learning rates, batch sizes, and model architectures to achieve the best segmentation results.
- **Model Interpretability**: Implement techniques for model interpretability, such as Grad-CAM, to visualize and explain the segmentation results.

Component 2: Medical OCR (Optical Character Recognition)

Task: Develop a deep learning model to perform OCR on medical documents, extracting text information from scans and reports. Relevant datasets and resources include:

- **Dataset Collection:** Identify and curate a dataset of medical documents for OCR. You can consider using datasets like:
    - **Radiology Report Dataset (Kaggle RANZCR CLiP - Catheter and Line Position Challenge)**:
        - Description: Radiology reports for the RANZCR CLiP Challenge, which includes text reports along with image data.
        - Link: Kaggle RANZCR CLiP - Catheter and Line Position Challenge
    - **CheXpert:**
        - Description: A dataset of chest radiography reports with mention of findings and uncertainties.
        - Link: CheXpert
- **Data Preprocessing:** Apply preprocessing steps to the document images, such as noise removal, deskewing, and resizing.
- **OCR Model:** Build a deep learning OCR model, such as Convolutional Recurrent Neural Networks (CRNN) or Transformer-based models, for recognizing text in medical documents.
- **Training and Validation:** Train the OCR model on the dataset, optimize the recognition process, and assess its performance using appropriate evaluation metrics.
- **Post-processing**: Implement post-processing techniques to enhance the accuracy of text extraction and correct any recognition errors.
- **Integration:** Integrate the segmentation and OCR components to process both medical images and associated reports, allowing for the extraction of structured information.

Deliverables:

The project should result in a well-documented, end-to-end solution for medical image segmentation and OCR. The core deliverables should include:

- Trained deep learning models for image segmentation and OCR.
- Documentation on data processing, model architectures, and results.
- Integration of the two components to create a cohesive solution.
- Source code for model training, data preprocessing, and integration.
- Evaluation metrics and validation results.
- Instructions for deploying the solution in a healthcare environment.

Problem Statement-2:

Objective: The objective of this assignment is to evaluate the candidate's ability to extract information from research paper data, create a knowledge graph using natural language processing (NLP) techniques, and build a machine learning model to extract valuable insights from the knowledge graph.
Dataset: https://www.kaggle.com/datasets/anandhuh/covid-abstracts
Instructions:
Part 1: Data Acquisition and Preprocessing

- **Data Collection:** You will be provided with a set of research papers in a specific domain. Begin by collecting these research papers and organizing them into a dataset.
- **Text Preprocessing:** Preprocess the text data to remove noise, including stop words, punctuation, and special characters. Tokenize the text and perform stemming or lemmatization as needed.
- **Data Exploration:** Perform basic exploratory data analysis to gain insights into the dataset, such as word frequency distributions, common keywords, and any patterns within the research papers.

Part 2: Building the NLP Knowledge Graph

- **Entity Recognition:** Implement NER (Named Entity Recognition) to identify entities such as authors, organizations, and key terms within the research papers. Use an NLP library or tool of your choice for this task.
- **Relationship Extraction:** Identify and extract relationships between entities within the text. For example, identify which authors collaborated on the same papers, which organizations are affiliated with which authors, and the key terms associated with specific papers.
- **Constructing the Knowledge Graph:** Create a knowledge graph that represents the extracted entities and their relationships. You can use a graph database or a data structure of your choice to build the knowledge graph.

Part 3: Building the NLP Model

- **Feature Engineering:** Select and engineer relevant features for building a machine learning model that can extract insights from the knowledge graph. Consider features like entity centrality, keyword relevance, or co-authorship relationships.
- **Model Selection and Training:** Choose an appropriate machine learning algorithm and train a model to predict or extract specific information from the knowledge graph. This could be, for example, predicting which authors are likely to collaborate in the future or identifying the most influential papers in the dataset.
- **Evaluation:** Evaluate the performance of your NLP model using appropriate metrics. Report the accuracy, precision, recall, and any other relevant evaluation criteria.

Part 4: Presentation and Documentation

- **Results and Analysis:** Prepare a report or presentation summarizing your findings, insights, and the NLP model's performance. Include visualizations and examples to illustrate your results.
- **Code and Documentation:** Provide well-documented code with clear comments explaining each step of the process, from data preprocessing to model development. Include instructions for running the code and reproducing the results.

Submission Guidelines:

- You should submit your assignment as a well-organized report or presentation along with the code in a format that the evaluator can run and reproduce your results.
- Submit your assignment by [Due Date]. Late submissions will be penalized.
- If you use external libraries or tools, make sure to specify them in your submission.

Grading Criteria:

Your assignment will be graded based on the following criteria:

- Data collection and preprocessing
- Entity recognition and relationship extraction
- Construction of the knowledge graph
- Feature engineering
- Model selection and training
- Evaluation of the NLP model
- Clarity and quality of documentation and presentation