

---

## Part 1: Text Processing and Exploratory Data Analysis

---

You are provided with a document corpus, which is an e-commerce fashion products dataset. You can see an example document in the appendix.

### PART 1: Data preparation

**1.** As a first step, you must pre-process the documents. In particular, for the text fields (title, description) you should:

- Removing stop words
- Tokenization
- Removing punctuation marks
- Stemming
- and... anything else you think it's needed (bonus point)

**2.** Take into account that for future queries, the final output must return (when present) the following information for each of the selected documents:

pid, title, description, brand, category, sub\_category, product\_details, seller, out\_of\_stock, selling\_price, discount, actual\_price, average\_rating, url

**3.** Decide how to handle the fields *category*, *sub\_category*, *brand*, *product\_details*, and *seller* during pre-processing. Should they be merged into a single text field, indexed as separate fields in the inverted index or any other alternative? Justify your choice, considering how their distinctiveness may affect retrieval effectiveness. What are pros and cons of each approach?

**4.** Consider the fields *out\_of\_stock*, *selling\_price*, *discount*, *actual\_price*, and *average\_rating*. Decide how these should be handled during pre-processing to use in further search. Should they be indexed as textual terms?

#### HINTS:

1. As guidance, refer to *validation\_labels.csv*, which will play a pivotal role in the project's second phase. This file contains search results for two different queries, with each document labeled as relevant (1) or not relevant (0). Reflect on how this evaluation context might influence your strategy for processing and weighting different document fields. For your reference: *query\_1*: women full sleeve sweatshirt cotton, *query\_2*: men slim jeans blue
2. Suggested library that may help you in stemming and stop words: **nltk**
3. Make sure you keep pid as it is going to be used for evaluation.

## PART 2: Exploratory Data Analysis

When working with data, it is important to have a better understanding of the content and some statistics. Provide an exploratory data analysis to describe the dataset you are working on in this project and explain the decisions made for the analysis. For example, word counting distribution, average sentence length, vocabulary size, ranking of products based on rating, price, discount, top sellers and brands, out\_of\_stock distribution, word clouds for the most frequent words, and entity recognition. Feel free to do the exploratory analysis and report your findings in the report.

---

## Appendix

---

### *Example:*

```
{  
  "_id": "fa8e22d6-c0b6-5229-bb9e-ad52eda39a0a",  
  "actual_price": "2,999",  
  "average_rating": "3.9",  
  "brand": "York",  
  "category": "Clothing and Accessories",  
  "crawled_at": "02/10/2021, 20:11:51",  
  "description": "Yorker trackpants made from 100% rich combed cotton giving it a rich look. Designed for  
Comfort, Skin friendly fabric, itch-free waistband & great for all year round use. Proudly made in India",  
  "discount": "69% off",  
  "images": [  
  
    "https://rukminim1.flixcart.com/image/128/128/jr3t5e80/track-pant/z/y/n/m-1005combo2-yorker-origi  
nal-imafczg3xfh5qqd4.jpeg?q=70",  
  
    "https://rukminim1.flixcart.com/image/128/128/jr58l8w0/track-pant/w/d/a/l-1005combo8-yorker-origi  
nal-imafczg3pgtxgraq.jpeg?q=70"  
,  
  "out_of_stock": false,  
  "pid": "TKPFCZ9EA7H5FYZH",  
  "product_details": [  
    {  
      "Style Code": "1005COMBO2"  
    },  
  ]}
```

```
        "Closure": "Elastic"
    },
    {
        "Pockets": "Side Pockets"
    },
    {
        "Fabric": "Cotton Blend"
    },
    {
        "Pattern": "Solid"
    },
    {
        "Color": "Multicolor"
    }
],
"seller": "Shyam Enterprises",
"selling_price": "921",
"sub_category": "Bottomwear",
"title": "Solid Men Multicolor Track Pants",
"url":
"https://www.flipkart.com/yorker-solid-men-multicolor-track-pants/p/itmd2c76aadce459?pid=TKPFCZ9EA7H5FYZH&lid=LSTTKPFCZ9EA7H5FYZHVVXWP0&marketplace=FLIPKART&srno=b_1_1&otracker=browse&fm=organic&iid=177a46eb-d053-4732-b3de-fcad6ff59cbd.TKPFCZ9EA7H5FYZH.SEARCH&ssid=utkd4t3gb40000001612415717799"
}
```