Product matching

Presented by Zeyad Tarek Ahmed

Problem Statement:

Isupply platform faces challenges in matching products from different sellers due to variations in naming and pricing.

Objective:

Develop an automated method to match products using a combination of text similarity and price comparison.

Enhance User Experience by quickly matching

Faster Data review with high accuracy

Data Collection & Preprocessing

- Master Product List (Official Product Names & Prices)
- Seller Dataset (Seller's Product Names & Prices)

Data Cleaning Techniques:

- Standardizing Arabic text
- Removing unwanted words (e.g., discounts, promotions)
- Handling missing values & formatting numbers

Matching Methodology:

As we found that the price is more unique due to dosage interference in text so we decided to make a grid search to get the perfect distribution of weights so,

- Text Similarity (25%)
  - TF-IDF vectorization (Character n-grams: 2-4)
  - Cosine Similarity for text matching
- Price Similarity (75%)
  - Formula: 1 - abs(price1 - price2) / max(price1, price2)
  - Normalization to ensure fair contribution
- Final Score Calculation:
  - Combined Score = (0.25* Text Similarity) + (0.75* Price Similarity)

Implementation & Workflow

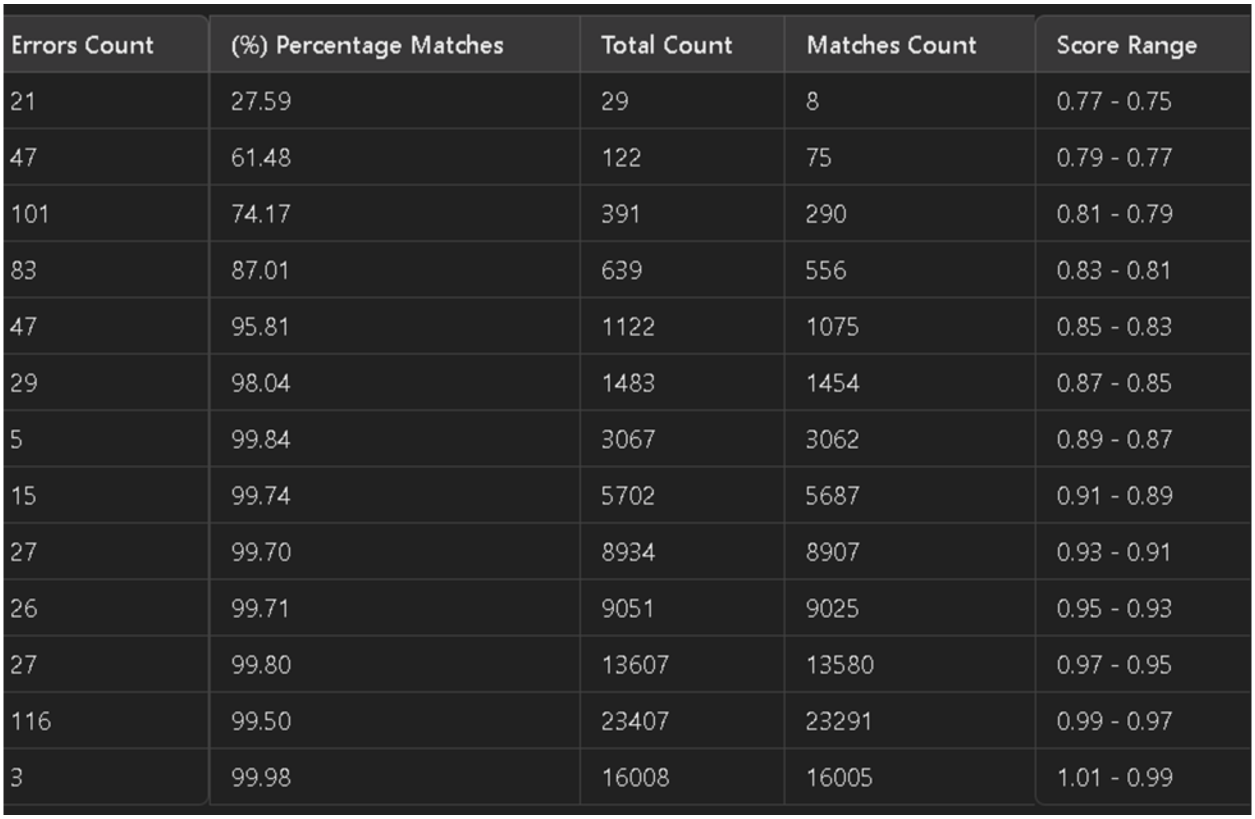
- Pipeline Overview:
  - Load & preprocess data
  - Compute text similarity (TF-IDF + Cosine Similarity)
  - Compute price similarity

- Generate final matching scores
- Filter & evaluate matches
- Extract a matched data file

Performance Analysis

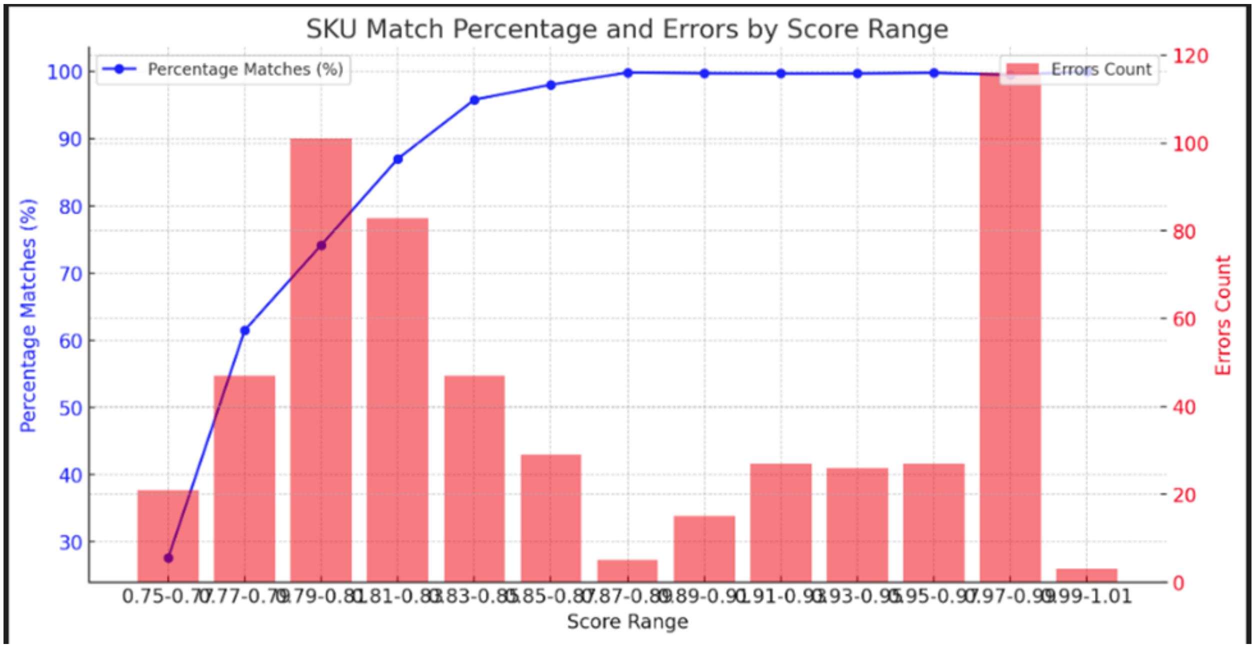
The model has a 99.3% a total accuracy labeled by the true sku

The distribution of correct matches according to sku and scores is this error counts at the worst cases some skus where wrong in the dataset



Errors Count	(%) Percentage Matches	Total Count	Matches Count	Score Range
21	27.59	29	8	0.77 - 0.75
47	61.48	122	75	0.79 - 0.77
101	74.17	391	290	0.81 - 0.79
83	87.01	639	556	0.83 - 0.81
47	95.81	1122	1075	0.85 - 0.83
29	98.04	1483	1454	0.87 - 0.85
5	99.84	3067	3062	0.89 - 0.87
15	99.74	5702	5687	0.91 - 0.89
27	99.70	8934	8907	0.93 - 0.91
26	99.71	9051	9025	0.95 - 0.93
27	99.80	13607	13580	0.97 - 0.95
116	99.50	23407	23291	0.99 - 0.97
3	99.98	16008	16005	1.01 - 0.99

And there is graph representation



Challenges & Solutions

- **Challenges:**
  - Variations in product naming conventions
  - Handling promotions & misleading text
  - Different currency formats & price fluctuations
  - Algorithms mixing between embedded prices and dosages
  - Make a good confidence score
- **Solutions Implemented:**
  - Enhanced text preprocessing
  - Adjusted weights for similarity calculation
  - Used SKU-based ground truth for validation

Conclusion & Future Work

- Conclusion:
    - Effective method for automated product matching
    - Balanced approach using text and price similarity
    - Need for further fine-tuning on datasets
- Future Work:
    - Incorporating deep learning (BERT for Arabic text)
    - Improving handling of synonyms & abbreviations
    - Expanding dataset for better generalization

Thank you!

-
    - LinkedIn: [https://www.linkedin.com/in/zeyad-el-bagouri-90499524b/?lipi=urn%3Ali%3Apage%3Ad_flagship3_profile_view_base%3B4Ng2ReJ1SliB62%2BQBWs6Jg%3D%3D]
    - Email: [tarekzeyad858@gmail.com]
    - GitHub: [ZeyadTarekAhmed (Zeyad EL-Bagouri)]