**Dataset Information for Product Recommender System using NLP:**

The table in your screenshot shows dataset stats from Amazon Review Data.

McAuley Lab. *Amazon Reviews 2023*. Retrieved from [https://amazon-reviews-2023.github.io/](https://amazon-reviews-2023.github.io/?utm_source=chatgpt.com); https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

Each “category” contains two JSON files. For appliances, these are the files:

| **Type** | **File Name** | **What It Contains** | **Size (approx)** |
| --- | --- | --- | --- |
| **Review data** | Appliances.json.gz | 2 million+ User reviews (ratings, text, timestamps, user\_id, asin) | 258 MB |
| **Metadata** | meta\_Appliances.json.gz | 93,000+ Product info (titles, categories, brands, prices, descriptions) | 63.3 MB |

Json file features example from user review dataset (Appliances.json.gz):

| **Field** | **Description** | **Example / Interpretation** |
| --- | --- | --- |
| "rating" | The star rating (1.0–5.0) | 5.0 → extremely satisfied |
| "title" | Short summary or headline of the review | "The hardest part is pulling(disconnecting & reconnecting) dishwasher" |
| "text" | The main body of the review | "Love this because it's the complete assembly" |
| "images" | List of review images (if any) | [] → no uploaded photos |
| "asin" | Amazon Standard Identification Number (unique product ID) | "B07B4MSV7J" |
| "parent\_asin" | Grouping ID (same for product variations) | "B07B4MSV7J" (identical, meaning this product has no variations or this review applies to the main listing) |
| "user\_id" | Unique reviewer identifier | "AEUL23LUBSPLLOKANUHAUY37SG5A" |
| "timestamp" | Time of review in Unix epoch (milliseconds) | 1565982989421 → corresponds to Aug 16, 2019, 16:16:29 UTC |
| "helpful\_vote" | How many people marked it as helpful | 0 → no votes |
| "verified\_purchase" | Whether the reviewer actually bought the product | true → verified customer |

Review data example from product info dataset (meta\_Appliances.json.gz

):

| **Field** | **Meaning** | **Example / Use** |
| --- | --- | --- |
| "main\_category" | Broad department the product belongs to | "Tools & Home Improvement" — useful for category-based filtering |
| "title" | Product name | "Agitator Cam Kit W/ Dogs for Whirlpool Sears Kenmore 285748 3347085" |
| "average\_rating" | Mean star rating across all users | 4.1 |
| "rating\_number" | Total number of ratings | 34 |
| "features" | Bullet-point product highlights | ["Agitator Cam Kit", "285748"] — could feed into NLP embeddings |
| "description" | Text description | "Appliance Replacement Part ER285748" — can be tokenized for content similarity |
| "price" | Product price | null (missing here, but normally numeric) |
| "images" | Image URLs | You can embed or use CNN-based vision embeddings for multimodal recommendations |
| "store" | Seller or brand | "Whirlpool" |
| "categories" | Full taxonomy hierarchy | ["Appliances", "Parts & Accessories", "Dryer Parts & Accessories", "Replacement Parts"] |
| "details" | Manufacturer specs | Manufacturer, weight, dimensions, etc. |
| "parent\_asin" | Identifier for grouped variations | "B00U1SCKNQ" — links to variations like color/size |
| "bought\_together" | Amazon’s own co-purchase data | null here, but usually includes “Customers also bought…” info |

**Strategy**

Join the datasets on the product ID: “asin”

Load raw json data and store in a MySQL database (from Python)

Can be used for exporting subsamples and easy to use in Jupyter environments