Task Description: Building a Topic-based Post Recommendation System using Generative AI

Overview:

Your challenge is to develop a system that recommends posts to users on a platform like Reddit or Quora. Unlike traditional methods such as collaborative filtering, this system will use advanced Generative AI techniques to understand and match user interests with relevant posts. Example, if a post talks about “Blockchain” and “Cryptocurrency” then it needs to be recommended to people who are interested in those topics and not to the ones interested in widely different topics such as “Retirement Savings” or “Fixed Deposits”.

Objective:

Create a system that uses Generative AI to recommend posts based on their topics. You will be provided with datasets for training and testing.

NOTE:

1. Please do not use the following conventional (pre GenAI era) NLP techniques: collaborative filtering, content-based filtering, neural collaborative filtering, topic models etc.
2. Feel free to use GenAI models (aka LLMs) such as GPT-4/3.5, Llama-2, Claude-2, Mixtral, or LLMs from websites/providers such as openai, together.ai, anyscale, Ollama, etc.
3. In the context of this challenge, a topic is nothing but the title of a Wikipedia Article. For Example,
    a. **Exchange Traded Funds** is a valid topic as it has a page in Wiki: https://en.wikipedia.org/wiki/Exchange-traded_fund
    b. **Parag Parikh Fund House** is not a valid topic as it does not have an associated Wikipedia page
4. Feel free to use AI coding assistants like ChatGPT plus or GitHub Copilot.
5. Sure bonus points will be given if you attempt advanced prompt engineering ideas like:
    a. Chain of Thought Prompts
    b. ReAct Agents (LangChain)
    c. Self-reflection
    d. Meta-prompting
6. Sure bonus points will be given if you choose to use advanced LLM agents and tools in pyautogen, autogen studio, LangChain and LlamaIndex.

Datasets:

Training Data (Reddit_data_train.json): Contains users and their posts, used to understand user topics of interest.

Testing Data (Reddit_data_test.json): Contains posts with ground truth data (i.e., which users interacted with a particular test post). This indicates which posts should be recommended to which users.

Core Components:

User Topic Profiling:

Objective: Create a user profile based on their interest, derived from their posts.

Method: Use Generative AI to analyze post content and identify up to 20 key topics (maximum 10 topics per user post).

Post Topic Profiling:

Objective: Assign topics to each post.

Method: Analyze posts to extract relevant topics, with the same topic limits as user profiles.

Recommendation Engine:

Objective: Match posts with the top 10 most relevant users based on their interests.

Approach: Use Generative AI to understand post and user profile topics and recommend posts to matching users.

Evaluation:

Metrics: Use Normalized Discounted Cumulative Gain (NDCG) and Jaccard similarity to evaluate the effectiveness of the recommendations.
Procedure: Compare the system's recommendations with the ground truth data in the testing set to measure accuracy and relevance.

Deliverables:
- A functional topic-based post recommendation system created using Generative AI.
- A report detailing the system's methodology, including the approach for user and post profiling.
- Code and Resources: All code developed for this task, along with any additional resources used (e.g., libraries, tools).

-If possible, include a small screen recording of the Topic based post Recommendation System demo

Finally, and most importantly, what do we look for in a GenAI Data Science Intern through this task:

As you engage with the task at hand, please be aware that our assessment will primarily focus on your approach to problem-solving and the appropriateness of the solutions you propose. We are particularly interested in seeing how you apply AI and data science principles and methodologies to address the given challenge. While technical accuracy and coding efficiency are valued, our evaluation will place greater emphasis on your creative and analytical thinking within the GenAI context. We encourage you to think innovatively, demonstrate your understanding of data-driven AI-assisted decision-making, and showcase your ability to extract insights and formulate strategies from complex datasets. Best of luck!