Assignment 1: Text Clustering and Topic Modeling on GPT-5 Reddit Discussions

Course: Vizuara's Large Language Models Production and Deployment

Deadline: 25th August 2025

1 Objective

The goal of this assignment is to extract, cluster, and model topics from Reddit discussions related to GPT-5, specifically from the AMA with Sam Altman and others (link here). The main objective is to identify the key pain points users expressed after GPT-5 was launched.

2 Tasks Overview

You will:

1. Gather discussion data (threads and comments) from the given Reddit link.
2. Build a text clustering pipeline.
3. Perform topic modeling.
4. Synthesize findings in form of a PDF document to identify and summarize GPT-5 pain points.

3 Part 1: Data Collection

- Scrape all threads and comments from the provided Reddit link.
- Decide what chunks you will consider for the dataset, to finally identify the pain points.
- Store the data in a structured format (e.g., CSV, JSON).

4 Part 2: Text Clustering Pipeline

The pipeline must include the following steps:

4.1 Step 1: Embedding Generation

- Use any high-quality sentence embedding model from HuggingFace.
- You will experiment with at least two embedding models from the HuggingFace MTEB Leaderboard.

4.2 Step 2: Dimensionality Reduction (Compression Techniques)

Experiment with at least two techniques:

- **PCA (Principal Component Analysis)**
- **UMAP (Uniform Manifold Approximation and Projection)**

4.3 Step 3: Clustering Algorithms

Experiment with at least three clustering algorithms:

- **k-Means**
- **DBSCAN**
- **HDBSCAN**

4.4 Step 4: Modular Experimentation

You must try multiple combinations:

- PCA + k-Means
- PCA + DBSCAN
- PCA + HDBSCAN
- UMAP + k-Means
- UMAP + DBSCAN
- UMAP + HDBSCAN

5 Part 3: Topic Modeling

After clustering, perform topic modeling on each cluster to interpret what is being discussed.

5.1 Required Approaches

1. **BERTopic with TF-IDF**
   Use BERTopic with TF-IDF for initial topic extraction.
2. **Representational Model Re-ranking**
   Use an additional model to re-rank extracted topics for better coherence.
3. **Generative Model (Flan-T5)**
   Summarize and refine topics using Flan-T5.
4. **OpenAI API**
   Summarize and refine topics using OpenAI API
5. **GPT-OSS (Open Source)**
   Download and run GPT-OSS locally to perform topic modeling and compare with previous results.

5.2 Expected Output

- Final list of clusters with their top representative sentences.
- For each cluster: one or more topic labels.
- All results must be presented for all different combinations you have tested in Sections 4.1, 4.2, 4.3, 4.4, 5.1.

6 Part 4: Final Deliverables

- **Report** (PDF) containing:
  1. Description of data collection process.
  2. Methodology and experiments (all modular combinations for clustering and topic modeling).
  3. Comparative analysis of topic modeling methods.
  4. Identified pain points for GPT-5 users from the Reddit threads.
- **Code Notebook(s)** with all implementation details.
- **Data files** (processed and raw).

7 Submission Details

- Upload the final PDF file and your code to the course dashboard.
- Ensure your code is reproducible and clearly commented.
- **Deadline:** 25th August 2025, 11:59 PM.

8 Evaluation Criteria

- **Completeness** – All required steps and variations attempted.
- **Technical Correctness** – Proper implementation of methods.
- **Analysis Quality** – Depth of insights and clarity in identifying pain points.
- **Presentation** – Clarity, structure, and formatting of the report.