COMP9727: Recommender Systems

Project Design: E-commerce Product Recommender System

Name: [Xiaoyue Hou]
zID: [z5540746]

Timeline and Milestones

1. **Data Exploration & Preprocessing** (Week 6): Sample the Amazon dataset, perform exploratory data analysis (EDA), filter interactions, implement feature engineering, construct user/item profiles using TF-IDF
2. **Baseline Implementation** (Week 7): Implement baseline collaborative filtering models, including item-based collaborative filtering and matrix factorization using singular value decomposition
3. **Content-Based Method** (Week 8): Develop content-based recommendation module using TF-IDF similarity between user and item profiles
4. **Hybrid Integration** (Week 9): Integrate the hybrid model, apply temporal decay weighting and diversity-aware reranking to refine recommendations
5. **Evaluation** (Week10):Conduct offline evaluation, simulate user study (A/B testing), and finalize the written report and presentation

1. Scope

Domain and Target Users

The recommender system will operate in the e-commerce product recommendation domain, specifically focusing on Amazon's marketplace ecosystem. The system targets:

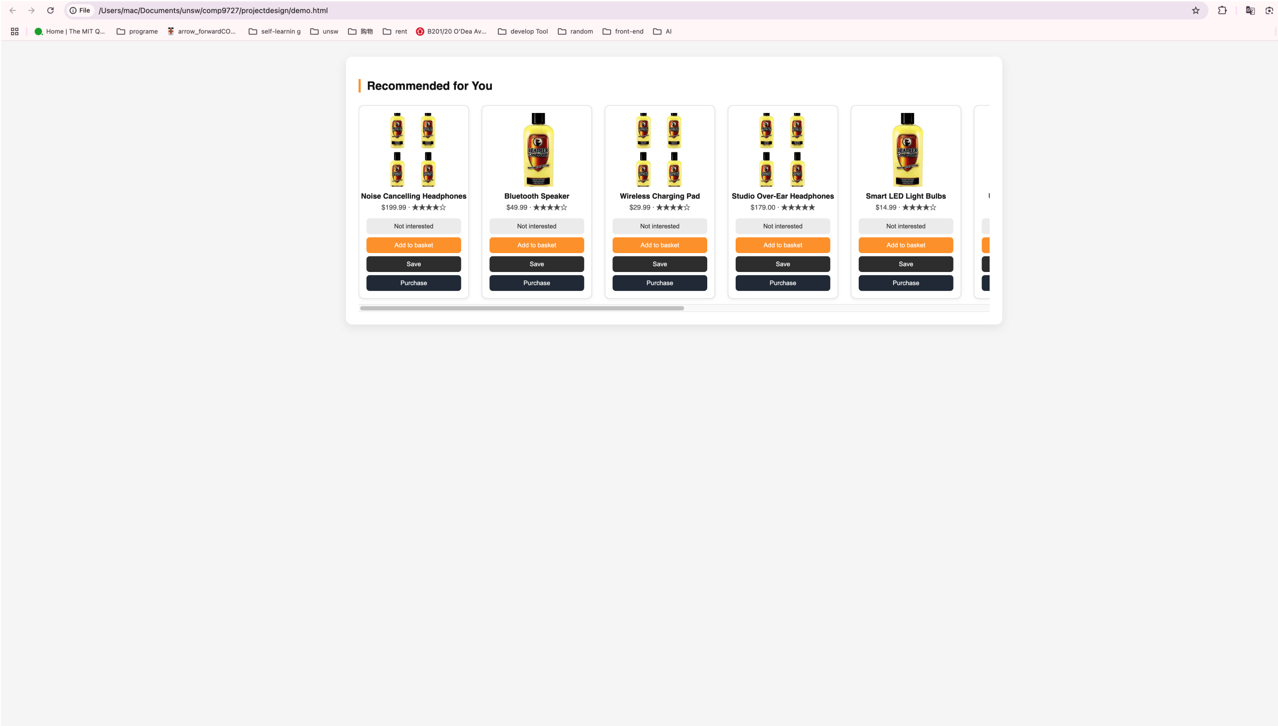
- **Primary Users**: Individual online shoppers who seek personalized product recommendations
- **Use Case**: For the user, the recommender system is to improve users' shopping experience by presenting personalized and relevant products, reducing decision effort and increasing the likelihood of engagement. For the platform, the system is to drive conversions, promote diverse product discovery, and support business goals such as sales and user retention.

User Interface and Interaction Design

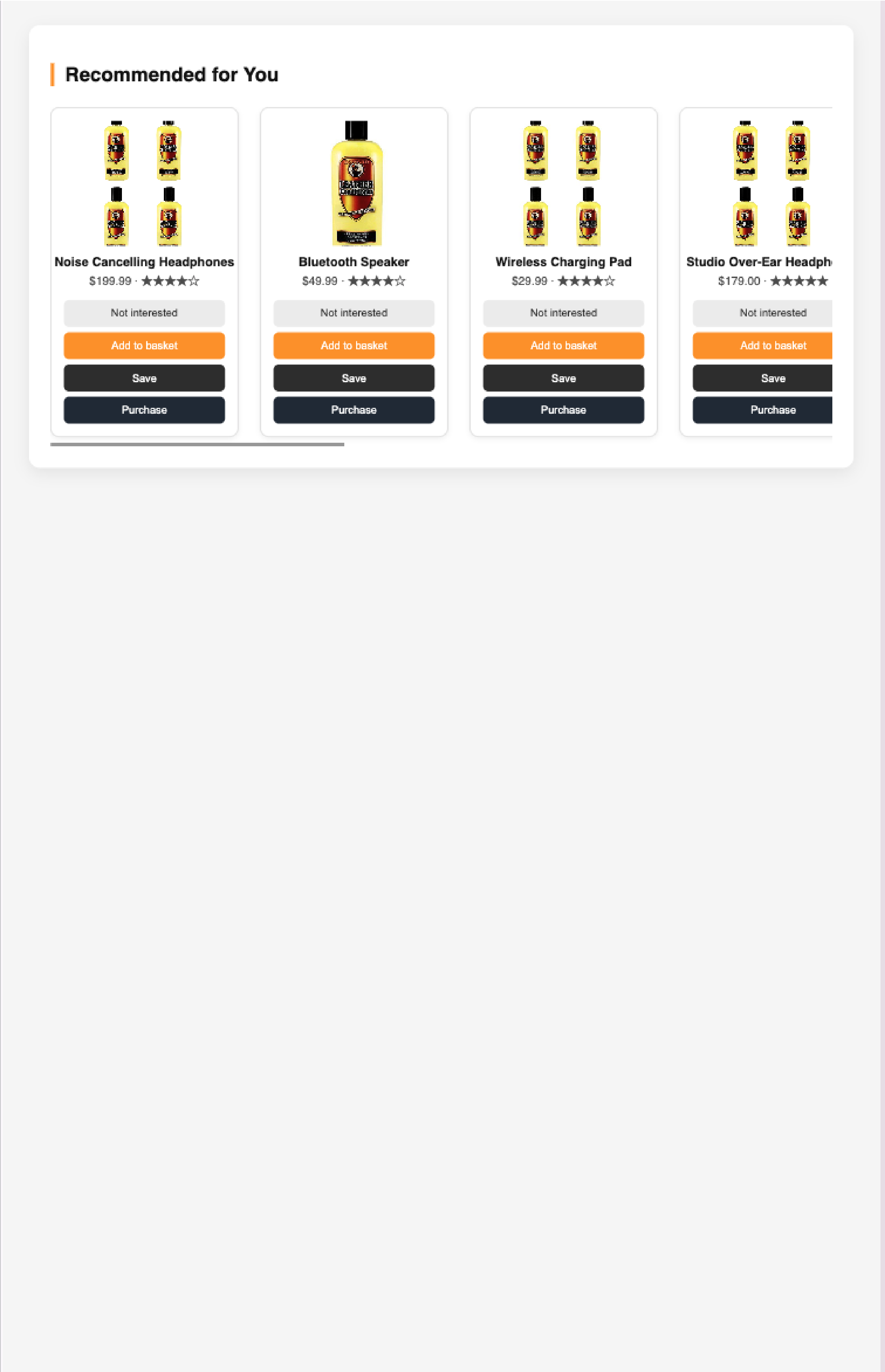
Display Strategy: Present **10 product recommendations** per page through a responsive web interface (desktop and mobile). Products shown in a single row on both interface.

Interface Mockup Description:

- desktop



- mobile


Recommended for You
Noise Cancelling Headphones
$199.99 · ★★★★☆
Not interested
Add to basket
Save
Purchase
Bluetooth Speaker
$49.99 · ★★★★☆
Not interested
Add to basket
Save
Purchase
Wireless Charging Pad
$29.99 · ★★★★☆
Not interested
Add to basket
Save
Purchase
Studio Over-Ear Headph
$179.00 · ★★★★★
Not interested
Add to basket
Save
Purchase

Demo

User Interaction Flow:

1. User views recommendations .
2. Click\Ignore\Save\Not intereste\add to Basket\Purchase.
3. Records these interection information.
4. Next page loads with updated recommendations.

Feedback Collection:

- **Explicit**: Save,Not intereste, Ratings, Purchase, Add to basket action.
- **Implicit**: Click-through, Time spent viewing.
- **Feedback Integration**: Transfer these metrics into the same data structure as the original relevant dataset structure and immediate weight adjustment in recommendation algorithm.

Expected Challenges and Mitigation:

- **Data Size**: Address through strategic sampling such as Stratified Sampling and System Sampling while preserving data characteristics.
- **Cold Start**: Content-based component specifically handles this scenario.
- **Evaluation Complexity**: Start with standard and simple metrics, then add more metrics step by step to evaluate recommender system comprehensively, and add diversity measure.

Business

Revenue:

- **Conversion Optimization**: Personalized recommendations increase purchase likelihood by digging the unawared demand of the users.
- **Cross-selling**: Suggest complementary products to increase order value
- **Customer Retention**: Improved satisfaction through relevant recommendations leads to repeat purchases and the loyalty of users.

2. Datasets

Primary Dataset: Amazon Reviews 2023

Source: McAuley Lab Amazon Reviews 2023 dataset from Hugging Face
URL: https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

The example of dataset:

- **User Reviews**

- {'rating': 5.0,
'title': 'Such a lovely scent but not overpowering.',
'text': "This spray is really nice. It smells really good, goes on really fine, and does the trick. I will say it feels like you need a lot of it though to get the texture I want. I have a lot of hair, medium thickness. I am comparing to other brands with yucky chemicals so I'm gonna stick with this. Try it!",
'images': [],
'asin': 'B00YQ6X8EO',
'parent_asin': 'B00YQ6X8EO',
'user_id': 'AGKHLEW2SOWHNMFQIJGBECAF7INQ',
'timestamp': 1588687728923,
'helpful_vote': 0,
'verified_purchase': True}

- **Item Metadata**

- {'main_category': 'All Beauty',
'title': 'Howard LC0008 Leather Conditioner, 8-Ounce (4-Pack)',
'average_rating': 4.8,
'rating_number': 10,
'features': [],
'description': [],
'price': 'None',
'images': {'hi_res': [None, 'https://m.media-amazon.com/images/I/71i77AuI9xL.*SL1500*.jpg'],
'large': ['https://m.media-amazon.com/images/I/41qfjSfqNyL.jpg',
'https://m.media-amazon.com/images/I/41w2yznfuZL.jpg'],
'thumb': ['https://m.media-amazon.com/images/I/41qfjSfqNyL.*SS40*.jpg',
'https://m.media-amazon.com/images/I/41w2yznfuZL.*SS40*.jpg'],
'variant': ['MAIN', 'PT01']},
'videos': {'title': [], 'url': [], 'user_id': []},
'store': 'Howard Products',
'categories': [],
'details': '{"Package Dimensions": "7.1 x 5.5 x 3 inches; 2.38 Pounds", "UPC": "617390882781"}',
'parent_asin': 'B01CUPMQZE',
'bought_together': None,
'subtitle': None,
'author': None}

- **Links:** (user-item / bought together graphs).

Dataset information(sufficient quality and quantity):

- **Larger Dataset**: Collected 571.54M reviews.
- **Newer Interactions**: Current interactions range from May. 1996 to Sep. 2023, sufficient scale for robust collaborative filtering
- **Richer Metadata**: More descriptive features in item metadata, and include 33 categories which enabling the diverse recommendation scenarios.
- **Fine-grained Timestamp**: Interaction timestamp at the second, enabling temporal recommendation modeling
- **Cleaner Processing**: Cleaner item metadata than previous versions;
- **Standard Splitting**: Standard data splits to encourage RecSys benchmarking.
- **Rich user behavior data** including verified purchases, helpfulness votes and rating.

Dataset Process

- **Data size:** Because this dataset is very large, this project will handle data by addressing through strategic sampling such as Stratified Sampling and System Sampling while preserving data characteristics.
- Filter for users/items with more than 5 interactions to ensure collaborative filtering stability and reduce the sparcity of the metrxes.
- Handle missing metadata through category-based imputation.
- 80/10/10 train/validation/test split with temporal ordering preserved.

Review Data Fields:

- `user_id` + `parent_asin` + `rating`: Core user-item-rating matrix for collaborative filtering
- `timestamp`: Temporal decay weighting (recent reviews weighted higher)
- `text`: Review sentiment analysis for implicit preference strength
- `verified_purchase`: Quality filtering (verified purchases weighted 2x)
- `helpful_vote`: Social proof indicator for review quality assessment

Metadata Fields:

- `title` + `description` + `features`: TF-IDF vectors for content-based similarity
- `main_category` + `categories`: Hierarchical clustering for cold-start recommendations
- `price`: Price-range filtering and value-based recommendations
- `store`: Brand loyalty modeling and seller reputation
- `bought_together`: Explicit item-item associations for cross-selling

Dataset Limitations and Mitigation

Limitations:

1. **Temporal Concentration**: Data from 2016 to 2023 occupies the large majority of the dataset, the class high-quality products could be ignored.
    - *Mitigation*: Implement temporal decay weighting which will handle with the problem of ignoring old products, and separatly model products in 2023 of reviews.

2. **Review Selection Bias**: Users more likely to review extreme rating which could not align with their review text.
   - *Mitigation*: Apply sentiment analysis of review texts to normalize rating distributions.
3. **Subset Bias**: Dataset only concentrates on reviewed products, and ignores unreviewed purchases actions.
   - *Mitigation*: Introduce category-based exploration to find the unreviewed products that interest user.
4. **Competition-Oriented Processing**: Concentrate more on prediction tasks but less on diversity
   - *Mitigation*: Explicitly implement diversity metrics, avoid overfitting on rating prediction accuracy
5. **US Market Focus**: Geographic limitation reduces global applicability
   - *Mitigation*: Explicitly scope system for US market especially for Amazon.com. Because of the different product taste for different regions, we should only use this recommender system in US market.

3. Methods

Recommendation Method Overview

Method 1: Collaborative Filtering with Matrix Factorization (Baseline)

Implementation: Implicit feedback matrix factorization using Alternating Least Squares (ALS) **Rationale**:

- Large-scale user-item interactions make collaborative filtering the natural starting point.
- Matrix factorization handles sparse data effectively .
- Well-established baseline with clear evaluation metrics are very simple and explianable. **data implementation**: 50-dimensional user/item embeddings, confidence weighting based on rating frequency

Method 2: Content-Based Filtering with Text Embeddings

Implementation: Product similarity using TF-IDF vectors from metadata **Rationale**:

- Rich textual features (titles, descriptions) in Amazon dataset enable semantic matching
- Addresses cold-start problem for new products lacking interaction history
- Provides interpretable recommendations through feature similarity **data implementation**:
- TF-IDF vectors from product titles + descriptions (5000 features)
- Category and price tier as additional features
- Cosine similarity for product matching

Method 3: Hybrid Linear Combination Model

Implementation: Weighted combination of collaborative and content-based scores **Rationale**:

- Combines strengths of both approaches while remaining implementable in course timeframe
- Simple linear weighting allows clear ablation studies
- Provides flexibility to handle different user scenarios (new vs returning users) **data implementation**:
- CF score × 0.6 + Content score × 0.4 for regular users
- CF score × 0.3 + Content score × 0.7 for users with <10 interactions
- Grid search for optimal weights using validation set

Hybrid Method

Weighted Linear Combination:

- Collaborative Filtering: 45% weight (primary preference modeling)
- Content-Based: 35% weight (similarity and cold-start)
- Temporal Model: 20% weight (trend awareness)

Dynamic Weight Adjustment:

- New users (≤5 interactions): Content-based 65%, CF 25%, Temporal 10%
- Power users (≥100 interactions): CF 55%, Content 25%, Temporal 20%
- Seasonal periods: Increase temporal weight to 30%

Justification for Method Selection:

- **Dataset Compatibility**: Each method leverages different aspects of the Amazon dataset
- **Complementary Coverage**: CF handles user preferences, content handles new items, temporal handles trends
- **Scalability**: All methods can handle millions of users/items with proper implementation
- **Evaluation Feasibility**: Clear baselines and incremental improvements enable systematic evaluation be clear and easy.

4. Evaluation

Evaluation Overview

Note:high prediction accuracy doesn't always translate to user satisfaction.

Model Performance Metrics

Primary Accuracy Metrics

- **Precision@10 & Recall@10** (40% weight): Core recommendation accuracy metrics. The former avoid irrelevant recommendations and reduce unwanted recommendations, the later avoid missing wanted products and discover potential purchase opportunities. They are the most important metrics in this recommender system.
- **NDCG@10** (30% weight): Position-aware ranking quality what is whether the best products are ranked at the top. It improve user experience, increase click-through rates.
- **Mean Absolute Error (MAE)** (20% weight): Rating prediction accuracy. The smaller the better, it measures personalization level and user trust.
- **Coverage** (10% weight): Percentage of items that appear in any recommendation. Avoid system from recommending only popular products, which leads to long-tail product exposure. It avoids the Matthew effect

Method Comparison Framework

Baseline Comparison: Each method evaluated against random and popularity-based baselines **Ablation Study**: Hybrid model with different weighting schemes (CF vs Content emphasis) **Statistical Significance**: Paired t-tests between methods on per-user metrics

Evaluation Protocol:

- **Leave-one-out**: For each user, hide most recent purchase for testing
- **Cross-validation**: 5-fold validation on user level to ensure robustness
- **Cold-start analysis**: Separate evaluation for users with ≤5 historical interactions

System Effectiveness Metrics (User-Centric Evaluation)

Behavioral Metrics

- **Click-Through Rate (CTR)**: Percentage of recommendations clicked
- **Conversion Rate**: Clicks leading to purchases
- **Session Engagement**: Time spent interacting with recommendations
- **Return Usage**: Users returning to use recommendation system

User Experience Metrics

- **Recommendation Relevance Score**: User-reported appropriateness (5-point scale)
- **Diversity Satisfaction**: "Recommendations show good variety" (5-point scale)
- **System Trust**: "I trust these recommendations" (5-point scale)

- **Overall Satisfaction**: Net Promoter Score for recommendation system

Trade-off Management: If accuracy conflicts with diversity, prioritize user engagement metrics over pure prediction accuracy. A slightly less accurate but more diverse system often performs better in real deployment.

Computational Requirements and Real-Time Constraints

Performance Requirements:

- **Response Time**: Recommendations delivered within 150ms of request
- **Model Training**: Batch updates feasible within 2-hour overnight window
- **Memory Usage**: Embedding storage optimized for sub-second retrieval

Update Strategy:

- **Real-time**: Immediate preference updates based on clicks/purchases
- **Batch Processing**: Full model retraining weekly with accumulated feedback
- **Trending Integration**: Hourly popularity updates for new items

User Study Design

Study Framework

Participants: 20-25 peoples with online shopping experience **Setting**: Laptop-based interface showing recommendation results

Experimental Design

A/B Test:

- **Condition A**: Baseline collaborative filtering recommendations
- **Condition B**: Hybrid system recommendations
- **Condition C**: Content-based recommendations (for comparison)

User Tasks:

1. **Evaluation Task**: Rate relevance of 10 recommendations (1-5 scale)
2. **Preference Task**: "Which system would you prefer for real shopping?"
3. **Discovery Task**: "Did you see any interesting products you hadn't considered?"

Data Collection Protocol

Quantitative Measures:

- Relevance ratings for each recommended item
- Task completion time
- System preference choice

Practical Considerations:

- Use mock interface with actual recommendation outputs
- Anonymous participation, voluntary basis
- Results used solely for course evaluation and learning