Report

Rajith Jarang

Data filtering :

First we will filter the data and select only 'shorts category'

Run the python file:

For getting shorts.py

This will output a file 'Shorts.csv' which we will use

To download the images (we only download 2nd link in the imageUrl column):

Just run all the 6 python files in the folder name at a time :

"For images run this codes"

All the images can be found in "image" directory

We extract feature vector for 'images' and for 'text' separately and finally we try to find the similarity using both

Algorithm for vector extraction for text:

Assumption :

I assumed these column are useful for finding the similarity of the products

1.Title (Similar titles products can be similar)

2.Description('')

3.Brand

3. Mrp as mrp across different platform doesn't change souseful checking both are the same items or not)

4. product family

5.color

6.key specs, detailed key specs

Vectorize all these columns so that we can find the similarity using these vectors

Note: NaN cases should be handled here

Run the features.py file and direct the output to a file those are the feature vectors

Command:

1. python features.py

'features' file has the features vector of the text

Algorithm for vector extraction for image:

Try to use the "**BVLC GoogLeNet** model" due to error installing caffe

It should have given better resultsI cannot able to do so here is the link:

http://www.marekrei.com/blog/transforming-images-to-feature-vectors/

So I used some vector extraction library "Graphlab" to get the feature vectors of a image

Run the file

1. python graphlab_image_feature_extraction.py

'glfeature' has the image vector for each image

Similarity calculation:

Use cosine similarity by giving optimal weight to image vector and to text vector

Sort the similarity score and get the top 25 similar products and print it in a json format with the similarity score

This way we can achieve that state where we can see the similar products related to the product

But unfortunately I cannot complete the whole thing because I don't have gpu support and these algorithm are taking so much time no output file

This is all can do in one day so much more to fix