UE Lab RA Evaluation Project

Project Overview

You'll analyze Reddit posts that express opinions, frustrations, or rants about AI in different situations. The goal is to discover main topics people discuss and organize these into easy-to-understand categories.

Dataset Provided

A CSV file named `reddit_posts_2025-03-02_204006.csv`. This file contains two important columns:

- `title`: The title of the Reddit post.
- `text`: The main content of the post.

Methods and Techniques

You'll start with simple methods and move to more advanced methods, using Python and provided example code to help you.

Easy (Basic) Techniques

1. **Word Frequency Analysis**

    - Count and visualize the most common words in the posts.
    - Tools: Word clouds or simple bar graphs.

2. **TF-IDF + K-Means Clustering**

    - Use TF-IDF to represent text numerically.
    - Cluster similar posts together using K-Means.

3. **Word2Vec Embeddings + HDBSCAN**

    - Convert words into vectors using Word2Vec.
    - Group similar posts together using the HDBSCAN clustering method.

Advanced Techniques (Optional but encouraged!)

1. **BERTopic**

    - Use advanced language models to identify more accurate topics.

2. **Latent Dirichlet Allocation (LDA)**

    - A popular method to discover hidden topics in text.

How to Get Started

1. **Load and Explore Data**

    - Start by loading your data into Python with pandas.
    - Get familiar with the data by looking at a few rows first.

2. **Preprocess Data**

    - Clean your text data using provided code examples.
    - Tokenize the text into simpler pieces for analysis.

3. **Apply Basic Techniques**

    - First, perform a word frequency analysis using simple Python scripts.
    - Then, follow the provided examples to apply TF-IDF and K-Means clustering.
    - Next, use Word2Vec embeddings with HDBSCAN clustering using provided examples.

4. **Advanced Methods (Optional)**

    - Use the provided examples to apply BERTopic and LDA if you're comfortable.

Feel Free to Use GPT!

If you're stuck or unsure about anything, don't hesitate to use GPT (like ChatGPT) to help understand concepts, troubleshoot code, or interpret your results.

Deliverables (What You Need to Submit)

Submit the code and Create a clear summary for each method used. Your summary should include:

- Method name (e.g., Word Frequency Analysis, TF-IDF + K-Means)
- Topics you found (short descriptions or keywords)
- Example Reddit posts for each topic

Final Comparison and Analysis

Finish your project by writing a short comparison, including:

- What topics each method discovered
- How clear and useful these topics were
- Which methods you preferred and why

If you need any more help, just ask! Good luck!