**Building a Topic-based Post Recommendation System using Generative AI**

**Text Preprocessing:**

* Replace newline(“\n”) characters
* Replace html tags(“&gt”)
* Remove extra whitespaces and tabs
* Convert all characters to lowercase

**User Profiling:**

* Using PaLM to get top 10 wikipedia articles for each post.
* Using mediawikiapi to verify the article titles are in Wikipedia database to minimise the error.
* Using **HuggingFace** transformers, create a pipeline for sending user posts to transformer along with candidate topics and receive the title class the posts belong to along with the similarity scores to rank them.
* LLM used in this project, **facebook/bart-large-mnli (Zero-short classification).**
* Get top 20 topics for each user by aggregating the topics of all the posts of the user into a single array and using this array as candidate topics for the LLM and then create a dictionary with username as keys and the array of 20 topics as values.

**Aggregating dataset before training model:**

* **Input\_texts**: the list of 10 topics of each post
* **Output\_labels**: List of lists of usernames, where each inner list contains the 10 usernames associated with a particular input text.
* One hot encoding of output labels before training the logistic regression model

**NOTE!** Due to longer response time of **PaLM API** and **facebook/bart-mnli LLM** created a sample dataset with a list of 50 users and 250 topics.

DataAggregation of sample dataset and training of model provided in **SampleDataAggregation\_trainGenAImodel.ipynb** attached in zip file

**Model Training and Prediction:**

**VariationalAutoEncoder model**

* Define encoder, decoder and VariationalAutoEncoder functions
* Define hyperparameters vocab\_size, latent\_dim, etc..
* Extract latent representations
* Train logistic regression classifier with softmax for multi-label classification
* Predict probabilities for all usernames
* Get top 10 usernames with highest probabilities (across all labels)

**Post Topic Analysis:**

**Extract topic features from new posts**:

Preprocess the posts.

Applying the same text analysis techniques used for user interest modelling to identify the dominant topics of each new post.

**Model Testing:**

Giving topics of the testing data posts as input to model to predict the top 10 users relevant to it.

Evaluating the model by performing Jaccard similarity on predicted users and ground truths to determine the level of accuracy or correctness of the model.