**SI670 Final Project Proposal**

**Motivation:** Accurately predicting query–product relevance is essential for improving search experiences. Traditional methods like BM25 rely on keyword matching and miss semantic meaning. With eBay apparel dataset with graded relevance labels, we train a model to predict relevance by combining text-based and simple product features to build an interpretable and efficient model.

**Methods:** We formulate a regression problem, which we predict a continuous relevance score between each query–item pair . We use TF-IDF vectorization of concatenated text (question + title + contents), and Random Forest Regressor (baseline) and Linear Regression for comparison. The training objective is minimizing RMSE between predicted and true relevance labels. The base comparison will beBM25 relevance_score from the hybrid retriever.

**Datasets:** We have 20,000 clothing listings from the eBay Browse API, retrieved from Browse API resources | eBay Developers Program. Each record contains metadata including title, price, seller rating, and description text. We used a retrieval system to produce a dataset of query–item pairs. For each query, candidate items were assigned a relevance level (1–5) based on ranking position and ensemble scoring. This file includes question, title, contents, relevance_score , relevance_label.

**Evaluation:** We will use the eBay apparel relevance dataset to test our model's ability to predict and rank
**Experiments**: Compare the Random Forest Regressor with the BM25 baseline; Examine the effect of extra features; Test other lightweight models for robustness.
**Metrics:** Regression: RMSE and $R^2$ for prediction accuracy;
**Ranking:** nDCG@5 and MRR for ranking quality. Baseline: Improvement over BM25 relevance scores.
**Success Criteria:** Achieving lower RMSE and higher $R^2$, nDCG@5, and MRR than BM25, proving the model offers more accurate and semantically informed relevance prediction while remaining lightweight and efficient.

**Computing**: We will use some amount of computational resources since the dataset is quite large. The main workloads include TF-IDF vectorization. Even though sparsing text representations may increase memory usage, but can be handled efficiently using scikit-learn. Random Forest and Linear Regression are considered manageable. If necessary, experiments will be run on Great Lakes or Google Colab for faster execution.

**Existing Work:** Nogueira & Cho (2019) on passage re-ranking with BERT, highlight this trade-off between performance and efficiency. eBay's AI blog (2019) and Amazon Search papers (Ai et al., 2021) have also discussed balancing retrieval accuracy with latency constraints in production search systems. Our work bridges these two extremes by proposing an interpretable regression model that integrates semantic features. This approach will make it simple enough for transparent analysis and quick deployment.

**Duty of each group member:**
**Xinyi Huang:** Responsible for **data preparation and feature extraction**, including cleaning the eBay apparel dataset, implementing TF-IDF vectorization, and constructing additional features .
**Jiaxuan Xu:** In charge of **model training and hyperparameter tuning**, building and optimizing the Random Forest Regressor and alternative baseline models.
**Merrila Liu:** Handles **evaluation and report writing**, assessing model performance using regression and ranking metrics, comparing with the BM25 baseline, and compiling the final project report.