Final Project Proposal, Fall 2025, 10/12/2025

Title: **Product Recommender System Using NLP**

Team Members: Laine Close, Marcos Fernandez, Owen Randolph

**Project Description:**

The purpose of this project is to build and experiment with a fully functioning product recommender system. Recommender systems can have powerful effects on how companies improve sales, offer customers better options and improve customer experience. It’s a common feature of the online shopping experience at present. The purpose of this is both to build a functional product recommender application, but also to demonstrate best practices when building a product recommender system using Natural Language processing and machine learning.

**Dataset and Scope:**

For our dataset, we aim to use ecommerce product review data, maybe an Amazon Review dataset for a niche product type (ex. Electronics, automotive, beauty, etc.) as these datasets tend to be quite large and we need to control the scope of our work. To maintain a manageable project scope, we will focus on a single product domain (ex. “Electronics Accessories”) within the Amazon Reviews dataset. This keeps data volume tractable while allowing us to evaluate model quality meaningfully. We plan to work with roughly 50k–100k reviews depending on availability after preprocessing. We aim to use state of the art NLP techniques and machine learning algorithms to build our system.

**Methodology:**

Phase 1: Data cleaning and preprocessing techniques to ready out dataset for experimentation.

Phase 2: Feature Engineering: Extract key linguistic features from product reviews and descriptions. Experiment with different frameworks: BERT, transformer-based encoders

Phase 3: Build hybrid recommendation system which uses: content-based filtering and collaborative based filtering, matrix factorization or neural collaborative filtering, user-item and/or item-item interaction.

Phase 4: We will compare classical NLP representations (TF-IDF, Word2Vec) against transformer-based embeddings (BERT, Sentence-BERT) to analyze the trade-offs in performance, interpretability, and computation time. Simulate and test each system and record metrics for performance (precision, recall, F1, etc. Tuning models to improve performance.

Phase 5: Create an interactive Streamlit application with input of product preferences

**Example datasets:** [Amazon Reviews'23](https://amazon-reviews-2023.github.io/); <https://amazon-reviews-2023.github.io/>,

https://shuaizhang.tech/posts/2019/08/blog-post-3/

https://cseweb.ucsd.edu/~jmcauley/datasets.html