AI Engineer- NLP Skills Assessment

The Enigmatic Research of Dr. X

The Story:
Dr. X, a researcher known for groundbreaking work, abruptly vanished, leaving behind a collection of publications. These publications are in two formats: digital documents (.docx) and digital PDF documents (.pdf). Your task is to analyze these publications, potentially uncovering the reason for Dr. X's disappearance and understanding the nature of their research.

Your Task:
You are a NPL specialist, tasked with processing Dr. X's publications and creating a Q&A system.

1. Reading the Publications:
   - You have different types of publications: Ms Word files (.docx), PDF files (.pdf), and 'csv', 'xlsx', 'xls', 'xlsm' files.
   - Write a Python function to extract the text from these files.
   - Remember, the files might contain tables. Extract the text from these tables in a readable format, but do not attempt to reconstruct their visual layout.

2. Breaking Down the Publications:
   - The publications are extensive, so you need to divide them into smaller, manageable parts.
   - Use the cl100k_base tokenizer to break the text into chunks.
   - For each chunk, record:
     - The file name (source).
     - The page number.
     - The chunk number (hint: add a counter).
     - The text of that chunk.

3. Building a Vector Database:
   - Create a vector database from the chunked publications.
   - Use nomic embedding model to generate vector embeddings for each chunk.
   - Store the embeddings along with the chunk metadata in a vector database.

4. Creating a RAG Q&A System:
   - Develop a RAG-based Q&A system that can answer questions about the publications.
   - When a user asks a question, the system should:
     - Generate a vector embedding for the question.
     - Retrieve the most relevant chunks from the vector database.
     - Use llama LLM to generate an answer based on the retrieved chunks.
   - Hint: Can your code answer to user's questions based of the previous question?

5. Translating the publications:
   - Dr. X wrote some publications in different languages.
   - Build a tool (using any LLM) to translate between any language to English or Arabic.
   - A plus: Strive to maintain the original structure and formatting of the publications after translation.
   - Find creative ways to improve the translation accuracy and fluency.

6. Finding the Main Ideas:
   - Create a tool (using any LLM) to summarize the publications.
   - Evaluate the quality of your summaries using the ROUGE metric.
   - Experiment with different summarization techniques and prompt strategies and record the result.

7. Performance Measurement:
   - During the embedding generation, translation, summarization, and RAG processes, record the "tokens per second" processed by the LLM. This will help us understand the efficiency of your algorithms.

8. Be Creative:
   - Demonstrate your creativity. For example:

    - Develop advanced chunking methods.
    - Enhance the accuracy and clarity of your translations.
    - Create unique evaluation metrics for the algorithms.
    - Implement create algorithms for tables and charts text extraction.

9. Your Work:
   - Upload your work on Github and share the link (code should be in python).
   - Include a README.md file explaining your methodology, any significant discoveries, and the LLM used, and how you obtained it. Include information about the embedding model.
   - Provide a requirements.txt listing all required libraries.
   - Provide a maximum of 6 pages documentation of your work.

10. Rules:
    - Only use text-based NLP techniques. Do not use computer vision libraries.
    - Use local LLMs.
    - Use local vector databases, or those that can run offline.

How We Will Evaluate Your Work:

- Does your code execute correctly?
- Is your code well-structured and readable?
- Is your code efficient (including tokens per second)?
- Are your translations and summaries accurate?
- Does your RAG system provide accurate and relevant answers?
- Does your code handle different file formats and potential errors effectively?
- Do you demonstrate creativity and innovation?
- Is your documentation clear and thorough?
- Did you select appropriate LLMs for the tasks?
- Extra credits: Can your code handle 'csv', 'xlsx', 'xls', and 'xlsm' files?

Submission:

- Submit your solution as a Jupyter Notebook (.ipynb), or by attaching a link to your GitHub repository by replying to the challenge email.
- It is recommended to submit your solution even if you haven't completed all parts of the challenge, as there is always a chance to receive feedback, or be considered for future opportunities.

Conclusion:

As you complete your analysis, you've not only processed the publications but also pieced together fragments of Dr. X's work. The RAG system you've built serves as a digital conduit, allowing us to ask questions and gain insights that were once locked away. However, the mystery of Dr. X's disappearance may still linger.

Remember, no AI application is perfect, and even the most advanced systems can only reveal what is present in the data. The journey of discovery is ongoing, and your dedication and creativity are invaluable in pushing the boundaries of what's possible. We hope that your work reveals important information about Dr. X's research and perhaps provides clues to their whereabouts. We are excited to see your innovative solutions and wish you the best in your journey!