Albert Jané Lardiés - U215114
Marc de los Aires Tello - U198732
Jordi Esteve Claramunt - U215108
https://github.com/jordiestevee/irwa-search-engine-g_019
TAG: IRWA-2025-part-1

IRWA Project - Part 1 Report

PART 1: Data Preparation

Overview

Preparing and cleaning the "Fashion Products Dataset" for subsequent information retrieval and analysis was the aim of this initial phase. The JSON-formatted product records that made up the raw dataset included text fields (like title and description), numeric fields (like price and rating), and categorical attributes (like brand and category).

We conducted structured preprocessing to standardize text, eliminate noise, and guarantee uniform formats across all fields before utilizing this data for modeling or search.

Data Loading

We created a custom data loader that can read newline-delimited JSON (NDJSON) files and standard JSON arrays. Compressed.gz files can also be supported. This made it possible for the preprocessing pipeline to manage big datasets effectively. After being transformed into a Python dictionary, each product record was saved in memory for later use.

Text Preprocessing

For the main textual fields — title, description, and product_details — we applied the following NLP preprocessing steps using the NLTK library:

Step	Description
Tokenization	Split text into individual tokens (words).
Lowercasing	Standardize capitalization to reduce redundancy.
Removing punctuation	Keep only alphanumeric tokens.
Removing stop words	Eliminate common, non-informative words such as “the”, “and”, “is”.
Stemming	Reduce words to their root forms using the Porter Stemmer (e.g., “running” → “run”).

Removing very short tokens	Discard tokens with length ≤ 2, which are typically not meaningful.

These steps were combined in a preprocessing function that produced a cleaned text string for each document, stored in a new field called “tokens”.

Additionally, the `product_details` field was converted from structured JSON (key–value pairs) into plain text, allowing descriptive attributes like “Color: Red” or “Material: Cotton” to contribute to the searchable text.

Field Handling Decisions

To ensure consistency with the assignment requirements, we preserved the following fields for each document:

```
pid, title, description, brand, category, sub_category,
product_details, seller, out_of_stock, selling_price, discount,
actual_price, average_rating, url, tokens
```

Categorical fields

- We decided to keep brand, category, sub-category, and seller as separate fields, instead of merging them into the main text.
- `product_details` was merged into the main text because it frequently contains descriptive, semantically useful information.
- Merging all fields may increase recall but introduce noise; keeping them separate allows for more precise weighting in retrieval models.

Numeric fields

- A helper function was used to convert numeric attributes (such as selling_price, discount, and average_rating) to numerical values.
- These fields were not indexed as text because they are more suitable for numerical filtering, sorting, or range-based ranking (e.g., price < 100, rating > 4).

PID handling

- The `pid` field was retained as a unique identifier for each product, required for later evaluation.

Output

After processing, all records were combined into a **pandas DataFrame**, with the product ID (`pid`) set as the index. The resulting dataset was exported to a CSV file for reuse in later phases (e.g., indexing, query retrieval).

PART 2: Exploratory Data Analysis

Overview

By examining the dataset's structure, contents, and underlying patterns, this section aims to provide a deeper understanding of the data. Later phases of the information retrieval system are built upon this exploratory phase, which guarantees that the features of the dataset inform data preprocessing and indexing strategies.

Overview of the Dataset

After preprocessing, the dataset contains the following main fields for each product:

- **Text fields:** `title`, `description`, `tokens` (preprocessed text).
- **Categorical fields:** `brand`, `category`, `sub_category`, `seller`.
- **Numeric fields:** `selling_price`, `actual_price`, `discount`, `average_rating`, `out_of_stock`.
- **Unique identifier:** `pid`.

pid	title	description	brand	category	sub_category	product_details	seller	out_of_stock	selling_price	discount	actual_price	average_rating	url	tokens
TKPFCZ9EA7H5FYZH	Solid Women Multicolor Track Pants	Yorker trackpants made from 100% rich combed c...	York	Clothing and Accessories	Bottomwear	[{'Style Code': '1005COMBO2'}, {'Closure': 'El...	Shyam Enterprises	0.0	921.0	69.0	2999.0	3.9	https://www.flipkart.com/yorker-solid-men-mult...	solid women multicolor track pant yorker track...
TKPFCZ9EJZV2UVRZ	Solid Men Blue Track Pants	Yorker trackpants made from 100% rich combed c...	York	Clothing and Accessories	Bottomwear	[{'Style Code': '1005BLUE'}, {'Closure': 'Draw...	Shyam Enterprises	0.0	499.0	66.0	1499.0	3.9	https://www.flipkart.com/yorker-solid-men-blue...	solid men blue track pant yorker trackpant mad...
TKPFCZ9EHFCY6Z4Y	Solid Men Multicolor Track Pants	Yorker trackpants made from 100% rich combed c...	York	Clothing and Accessories	Bottomwear	[{'Style Code': '1005COMBO4'}, {'Closure': 'El...	Shyam Enterprises	0.0	931.0	68.0	2999.0	3.9	https://www.flipkart.com/yorker-solid-men-mult...	solid men multicolor track pant yorker trackpa...
TKPFCZ9ESZZ7YWEF	Solid Women Multicolor Track Pants	Yorker trackpants made from 100% rich combed c...	York	Clothing and Accessories	Bottomwear	[{'Style Code': '1005COMBO3'}, {'Closure': 'El...	Shyam Enterprises	0.0	911.0	69.0	2999.0	3.9	https://www.flipkart.com/yorker-solid-men-mult...	solid women multicolor track pant yorker track...

Table 1: *First four rows of the processed dataset.*

Table 1 shows a preview of the processed dataset, highlighting the main fields preserved for each product and the new `tokens` field containing cleaned text.

Textual Analysis

Product Text Length Distribution

We calculated the number of tokens per product using the preprocessed `tokens` field to understand text density.

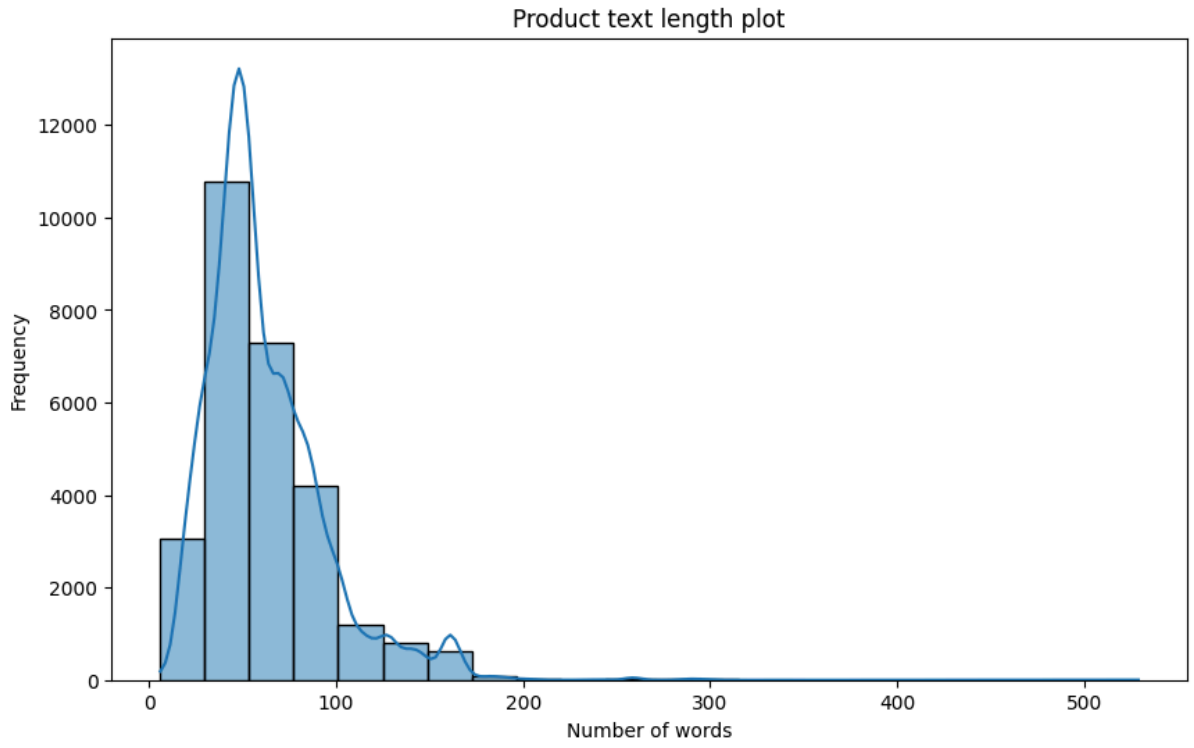


Figure 1: *Histogram of product text lengths.*

The histogram shows the distribution of the number of words per product. Most products have between *20* and *100* tokens, indicating concise, structured descriptions. A few outliers have longer texts, representing products with more detailed descriptions.

Vocabulary Size and Frequent Words

We compiled all tokens from the dataset to calculate:

- Vocabulary size: the number of unique tokens.
- Top 5 most frequent words.

```
Product vocabulary size is 15877

The top 5 most appearing words are:
('fabric', 57213)
('neck', 56553)
('sleev', 50876)
('fit', 39523)
('type', 38701)
```

Table 2: *Top 5 most frequent tokens and their counts.*

The vocabulary contains 15,877 unique tokens. The most commonly used words are `fabric`, `neck`, `sleeve`, `fit`, and `type`. These findings suggest a focus on product attributes such as clothing material, style, and fit.

Word Cloud of Product Tokens

We created a word cloud to visually represent the dataset's most frequently used tokens.

Figure 2: *Word cloud of product tokens.*
Insert the output from the WordCloud cell.

The most common terms in product descriptions are highlighted in the word cloud (Figure 2). Higher frequency is correlated with larger words. The most common material and style descriptors, which support the clothing domain focus, are `fabric`, `cotton`, and `sleeve`.

Numeric Field Analysis

Product Ratings

To determine which products had the highest ratings, we looked at the average rating of each item.

The following are the Top 5 Highest-Rated Products:

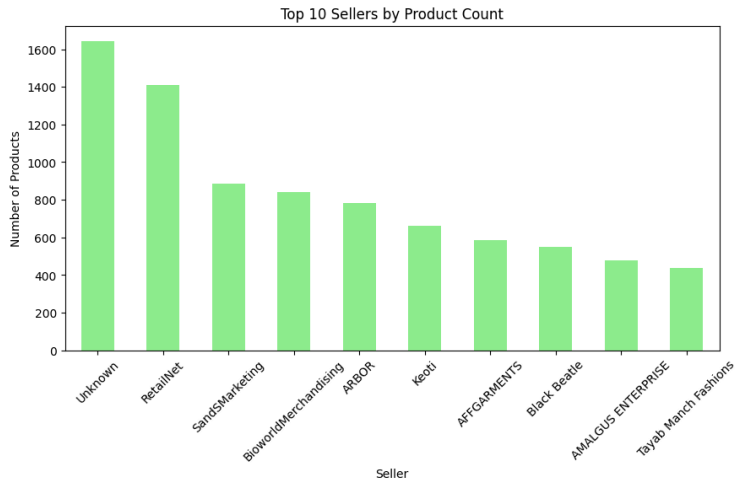
	title	description	brand	category	sub_category	product_details	seller	out_of_stock	selling_price	discount	actual_price	average_rating	tokens
0	Solid Women Round Neck Blue T-Shirt		Oka	Clothing and Accessories	Topwear	[{'Type': 'Round Neck'}, {'Sleeve': 'Half Slee...	OKANE	0.0	385.0	44.0	699.0	5.0	solid women round neck blue type round neck sl...
1	Printed Women Hooded Neck Black T-Shirt		ATTIITU	Clothing and Accessories	Topwear	[{'Type': 'Hooded Neck'}, {'Sleeve': 'Full Sle...	ATTIITUDE	0.0	549.0	60.0	1399.0	5.0	print women hood neck black type hood neck sle...
2	Printed Women Hooded Neck Grey T-Shirt		ATTIITU	Clothing and Accessories	Topwear	[{'Type': 'Hooded Neck'}, {'Sleeve': 'Full Sle...	Assiduus Distribution	0.0	909.0	35.0	1399.0	5.0	print women hood neck grey type hood neck slee...
3	Graphic Print Men Round Neck Blue T-Shirt		Free Authori	Clothing and Accessories	Topwear	[{'Type': 'Round Neck'}, {'Sleeve': 'Half Slee...	BioworldMerchandising	0.0	519.0	35.0	799.0	5.0	graphic print men round neck blue type round n...
4	Solid Women Round Neck White, Black T-Shirt	Loosen up in this perfectlloose-fit black vest...	ATTIITU	Clothing and Accessories	Topwear	[{'Type': 'Round Neck'}, {'Sleeve': 'Sleeveles...	Assiduus Distribution	0.0	649.0	35.0	999.0	5.0	solid women round neck white black loosen blac...

Table 3: *Top 5 highest-rated products.*

Table 3 shows the top 5 highest-rated products. All of the highest rated products belong to the clothing and accessories category and topwear subcategory. ANTITU is the brand that appears the most in the top(3 times).

Categorical Field Analysis

We looked at the top ten sellers and brands. This analysis sheds light on the possible bias and composition of the dataset. Since many products had blank or missing brand entries, the first bar was left unlabelled; these were changed to "Unknown" to guarantee that all brands were shown in the visualisation consistently. In order to maintain data completeness and avoid bias brought on by data loss, we chose to keep these records under the single label "Unknown" rather than delete them.



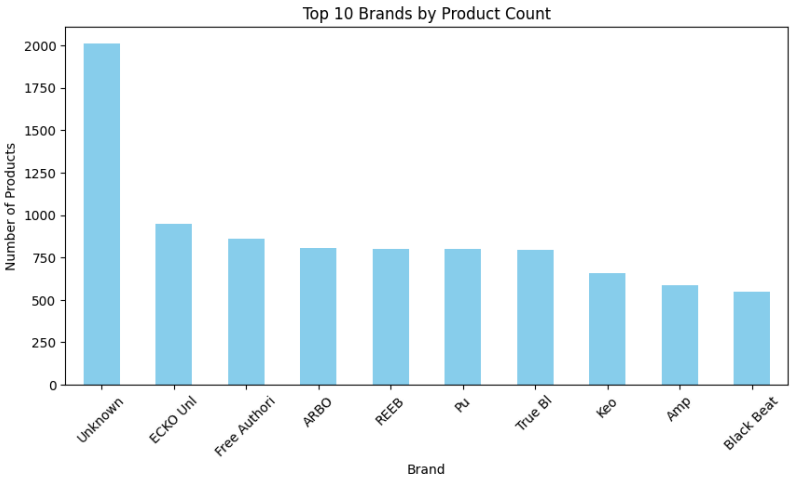


Figure 2 and 3: *Bar charts of top 10 brands AND sellers.*

These visualizations help identify the concentration of listings across different sellers and brands. The prevalence of “Unknown” entries in both charts points to incomplete data, which should be considered. Additionally, the skewed distributions may impact retrieval bias, where popular sellers or brands dominate search results.