DS 5720

Sovann Chang and Jiaying Liang

Project Proposal Update

Next Product Recommendation and Prediction

The first thing we are doing is limiting our scope to only products and sessions within the UK. The original dataset contains multiple countries' data in the 'locale' variable, which we are restricting to only 'UK'. With that in mind, all EDA is on the dataset after filtering out other countries.

EDA:

Products:

Total rows in the dataset: 500,180

Columns and null values:

```
<class 'pandas.core.frame.DataFrame'>
Index: 500180 entries, 913336 to 1413515
Data columns (total 9 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   id        500180 non-null  object
 1   title     500180 non-null  object
 2   price     500180 non-null  float64
 3   brand     495898 non-null  object
 4   color     378078 non-null  object
 5   size      301092 non-null  object
 6   model     243528 non-null  object
 7   material  298955 non-null  object
 8   desc      460922 non-null  object
dtypes: float64(1), object(8)
memory usage: 38.2+ MB
```

Columns we do not plan to utilize in the model: color, size, model

Columns we may use: material

Here is some information about the columns we may/will use.

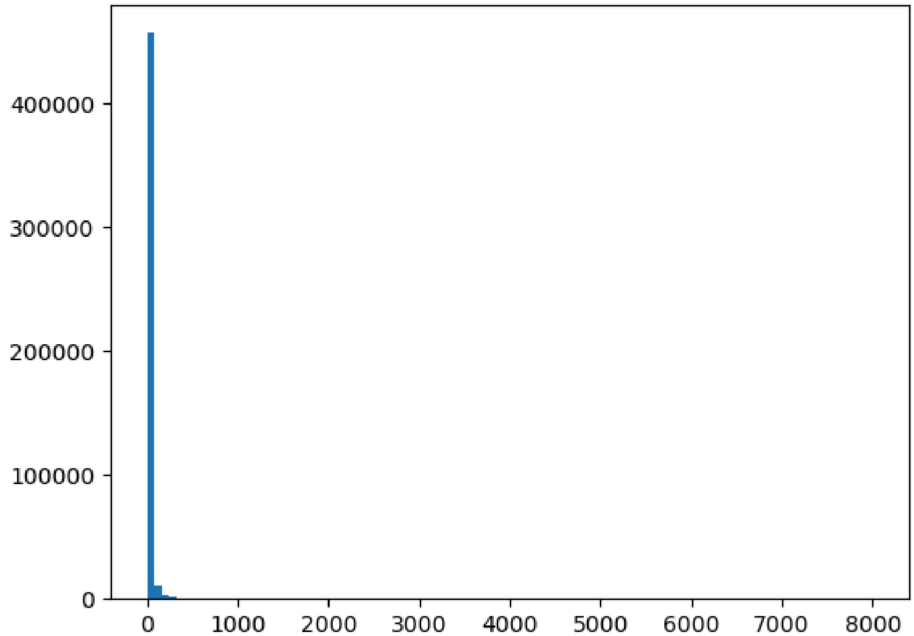
title:

Almost all products have different titles. Value counts are almost all 1.

price:

```
price
40000000.07    25280
9.99           18967
8.99           14592
7.99           13291
12.99          13140
               ...
122.41             1
89.20              1
64.44              1
321.97             1
273.72             1
Name: count, Length: 11311, dtype: int64
```

Obviously, we should not be recommending items that are $40,000,000. After removing items that are that expensive, our most expensive item is $8,000. However, as shown below, the vast majority of our items are under $500.



We will remove products that are over $500 from our dataset.

brand:

```
Generic            2283
L'Oreal            1508
Amazon Basics      1402
LEGO               1335
Morrisons          1287
                   ...
Bookends2pairUK       1
Lipton Iced Tea       1
Q&K                   1
HULY                  1
CRYSTALS              1
Name: brand, Length: 76349, dtype: int64
```

We have 76,349 brands, but the majority of them will have fewer than 5 products. We should filter out rare brands and set them as “other”, but where do we set the cut off for “rare”?

```
brand_counts[brand_counts > 100] # 439
```

```
Generic          2283
L'Oreal          1508
Amazon Basics    1402
LEGO             1335
Morrisons        1287
                 ...
ORETECH           101
Olay              101
Pecute            101
JUSTOTRY          101
POPRUN            101
Name: brand, Length: 439, dtype: int64
```

439 of the brands have 100 or more products. We will not make a final decision on what to do yet, since one-hot encoding 440 brands seems like too much.

material:

```
Plastic                      42638
Metal                        13343
Polyester                    13080
Stainless Steel              10226
Paper                        10029
                             ...
Gel, Silicone                    1
Polyester + Magnet               1
essential oil set                1
tissue paper                     1
Aluminium,Plastic,Resin          1
Name: material, Length: 14775, dtype: int64
```

Most of our nearly 15,000 unique materials appear to be of a few basic types. We'll look into this a bit more:

```
Plastic                          42638
Metal                            13343
Polyester                        13080
Stainless Steel                  10226
Paper                            10029
Cotton                            9552
Silicone                          8508
Wood                              7166
Aluminium                         6752
Rubber                            6716
```

The 10 most common materials are generally a combination of clothing material and building materials. We could do one-hot encoding and call all other materials "other". We are not sure how much this variable will positively affect our model, so we will again not make a final decision right now.

description:
We may combine title and description into one sentence embedding. We may also add size, color, model, and material, TBD.

Session:

Total rows in the dataset: 11,828,181

```
<class 'pandas.core.frame.DataFrame'>
Index: 1182181 entries, 2090535 to 3272715
Data columns (total 3 columns):
 #   Column      Non-Null Count    Dtype
---  ------      --------------    -----
 0   prev_items  1182181 non-null  object
 1   next_item   1182181 non-null  object
 2   locale      1182181 non-null  object
dtypes: object(3)
memory usage: 36.1+ MB
```

We want to know how long our sessions are

```
prev_items_len
2     449320
3     248849
4     152794
5      97431
6      65199
       ...
95         1
88         1
75         1
77         1
66         1
Name: count, Length: 79, dtype: int64
```

The vast majority of sessions are somewhat short, but there are a few long ones. For any session longer than 5 products, we will build our graph using all the products from the session, but we will train our neural network on ONLY the sequence of the last 5 products in the session.

Data cleaning strategies:

1. products:
    a. locale: only consider UK
    b. price: only consider under $500
    c. brand: we want to consider filtering out rare brands, but need to define what is “rare”
    d. title+desc(+color+size+model+material): we would like to encode all those words into word embeddings.
    e. author: drop it

2. sessions:
    a. locale: only consider UK
    b. create a column shows the length of prev_items
    c. prev_items_len: For any session longer than 5 products, we will build our graph using all the products from the session, but we will train our neural network on ONLY the sequence of the last 5 products in the session.

Timeline:

- Rough Timeline
    - 4/4-4/11: Setting up data, preliminary modeling
    - 4/11-4/16: Refining models, realistically fixing errors from previous week, creating presentation
    - 4/16-4/25: Clean repository, finalize models, write report