NIA Raúl Martín, 267819

NIA Noel Pedrosa, 269012

NIA Adrià Porta, 268513

# Part 1: Text Processing and Exploratory Data Analysis

## 1. Data Preparation

The dataset contained key product details like titles, descriptions, brands, categories, product specs, sellers, and some numeric values such as price, discount, and rating.

For text preprocessing, we worked with the title and description fields. The main steps included:

- Converting all text to lowercase
- Removing punctuation and special characters
- Breaking text into individual words
- Removing common stopwords like "the" and "and" using NLTK
- Applying stemming
- Cleaning extra spaces and unnecessary numbers

These steps helped simplify the text and reduce the vocabulary size, making the data more suitable for the following information retrieval tasks.

## 2. Working with Non-Text Fields

Fields like brand, category, sub-category, product details, and seller were kept separate instead of combining them into one field, to treat each type of information differently, which can improve search accuracy even though indexing and querying becomes a little bit more complex.

Numeric fields like selling price, actual price, discount, and average rating were treated as numbers, not text, which make them more useful for filtering.

The out_of_stock flag was treated as a simple true/false value (boolean).

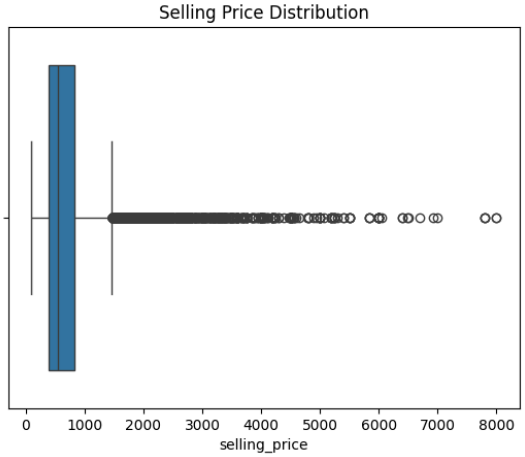
## 3. Exploratory Data Analysis

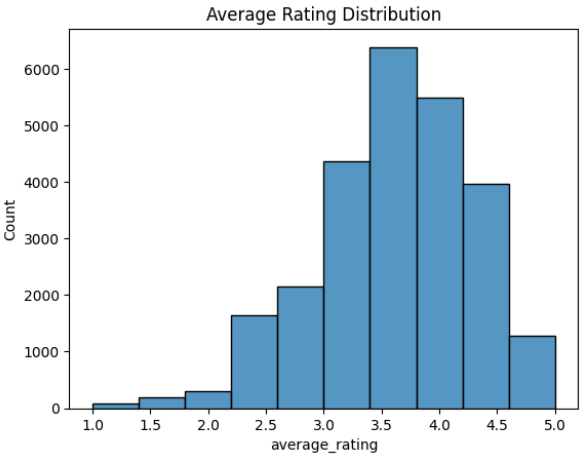
With the cleaned data, we explored patterns and trends in the dataset. Some key insights included:

- Word frequency in titles and descriptions to identify common product terms

- Word clouds to visually highlight frequent keywords
- Distributions of prices, discounts, and ratings
- Top brands and sellers based on frequency
- Out-of-stock analysis to understand availability patterns

The visualizations showed interesting patterns such as skewed price and rating distributions, and a few brands dominating the catalog.




## 4. Key Takeaways

The preprocessing steps effectively prepared the dataset for indexing. Keeping fields like brand and category separate should improve future search accuracy.

The exploratory analysis offered valuable insights into the dataset's structure, trends in pricing, and active sellers. It allowed us to see reliable distributions of the values of some variables, and some characteristics of the dataset, such as the average sentence length in product descriptions or the vocabulary size.