Name – Abhishek Krishna **Date – 26th Jan 2023**

Reg. No. – 20BDS0053

FarmwiseAI

Task 2: Topic-based Post Recommendation System using Generative AI

REPORT

1. Introduction:

- This report presents a methodology for developing a recommendation system using user posts data. The primary goal is to comprehend user posting behaviour, profile both users and posts, and ultimately create a system that suggests posts based on user preferences. The methodology covers key stages, including data cleaning, user profiling, post profiling, the recommendation system, and evaluation metrics.

2. Data Cleaning:

- The initial project phase involves loading and exploring user posts data from the provided JSON file (cleaned_user_posts.json). A thorough dataset examination reveals essential information, such as user IDs and post texts. Data cleaning ensures the removal of irrelevant or erroneous entries, establishing a more accurate foundation for subsequent analysis.

3. User Profiling:

- TF-IDF Vectorization – User profiling begins with TF-IDF vectorization of post texts, transforming text data into numerical vectors. The TfidfVectorizer from the sklearn.feature_extraction.text library quantifies the importance of each term in a post relative to the entire dataset, forming the basis for user profiling.

- Cosine Similarity – To measure user similarity based on posts, cosine similarity is calculated. This metric assesses the cosine of the angle between two non-zero vectors, representing each user as a vector in the TF-IDF space. The resulting similarity metric forms the foundation for user profiling, capturing the essence of posting styles.

4. Post Profiling:

- TF-IDF for Post Content – Each post is treated as an independent document for post profiling. TF-IDF vectorization captures term importance in posts relative to the entire dataset, laying the groundwork for comparing and recommending posts based on content similarity.

5. Recommendation System:

- The system suggests posts tailored to users' posting behaviour, employing cosine similarity to identify posts similar to those made by a given user. The implementation in Python utilizes libraries such as numpy and sklearn.metrics.pairwise.cosine_similarity, ensuring modularity for adaptation to different datasets.

6. Evaluation:

- The recommendation system's performance is evaluated using the Normalized Discounted Cumulative Gain (NDCG) metric, a widely accepted ranking quality metric. NDCG considers both the relevance and ranking of recommended posts, providing a comprehensive measure of the system's effectiveness.

- Interpretation of NDCG Scores – Interpreting involves considering the trade-off between relevance and position, with higher scores indicating better performance.

7. Resources and Code:

- Implemented in Python using libraries such as json, sklearn, numpy, and documentation resources which include pandas, scikit-learn, numpy, json, and ChatGPT by OpenAI.

- Model Reference – GPT-2 model used for text generation, inspired by OpenAI's GPT-3 model, chosen for its free-to-use availability.

8. Conclusion:

- The outlined methodology provides a comprehensive approach to building a recommendation system based on user posts data, ensuring adaptability to various datasets and user scenarios using TF-IDF vectorization, cosine similarity, and NDCG metrics.

GitHub Link: https://github.com/akrishna5/FarmwiseAI_TASK2_DataScience