Amazon Apparel Recommendations

[5.1] Missing data for various features.

Basic stats for the feature: product_type_name

In [0]:

```
# We have total 72 unique type of product_type_names
print(data['product_type_name'].describe())

# 91.62% (167794/183138) of the products are shirts,
```

```
count     183138
unique        72
top        SHIRT
freq      167794
Name: product_type_name, dtype: object
```

In [0]:

```
# names of different product types
print(data['product_type_name'].unique())
```

```
['SHIRT' 'SWEATER' 'APPAREL' 'OUTDOOR_RECREATION_PRODUCT'
 'BOOKS_1973_AND_LATER' 'PANTS' 'HAT' 'SPORTING_GOODS' 'DRESS' 'UNDERWEAR'
 'SKIRT' 'OUTERWEAR' 'BRA' 'ACCESSORY' 'ART_SUPPLIES' 'SLEEPWEAR'
 'ORCA_SHIRT' 'HANDBAG' 'PET_SUPPLIES' 'SHOES' 'KITCHEN' 'ADULT_COSTUME'
 'HOME_BED_AND_BATH' 'MISC_OTHER' 'BLAZER' 'HEALTH_PERSONAL_CARE'
 'TOYS_AND_GAMES' 'SWIMWEAR' 'CONSUMER_ELECTRONICS' 'SHORTS' 'HOME'
 'AUTO_PART' 'OFFICE_PRODUCTS' 'ETHNIC_WEAR' 'BEAUTY'
 'INSTRUMENT_PARTS_AND_ACCESSORIES' 'POWERSPORTS_PROTECTIVE_GEAR' 'SHIRTS'
 'ABIS_APPAREL' 'AUTO_ACCESSORY' 'NONAPPARELMISC' 'TOOLS' 'BABY_PRODUCT'
 'SOCKSHOSIERY' 'POWERSPORTS_RIDING_SHIRT' 'EYEWEAR' 'SUIT'
 'OUTDOOR_LIVING' 'POWERSPORTS_RIDING_JACKET' 'HARDWARE' 'SAFETY_SUPPLY'
 'ABIS_DVD' 'VIDEO_DVD' 'GOLF_CLUB' 'MUSIC_POPULAR_VINYL'
 'HOME_FURNITURE_AND_DECOR' 'TABLET_COMPUTER' 'GUILD_ACCESSORIES'
 'ABIS_SPORTS' 'ART_AND_CRAFT_SUPPLY' 'BAG' 'MECHANICAL_COMPONENTS'
 'SOUND_AND_RECORDING_EQUIPMENT' 'COMPUTER_COMPONENT' 'JEWELRY'
 'BUILDING_MATERIAL' 'LUGGAGE' 'BABY_COSTUME' 'POWERSPORTS_VEHICLE_PART'
 'PROFESSIONAL_HEALTHCARE' 'SEEDS_AND_PLANTS' 'WIRELESS_ACCESSORY']
```

In [0]:

```
# find the 10 most frequent product_type_names.
product_type_count = Counter(list(data['product_type_name']))
product_type_count.most_common(10)
```

Out[0]:

```
[('SHIRT', 167794),
 ('APPAREL', 3549),
 ('BOOKS_1973_AND_LATER', 3336),
 ('DRESS', 1584),
 ('SPORTING_GOODS', 1281),
 ('SWEATER', 837),
 ('OUTERWEAR', 796),
 ('OUTDOOR_RECREATION_PRODUCT', 729),
 ('ACCESSORY', 636),
 ('UNDERWEAR', 425)]
```

Basic stats for the feature: brand

In [0]:

```
# there are 10577 unique brands
print(data['brand'].describe())

# 183138 - 182987 = 151 missing values.
```

```
count     182987
unique     10577
top         Zago
freq         223
Name: brand, dtype: object
```

In [0]:

```
brand_count = Counter(list(data['brand']))
brand_count.most_common(10)
```

Out[0]:

```
[('Zago', 223),
 ('XQS', 222),
 ('Yayun', 215),
 ('YUNY', 198),
 ('XiaoTianXin-women clothes', 193),
 ('Generic', 192),
 ('Boohoo', 190),
 ('Alion', 188),
 ('Abetteric', 187),
 ('TheMogan', 187)]
```

Basic stats for the feature: color

In [0]:

```
print(data['color'].describe())


# we have 7380 unique colors
# 7.2% of products are black in color
# 64956 of 183138 products have brand information. That's approx 35.4%.
```

```
count      64956
unique      7380
top        Black
freq       13207
Name: color, dtype: object
```

In [0]:

```
color_count = Counter(list(data['color']))
color_count.most_common(10)
```

Out[0]:

```
[(None, 118182),
 ('Black', 13207),
 ('White', 8616),
 ('Blue', 3570),
 ('Red', 2289),
 ('Pink', 1842),
 ('Grey', 1499),
 ('*', 1388),
 ('Green', 1258),
 ('Multi', 1203)]
```

Basic stats for the feature: formatted_price

In [0]:

```


print(data['formatted_price'].describe())

# Only 28,395 (15.5% of whole data) products with price information
```

```
count       28395
unique       3135
top        $19.99
freq          945
Name: formatted_price, dtype: object
```

In [0]:

```
price_count = Counter(list(data['formatted_price']))
price_count.most_common(10)
```

Out[0]:

```
[(None, 154743),
 ('$19.99', 945),
 ('$9.99', 749),
 ('$9.50', 601),
 ('$14.99', 472),
 ('$7.50', 463),
 ('$24.99', 414),
 ('$29.99', 370),
 ('$8.99', 343),
 ('$9.01', 336)]
```

Basic stats for the feature: title

In [0]:

```
print(data['title'].describe())

# All of the products have a title.
# Titles are fairly descriptive of what the product is.
# We use titles extensively in this workshop
# as they are short and informative.
```

```
count                                                183138
unique                                               175985
top       Nakoda Cotton Self Print Straight Kurti For Women
freq                                                     77
Name: title, dtype: object
```

In [0]:

```
data.to_pickle('pickels/180k_apparel_data')
```

We save data files at every major step in our processing in "pickle" files. If you are stuck anywhere (or) if some code takes too long to run on your laptop, you may use the pickle files we give you to speed things up.

In [0]:

```
# consider products which have price information
# data['formatted_price'].isnull() => gives the information
#about the dataframe row's which have null values price == None|Null
data = data.loc[~data['formatted_price'].isnull()]
print('Number of data points After eliminating price=NULL :', data.shape[0])
```

```
Number of data points After eliminating price=NULL : 28395
```

In [0]:

```
# consider products which have color information
# data['color'].isnull() => gives the information about the dataframe row's which have
 null values price == None|Null
data =data.loc[~data['color'].isnull()]
print('Number of data points After eliminating color=NULL :', data.shape[0])
```

```
Number of data points After eliminating color=NULL : 28385
```

We brought down the number of data points from 183K to 28K.

We are processing only 28K points so that most of the workshop participants can run this code on thier laptops in a reasonable amount of time.

For those of you who have powerful computers and some time to spare, you are recommended to use all of the 183K images.

In [0]:

```
data.to_pickle('pickels/28k_apparel_data')
```

In [0]:

```
# You can download all these 28k images using this code below.
# You do NOT need to run this code and hence it is commented.


'''
from PIL import Image
import requests
from io import BytesIO

for index, row in images.iterrows():
        url = row['large_image_url']
        response = requests.get(url)
        img = Image.open(BytesIO(response.content))
        img.save('images/28k_images/'+row['asin']+'.jpeg')


'''
```

Out[0]:

```
"\nfrom PIL import Image\nimport requests\nfrom io import BytesIO\n\nfor i
ndex, row in images.iterrows():\n        url = row['large_image_url']\n
     response = requests.get(url)\n        img = Image.open(BytesIO(respon
se.content))\n        img.save('workshop/images/28k_images/'+row['asin']
+'.jpeg')\n\n\n"
```

[5.2] Remove near duplicate items

[5.2.1] Understand about duplicates.

In [0]:

```
# read data from pickle file from previous stage
data = pd.read_pickle('pickels/28k_apparel_data')

# find number of products that have duplicate titles.
print(sum(data.duplicated('title')))
# we have 2325 products which have same title but different color
```

```
2325
```

These shirts are exactly same except in size (S, M,L,XL)

:B00AQ4GMCK :B00AQ4GMTS

:B00AQ4GMLQ :B00AQ4GN3I

These shirts exactly same except in color

:B00G278GZ6 :B00G278W6O

:B00G278Z2A :B00G2786X8

In our data there are many duplicate products like the above examples, we need to de-dupe them for better results.

[5.2.2] Remove duplicates : Part 1

In [0]:

```
# read data from pickle file from previous stage
data = pd.read_pickle('pickels/28k_apparel_data')
```

In [0]:

```
data.head()
```

Out[0]:

	asin	brand	color	medium_image_url	product_type_name	title
4	B004GSI2OS	FeatherLite	Onyx Black/Stone	https://images-na.ssl-images-amazon.com/images...	SHIRT	Featherlite Ladies' Long Sleeve Stain Resistan...
6	B012YX2ZPI	HX-Kingdom Fashion T-shirts	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	Women's Unique 100% Cotton T - Special Olympic...
11	B001LOUGE4	Fitness Etc.	Black	https://images-na.ssl-images-amazon.com/images...	SHIRT	Ladies Cotton Tank 2x1 Ribbed Tank Top
15	B003BSRPB0	FeatherLite	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	FeatherLite Ladies' Moisture Free Mesh Sport S...
21	B014ICEDNA	FNC7C	Purple	https://images-na.ssl-images-amazon.com/images...	SHIRT	Supernatural Chibis Sam Dean And Castiel Short...

In [0]:

```
# Remove All products with very few words in title
data_sorted = data[data['title'].apply(lambda x: len(x.split())>4)]
print("After removal of products with short description:", data_sorted.shape[0])
```

```
After removal of products with short description: 27949
```

In [0]:

```
# Sort the whole data based on title (alphabetical order of title)
data_sorted.sort_values('title',inplace=True, ascending=False)
data_sorted.head()
```

Out[0]:

	asin	brand	color	medium_image_url	product_type_name	t
61973	B06Y1KZ2WB	Éclair	Black/Pink	https://images-na.ssl-images-amazon.com/images...	SHIRT	Éc Wome Prin Thin St Blo Blac
133820	B010RV33VE	xiaoming	Pink	https://images-na.ssl-images-amazon.com/images...	SHIRT	xiaom Wom Sleevel Lo Long shir
81461	B01DDSDLNS	xiaoming	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	xiaom Wome W L Sle Sir Bre
75995	B00X5LYO9Y	xiaoming	Red Anchors	https://images-na.ssl-images-amazon.com/images...	SHIRT	xiaom Stri T Patch/B Sle Anch
151570	B00WPJG35K	xiaoming	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	xiaom Sle Sh Lo Tas Kim Wom

Some examples of dupliacte titles that differ only in the last few words.

```
Titles 1:
16. woman's place is in the house and the senate shirts for Womens XXL White
17. woman's place is in the house and the senate shirts for Womens M Grey

Title 2:
25. tokidoki The Queen of Diamonds Women's Shirt X-Large
26. tokidoki The Queen of Diamonds Women's Shirt Small
27. tokidoki The Queen of Diamonds Women's Shirt Large

Title 3:
61. psychedelic colorful Howling Galaxy Wolf T-shirt/Colorful Rainbow Animal Pri
nt Head Shirt for woman Neon Wolf t-shirt
62. psychedelic colorful Howling Galaxy Wolf T-shirt/Colorful Rainbow Animal Pri
nt Head Shirt for woman Neon Wolf t-shirt
63. psychedelic colorful Howling Galaxy Wolf T-shirt/Colorful Rainbow Animal Pri
nt Head Shirt for woman Neon Wolf t-shirt
64. psychedelic colorful Howling Galaxy Wolf T-shirt/Colorful Rainbow Animal Pri
nt Head Shirt for woman Neon Wolf t-shirt
```

In [0]:

```
indices = []
for i,row in data_sorted.iterrows():
    indices.append(i)
```

In [0]:

```
import itertools
stage1_dedupe_asins = []
i = 0
j = 0
num_data_points = data_sorted.shape[0]
while i < num_data_points and j < num_data_points:

    previous_i = i

    # store the list of words of ith string in a, ex: a = ['tokidoki', 'The', 'Queen',
 'of', 'Diamonds', 'Women's', 'Shirt', 'X-Large']
    a = data['title'].loc[indices[i]].split()

    # search for the similar products sequentially
    j = i+1
    while j < num_data_points:

        # store the list of words of jth string in b, ex: b = ['tokidoki', 'The', 'Quee
n', 'of', 'Diamonds', 'Women's', 'Shirt', 'Small']
        b = data['title'].loc[indices[j]].split()

        # store the maximum length of two strings
        length = max(len(a), len(b))

        # count is used to store the number of words that are matched in both strings
        count  = 0

        # itertools.zip_longest(a,b): will map the corresponding words in both strings,
 it will appened None in case of unequal strings
        # example: a =['a', 'b', 'c', 'd']
        # b = ['a', 'b', 'd']
        # itertools.zip_longest(a,b): will give [('a','a'), ('b','b'), ('c','d'), ('d',
 None)]
        for k in itertools.zip_longest(a,b):
            if (k[0] == k[1]):
                count += 1

        # if the number of words in which both strings differ are > 2 , we are consider
ing it as those two apperals are different
        # if the number of words in which both strings differ are < 2 , we are consider
ing it as those two apperals are same, hence we are ignoring them
        if (length - count) > 2: # number of words in which both sensences differ
            # if both strings are differ by more than 2 words we include the 1st string
 index
            stage1_dedupe_asins.append(data_sorted['asin'].loc[indices[i]])

            # if the comaprision between is between num_data_points, num_data_points-1
 strings and they differ in more than 2 words we include both
            if j == num_data_points-1: stage1_dedupe_asins.append(data_sorted['asin'].l
oc[indices[j]])

            # start searching for similar apperals corresponds 2nd string
            i = j
            break
        else:
            j += 1
    if previous_i == i:
        break
```

In [0]:

```
data = data.loc[data['asin'].isin(stage1_dedupe_asins)]
```

We removed the dupliactes which differ only at the end.

In [0]:

```
print('Number of data points : ', data.shape[0])
```

```
Number of data points :  17593
```

In [0]:

```
data.to_pickle('pickels/17k_apperal_data')
```

[5.2.3] Remove duplicates : Part 2

```
In the previous cell, we sorted whole data in alphabetical order of  titles.The
n, we removed titles which are adjacent and very similar title

But there are some products whose titles are not adjacent but very similar.

Examples:

Titles-1
86261.  UltraClub Women's Classic Wrinkle-Free Long Sleeve Oxford Shirt, Pink, X
X-Large
115042. UltraClub Ladies Classic Wrinkle-Free Long-Sleeve Oxford Light Blue XXL

TItles-2
75004.  EVALY Women's Cool University Of UTAH 3/4 Sleeve Raglan Tee
109225. EVALY Women's Unique University Of UTAH 3/4 Sleeve Raglan Tees
120832. EVALY Women's New University Of UTAH 3/4-Sleeve Raglan Tshirt
```

In [0]:

```
data = pd.read_pickle('pickels/17k_apperal_data')
```

In [0]:

```
# This code snippet takes significant amount of time.
# O(n^2) time.
# Takes about an hour to run on a decent computer.

indices = []
for i,row in data.iterrows():
    indices.append(i)

stage2_dedupe_asins = []
while len(indices)!=0:
    i = indices.pop()
    stage2_dedupe_asins.append(data['asin'].loc[i])
    # consider the first apperal's title
    a = data['title'].loc[i].split()
    # store the list of words of ith string in a, ex: a = ['tokidoki', 'The', 'Queen',
 'of', 'Diamonds', 'Women's', 'Shirt', 'X-Large']
    for j in indices:

        b = data['title'].loc[j].split()
        # store the list of words of jth string in b, ex: b = ['tokidoki', 'The', 'Quee
n', 'of', 'Diamonds', 'Women's', 'Shirt', 'X-Large']

        length = max(len(a),len(a))

        # count is used to store the number of words that are matched in both strings
        count  = 0

        # itertools.zip_longest(a,b): will map the corresponding words in both strings,
 it will appened None in case of unequal strings
        # example: a =['a', 'b', 'c', 'd']
        # b = ['a', 'b', 'd']
        # itertools.zip_longest(a,b): will give [('a','a'), ('b','b'), ('c','d'), ('d',
 None)]
        for k in itertools.zip_longest(a,b):
            if (k[0]==k[1]):
                count += 1

        # if the number of words in which both strings differ are < 3 , we are consider
ing it as those two apperals are same, hence we are ignoring them
        if (length - count) < 3:
            indices.remove(j)
```

In [0]:

```
# from whole previous products we will consider only
# the products that are found in previous cell
data = data.loc[data['asin'].isin(stage2_dedupe_asins)]
```

In [0]:

```
print('Number of data points after stage two of dedupe: ',data.shape[0])
# from 17k apperals we reduced to 16k apperals
```

```
Number of data points after stage two of dedupe:  16042
```

In [0]:

```
data.to_pickle('pickels/16k_apperal_data')
# Storing these products in a pickle file
# candidates who wants to download these files instead
# of 180K they can download and use them from the Google Drive folder.
```

[10.2] Keras and Tensorflow to extract features

In [0]:

```
import numpy as np
from keras.preprocessing.image import ImageDataGenerator
from keras.models import Sequential
from keras.layers import Dropout, Flatten, Dense
from keras import applications
from sklearn.metrics import pairwise_distances
import matplotlib.pyplot as plt
import requests
from PIL import Image
import pandas as pd
import pickle
```

```
Using TensorFlow backend.
```

In [0]:

```
# https://gist.github.com/fchollet/f35fbc80e066a49d65f1688a7e99f069
# Code reference: https://blog.keras.io/building-powerful-image-classification-models-u
sing-very-little-data.html



# This code takes 40 minutes to run on a modern GPU (graphics card)
# like Nvidia  1050.
# GPU (NVidia 1050): 0.175 seconds per image

# This codse takes 160 minutes to run on a high end i7 CPU
# CPU (i7): 0.615 seconds per image.

#Do NOT run this code unless you want to wait a few hours for it to generate output

# each image is converted into 25088 length dense-vector


'''
# dimensions of our images.
img_width, img_height = 224, 224

top_model_weights_path = 'bottleneck_fc_model.h5'
train_data_dir = 'images2/'
nb_train_samples = 16042
epochs = 50
batch_size = 1


def save_bottlebeck_features():

    #Function to compute VGG-16 CNN for image feature extraction.

    asins = []
    datagen = ImageDataGenerator(rescale=1. / 255)

    # build the VGG16 network
    model = applications.VGG16(include_top=False, weights='imagenet')
    generator = datagen.flow_from_directory(
        train_data_dir,
        target_size=(img_width, img_height),
        batch_size=batch_size,
        class_mode=None,
        shuffle=False)

    for i in generator.filenames:
        asins.append(i[2:-5])

    bottleneck_features_train = model.predict_generator(generator, nb_train_samples //
 batch_size)
    bottleneck_features_train = bottleneck_features_train.reshape((16042,25088))

    np.save(open('16k_data_cnn_features.npy', 'wb'), bottleneck_features_train)
    np.save(open('16k_data_cnn_feature_asins.npy', 'wb'), np.array(asins))


save_bottlebeck_features()

'''
```

Assignment

In [1]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?cli
ent_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleuserconten
t.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&scope=email%20http
s%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2Fwww.google
apis.com%2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.
photos.readonly%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fpeopleapi.read
only&response_type=code

Enter your authorization code:
··········
Mounted at /content/drive
```

In [0]:

```
path='./drive/My Drive/Applied_AI_Workshop_Code_Data/pickels/16k_apperal_data_preproces
sed'
```

In [0]:

```
data=pd.read_pickle(path)
```

In [59]:

```
df.head()
```

Out[59]:

	asin	brand	color	medium_image_url	product_type_name	title
4	B004GSI2OS	FeatherLite	Onyx Black/ Stone	https://images-na.ssl-images-amazon.com/images...	SHIRT	Featherlite Ladies' Long Sleeve Stain Resistan...
6	B012YX2ZPI	HX-Kingdom Fashion T-shirts	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	Women's Unique 100% Cotton T - Special Olympic...
15	B003BSRPB0	FeatherLite	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	FeatherLite Ladies' Moisture Free Mesh Sport S...
27	B014ICEJ1Q	FNC7C	Purple	https://images-na.ssl-images-amazon.com/images...	SHIRT	Supernatural Chibis Sam Dean And Castiel O Nec...
46	B01NACPBG2	Fifth Degree	Black	https://images-na.ssl-images-amazon.com/images...	SHIRT	Fifth Degree Womens Gold Foil Graphic Tees Jun...

In [0]:

```
# Utility Functions which we will use through the rest of the workshop.


#Display an image
def display_img(url,ax,fig):
    # we get the url of the apparel and download it
    response = requests.get(url)
    img = Image.open(BytesIO(response.content))
    # we will display it in notebook
    plt.imshow(img)

#plotting code to understand the algorithm's decision.
def plot_heatmap(keys, values, labels, url, text):
        # keys: list of words of recommended title
        # values: len(values) ==  len(keys), values(i) represents the occurence of the
 word keys(i)
        # labels: len(labels) == len(keys), the values of labels depends on the model w
e are using
                # if model == 'bag of words': labels(i) = values(i)
                # if model == 'tfidf weighted bag of words':labels(i) = tfidf(keys(i))
                # if model == 'idf weighted bag of words':labels(i) = idf(keys(i))
        # url : apparel's url

        # we will devide the whole figure into two parts
        gs = gridspec.GridSpec(2, 2, width_ratios=[4,1], height_ratios=[4,1])
        fig = plt.figure(figsize=(25,3))

        # 1st, ploting heat map that represents the count of commonly ocurred words in
 title2
        ax = plt.subplot(gs[0])
        # it displays a cell in white color if the word is intersection(lis of words of
 title1 and list of words of title2), in black if not
        ax = sns.heatmap(np.array([values]), annot=np.array([labels]))
        ax.set_xticklabels(keys) # set that axis labels as the words of title
        ax.set_title(text) # apparel title

        # 2nd, plotting image of the the apparel
        ax = plt.subplot(gs[1])
        # we don't want any grid lines for image and no labels on x-axis and y-axis
        ax.grid(False)
        ax.set_xticks([])
        ax.set_yticks([])

        # we call dispaly_img based with paramete url
        display_img(url, ax, fig)

        # displays combine figure ( heat map and image together)
        plt.show()

def plot_heatmap_image(doc_id, vec1, vec2, url, text, model):

    # doc_id : index of the title1
    # vec1 : input apparels's vector, it is of a dict type {word:count}
    # vec2 : recommended apparels's vector, it is of a dict type {word:count}
    # url : apparels image url
    # text: title of recomonded apparel (used to keep title of image)
    # model, it can be any of the models,
        # 1. bag_of_words
        # 2. tfidf
        # 3. idf

    # we find the common words in both titles, because these only words contribute to t
he distance between two title vec's
    intersection = set(vec1.keys()) & set(vec2.keys())

    # we set the values of non intersecting words to zero, this is just to show the dif
ference in heatmap
    for i in vec2:
        if i not in intersection:
            vec2[i]=0

    # for labeling heatmap, keys contains list of all words in title2
    keys = list(vec2.keys())
    #  if ith word in intersection(lis of words of title1 and list of words of title2):
 values(i)=count of that word in title2 else values(i)=0
    values = [vec2[x] for x in vec2.keys()]

    # labels: len(labels) == len(keys), the values of labels depends on the model we ar
e using
        # if model == 'bag of words': labels(i) = values(i)
        # if model == 'tfidf weighted bag of words':labels(i) = tfidf(keys(i))
        # if model == 'idf weighted bag of words':labels(i) = idf(keys(i))

    if model == 'bag_of_words':
        labels = values
    elif model == 'tfidf':
        labels = []
        for x in vec2.keys():
            # tfidf_title_vectorizer.vocabulary_ it contains all the words in the corpu
s
            # tfidf_title_features[doc_id, index_of_word_in_corpus] will give the tfidf
 value of word in given document (doc_id)
            if x in  tfidf_title_vectorizer.vocabulary_:
                labels.append(tfidf_title_features[doc_id, tfidf_title_vectorizer.vocab
ulary_[x]])
            else:
                labels.append(0)
    elif model == 'idf':
        labels = []
        for x in vec2.keys():
            # idf_title_vectorizer.vocabulary_ it contains all the words in the corpus
            # idf_title_features[doc_id, index_of_word_in_corpus] will give the idf val
ue of word in given document (doc_id)
            if x in  idf_title_vectorizer.vocabulary_:
                labels.append(idf_title_features[doc_id, idf_title_vectorizer.vocabular
y_[x]])
            else:
                labels.append(0)

    plot_heatmap(keys, values, labels, url, text)


# this function gets a list of wrods along with the frequency of each
# word given "text"
def text_to_vector(text):
    word = re.compile(r'\w+')
    words = word.findall(text)
    # words stores list of all words in given string, you can try 'words = text.split
()' this will also gives same result
    return Counter(words) # Counter counts the occurence of each word in list, it retur
```

```
ns dict type object {word1:count}



def get_result(doc_id, content_a, content_b, url, model):
    text1 = content_a
    text2 = content_b

    # vector1 = dict{word11:#count, word12:#count, etc.}
    vector1 = text_to_vector(text1)

    # vector1 = dict{word21:#count, word22:#count, etc.}
    vector2 = text_to_vector(text2)

    plot_heatmap_image(doc_id, vector1, vector2, url, text2, model)
```

Model For IDF Based Features

In [0]:

```
# we need to convert the values into float
idf_title_features  = idf_title_features.astype(np.float)

for i in idf_title_vectorizer.vocabulary_.keys():
    # for every word in whole corpus we will find its idf value
    idf_val = idf(i)

    # to calculate idf_title_features we need to replace the count values with the idf
 values of the word
    # idf_title_features[:, idf_title_vectorizer.vocabulary_[i]].nonzero()[0] will retu
rn all documents in which the word i present
    for j in idf_title_features[:, idf_title_vectorizer.vocabulary_[i]].nonzero()[0]:

        # we replace the count values of word i in document j with  idf_value of word i

        # idf_title_features[doc_id, index_of_word_in_courpus] = idf value of word
        idf_title_features[j,idf_title_vectorizer.vocabulary_[i]] = idf_val
```

In [64]:

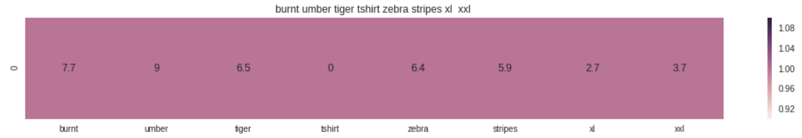
```
idf_title_features.shape
```

Out[64]:

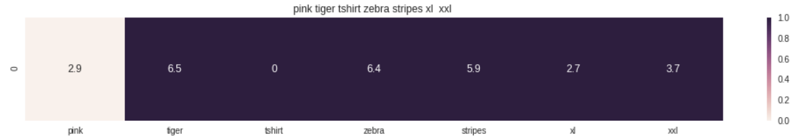
```
(16042, 11103)
```

In [65]:

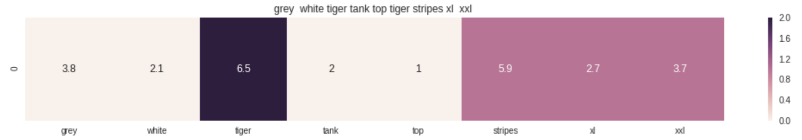
```
def idf_model(doc_id, num_results):
    # doc_id: apparel's id in given corpus

    # pairwise_dist will store the distance from given input apparel to all remaining a
pparels
    # the metric we used here is cosine, the coside distance is mesured as K(X, Y) = <
X, Y> / (||X||*||Y||)
    # http://scikit-learn.org/stable/modules/metrics.html#cosine-similarity
    pairwise_dist = pairwise_distances(idf_title_features,idf_title_features[doc_id])

    # np.argsort will return indices of 9 smallest distances
    indices = np.argsort(pairwise_dist.flatten())[0:num_results]
    #pdists will store the 9 smallest distances
    pdists  = np.sort(pairwise_dist.flatten())[0:num_results]

    #data frame indices of the 9 smallest distace's
    df_indices = list(data.index[indices])

    for i in range(0,len(indices)):
        get_result(indices[i],data['title'].loc[df_indices[0]], data['title'].loc[df_in
dices[i]], data['medium_image_url'].loc[df_indices[i]], 'idf')
        print('ASIN :',data['asin'].loc[df_indices[i]])
        print('Brand :',data['brand'].loc[df_indices[i]])
        print ('euclidean distance from the given image :', pdists[i])
        print('='*125)



idf_model(12566,20)
# in the output heat map each value represents the idf values of the label word, the co
lor represents the intersection with inputs title
```



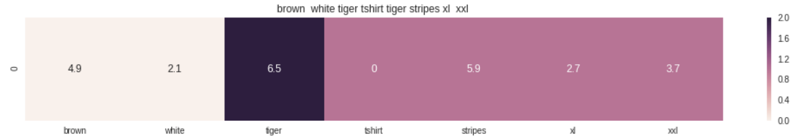
```
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 0.0
==========================================================================
===================================================
```



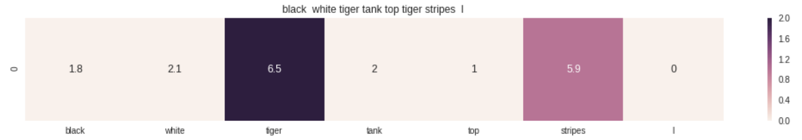
```
ASIN : B00JXQASS6
Brand : Si Row
euclidean distance from the given image : 12.20461230843029
==========================================================================
===================================================
```



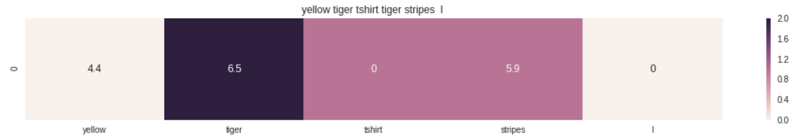
```
ASIN : B00JXQAFZ2
Brand : Si Row
euclidean distance from the given image : 14.432794112662998
==========================================================================
===================================================
```



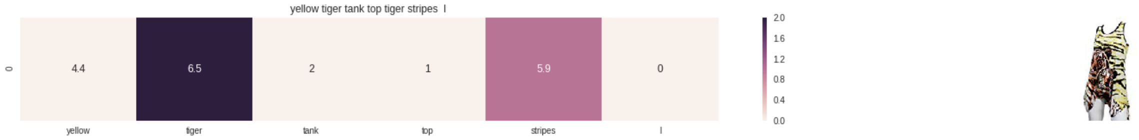
```
ASIN : B00JXQCWTO
Brand : Si Row
euclidean distance from the given image : 14.467956601978512
==========================================================================
===================================================
```



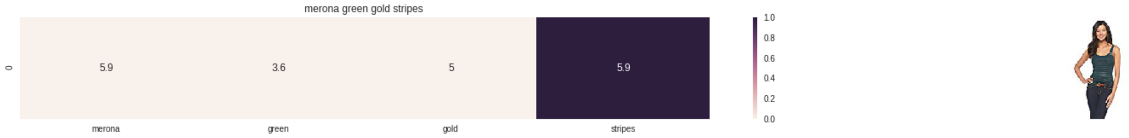
```
ASIN : B00JXQAO94
Brand : Si Row
euclidean distance from the given image : 14.780621107195545
==========================================================================
===================================================
```



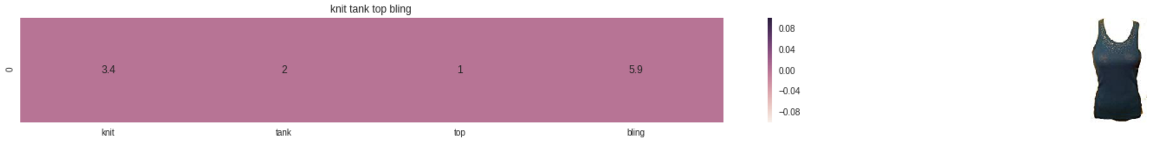
```
ASIN : B00JXQCUIC
Brand : Si Row
euclidean distance from the given image : 14.89835054151571
============================================================================
====================================================
```



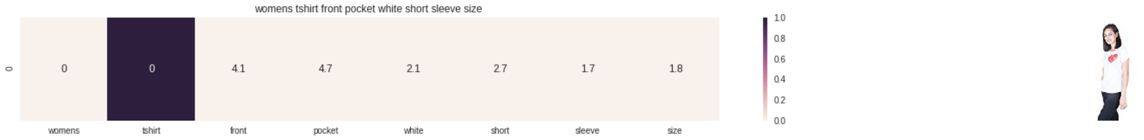
```
ASIN : B00JXQAUWA
Brand : Si Row
euclidean distance from the given image : 15.173045247524719
============================================================================
====================================================
```



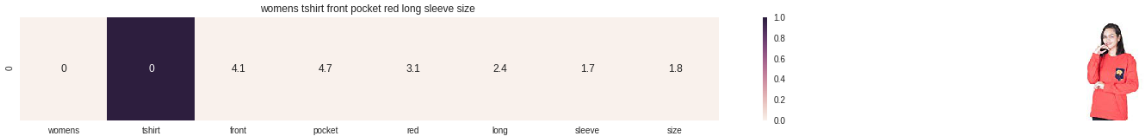
```
ASIN : B01KVZUB6G
Brand : Merona
euclidean distance from the given image : 17.927854989346454
============================================================================
====================================================
```



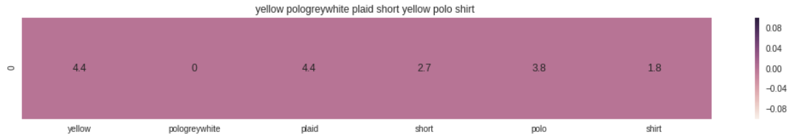
```
ASIN : B01NBQSBMN
Brand : Pink Cattlelac
euclidean distance from the given image : 18.26715981839026
============================================================================
====================================================
```



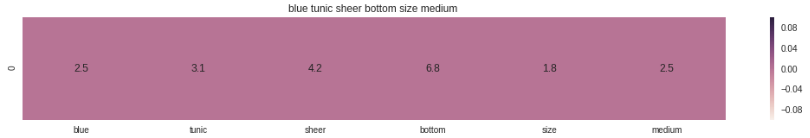
```
ASIN : B01JR73FSK
Brand : Lofbaz
euclidean distance from the given image : 18.519936260793127
============================================================================
====================================================
```



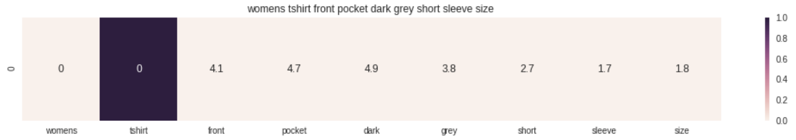
```
ASIN : B01JR72WHA
Brand : Lofbaz
euclidean distance from the given image : 18.61326206122322
============================================================================
====================================================
```



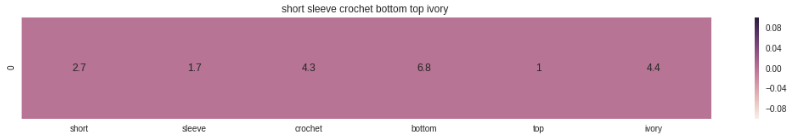
```
ASIN : B0755TBRM6
Brand : RuggedButts
euclidean distance from the given image : 18.948799587702137
===========================================================================
===================================================
```



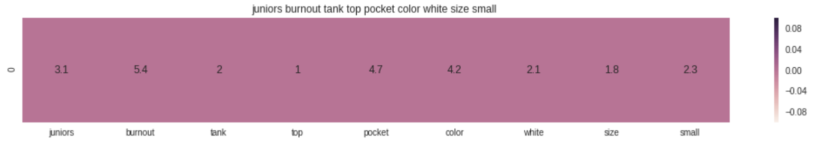
```
ASIN : B01NAOFEQE
Brand : Panhandle Slim
euclidean distance from the given image : 19.31873877117325
===========================================================================
===================================================
```



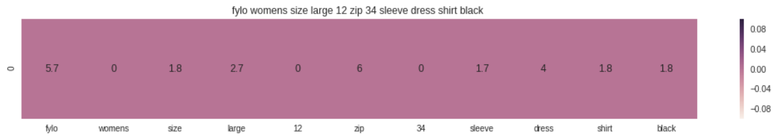
```
ASIN : B01JR73BMA
Brand : Lofbaz
euclidean distance from the given image : 19.423851016684544
===========================================================================
===================================================
```



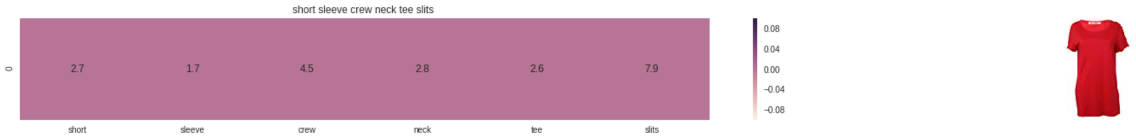
```
ASIN : B0749P1QFC
Brand : Heart and Hips
euclidean distance from the given image : 19.457896611876382
===========================================================================
===================================================
```



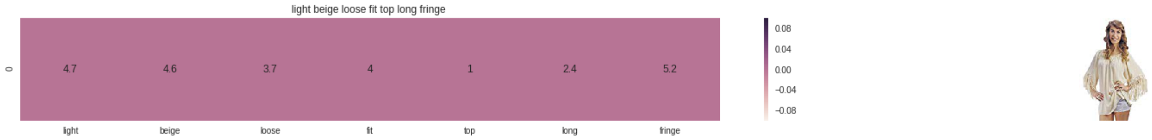
```
ASIN : B00FJR0VG2
Brand : The Blue Brand
euclidean distance from the given image : 19.490960103008423
===========================================================================
===================================================
```



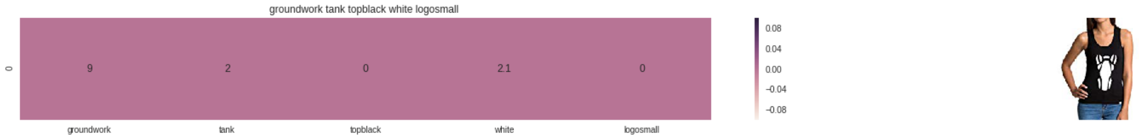
```
ASIN : B0718Y9J4M
Brand : f
euclidean distance from the given image : 19.663869330393368
==========================================================================
===================================================
```



```
ASIN : B01NAAIH0W
Brand : Michael Stars
euclidean distance from the given image : 19.741402922106168
==========================================================================
===================================================
```



```
ASIN : B01BX9G1HW
Brand : Luxury Divas
euclidean distance from the given image : 19.750829242543315
==========================================================================
===================================================
```



```
ASIN : B06XH59DYM
Brand : Privileged and Plaid
euclidean distance from the given image : 19.941744254435704
==========================================================================
===================================================
```

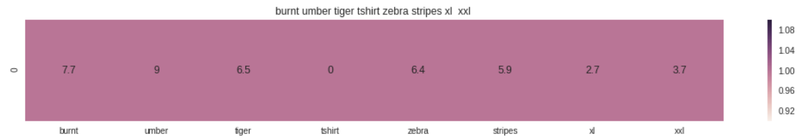
Model With brand , color and idf based Features with weighted pairwise similarities

In [0]:

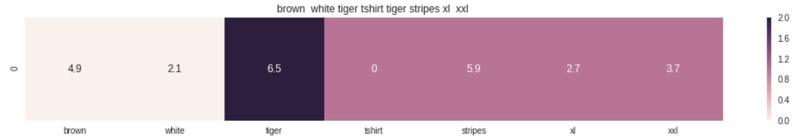
```
data['brand'].fillna(value="Not given", inplace=True )

# replace spaces with hypen
brands = [x.replace(" ", "-") for x in data['brand'].values]
types = [x.replace(" ", "-") for x in data['product_type_name'].values]
colors = [x.replace(" ", "-") for x in data['color'].values]

brand_vectorizer = CountVectorizer()
brand_features = brand_vectorizer.fit_transform(brands)

type_vectorizer = CountVectorizer()
type_features = type_vectorizer.fit_transform(types)

color_vectorizer = CountVectorizer()
color_features = color_vectorizer.fit_transform(colors)

extra_features = hstack((brand_features, type_features, color_features)).tocsr()
```

In [71]:

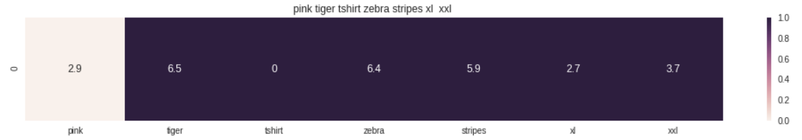
```
def idf_model(doc_id,w1,w2, num_results):

    idf_w2v_dist = pairwise_distances(idf_title_features,idf_title_features[doc_id])
    ex_feat_dist = pairwise_distances(extra_features, extra_features[doc_id])
    pairwise_dist   = (w1 * idf_w2v_dist +  w2 * ex_feat_dist)/float(w1 + w2)

    # np.argsort will return indices of 9 smallest distances
    indices = np.argsort(pairwise_dist.flatten())[0:num_results]
    #pdists will store the 9 smallest distances
    pdists  = np.sort(pairwise_dist.flatten())[0:num_results]

    #data frame indices of the 9 smallest distace's
    df_indices = list(data.index[indices])

    for i in range(0,len(indices)):
        get_result(indices[i],data['title'].loc[df_indices[0]], data['title'].loc[df_in
dices[i]], data['medium_image_url'].loc[df_indices[i]], 'idf')
        print('ASIN :',data['asin'].loc[df_indices[i]])
        print('Brand :',data['brand'].loc[df_indices[i]])
        print ('euclidean distance from the given image :', pdists[i])
        print('='*125)



idf_model(12566,1,5,20)
# in the output heat map each value represents the idf values of the label word, the co
lor represents the intersection with inputs title
```



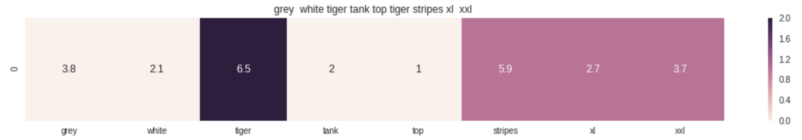
```
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 0.0
============================================================================
====================================================
```



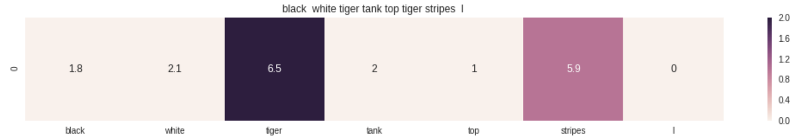
```
ASIN : B00JXQCWTO
Brand : Si Row
euclidean distance from the given image : 2.411326100329752
============================================================================
====================================================
```



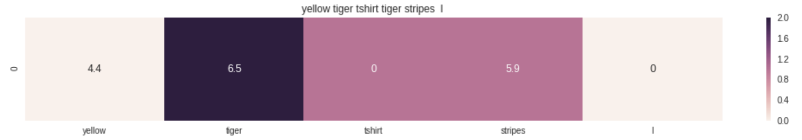
```
ASIN : B00JXQASS6
Brand : Si Row
euclidean distance from the given image : 3.2126133533826278
============================================================================
====================================================
```



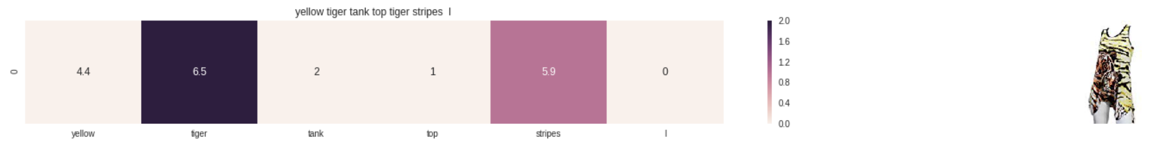
```
ASIN : B00JXQAFZ2
Brand : Si Row
euclidean distance from the given image : 3.583976987421412
============================================================================
====================================================
```



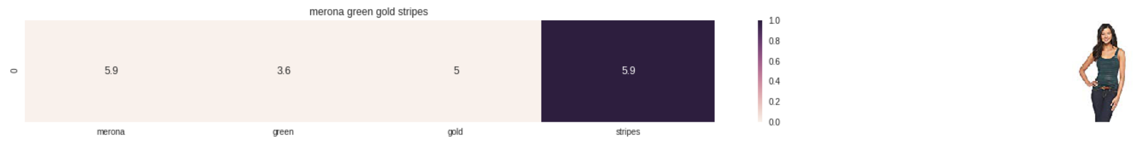
```
ASIN : B00JXQAO94
Brand : Si Row
euclidean distance from the given image : 3.6419481531768363
============================================================================
====================================================
```



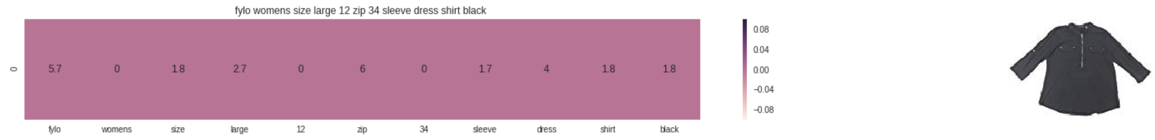
```
ASIN : B00JXQCUIC
Brand : Si Row
euclidean distance from the given image : 3.661569725563531
==========================================================================
===================================================
```



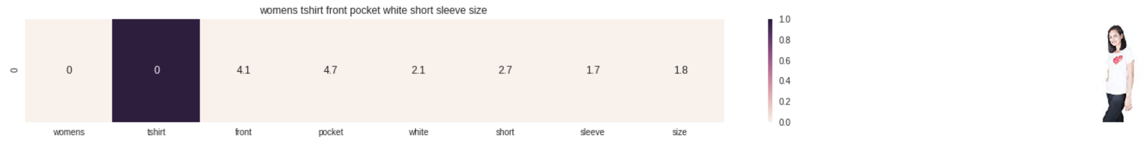
```
ASIN : B00JXQAUWA
Brand : Si Row
euclidean distance from the given image : 3.7073521765650326
==========================================================================
===================================================
```



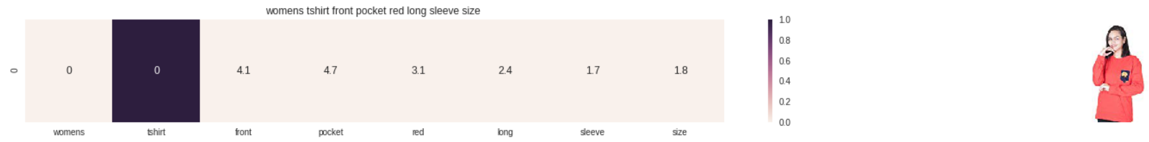
```
ASIN : B01KVZUB6G
Brand : Merona
euclidean distance from the given image : 4.851365812807567
==========================================================================
===================================================
```



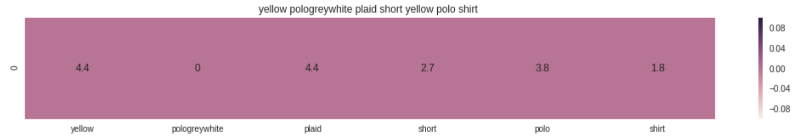
```
ASIN : B0718Y9J4M
Brand : f
euclidean distance from the given image : 4.943978221732228
==========================================================================
===================================================
```



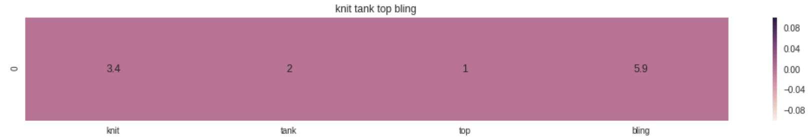
```
ASIN : B01JR73FSK
Brand : Lofbaz
euclidean distance from the given image : 4.950046024715346
==========================================================================
===================================================
```



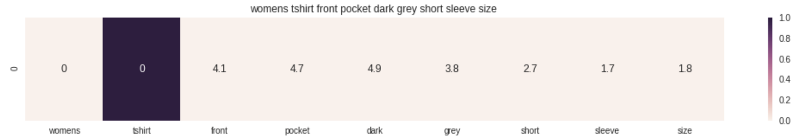
```
ASIN : B01JR72WHA
Brand : Lofbaz
euclidean distance from the given image : 4.965600324787029
==========================================================================
===================================================
```



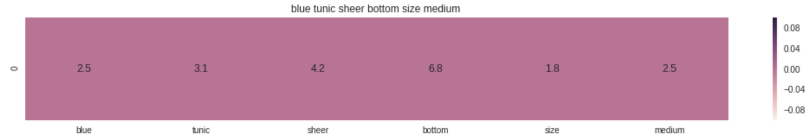
```
ASIN : B0755TBRM6
Brand : RuggedButts
euclidean distance from the given image : 5.021523245866848
==========================================================================
===================================================
```



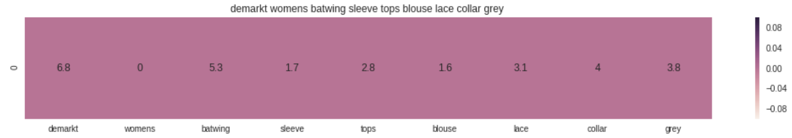
```
ASIN : B01NBQSBMN
Brand : Pink Cattlelac
euclidean distance from the given image : 5.085768088717692
==========================================================================
===================================================
```



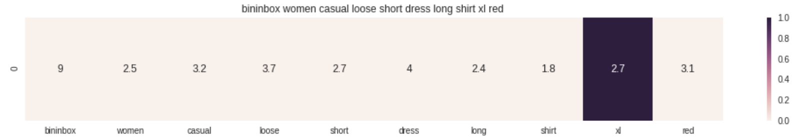
```
ASIN : B01JR73BMA
Brand : Lofbaz
euclidean distance from the given image : 5.100698484030582
==========================================================================
===================================================
```



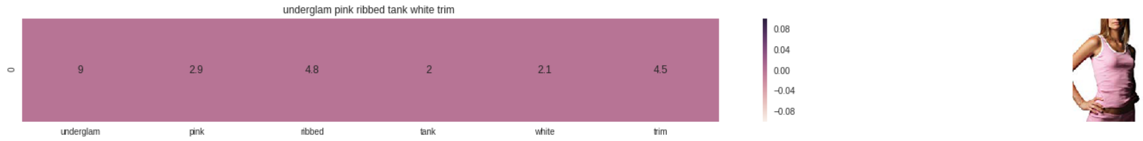
```
ASIN : B01NAOFEQE
Brand : Panhandle Slim
euclidean distance from the given image : 5.261031247514857
==========================================================================
===================================================
```



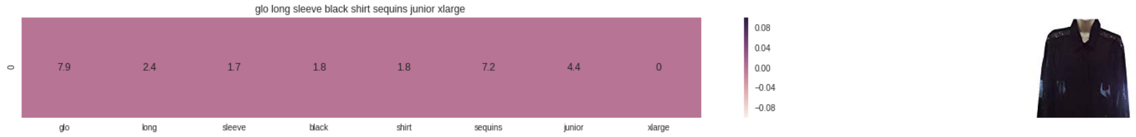
```
ASIN : B00VBAYU9U
Brand : Demarkt
euclidean distance from the given image : 5.2652592547015535
==========================================================================
===================================================
```



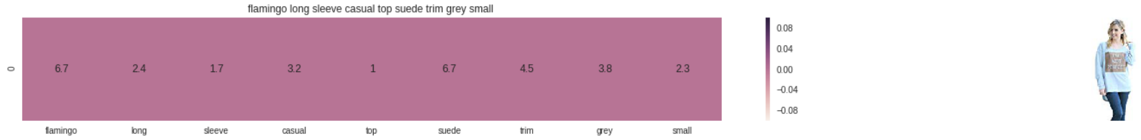
```
ASIN : B01BZXQ55O
Brand : BININBOX
euclidean distance from the given image : 5.285456384534895
===========================================================================
===================================================
```



```
ASIN : B008D4RMH4
Brand : Underglam
euclidean distance from the given image : 5.293480945897151
===========================================================================
===================================================
```



```
ASIN : B01NBLNC7J
Brand : Glo
euclidean distance from the given image : 5.326927007153286
===========================================================================
===================================================
```



```
ASIN : B017WSOZC6
Brand : FLAMINGO
euclidean distance from the given image : 5.329801624861757
===========================================================================
===================================================
```

Model With brand , color, idf based Features and Image vector Features(CNN) with weighted pairwise similarities

In [0]:

```
#load the features and corresponding ASINS info.
bottleneck_features_train = np.load('./drive/My Drive/Applied_AI_Workshop_Code_Data/16k_data_cnn_features.npy')
asins = np.load('./drive/My Drive/Applied_AI_Workshop_Code_Data/16k_data_cnn_feature_asins.npy')
asins = list(asins)

# load the original 16K dataset
#data = pd.read_pickle('pickels/16k_apperal_data_preprocessed')
df_asins = list(data['asin'])


from IPython.display import display, Image, SVG, Math, YouTubeVideo
```

In [0]:

```
def get_similar_products_cnn(doc_id,w1,w2,w3, num_results):
    doc_id = asins.index(df_asins[doc_id])
    image_pairwise_dist = pairwise_distances(bottleneck_features_train, bottleneck_features_train[doc_id].reshape(1,-1))
    idf_dist = pairwise_distances(idf_title_features,idf_title_features[doc_id])
    ex_feat_dist = pairwise_distances(extra_features, extra_features[doc_id])
    pairwise_dist   = (w1 * idf_dist +  w2 * ex_feat_dist+w3*image_pairwise_dist)/float(w1 + w2+w3)

    indices = np.argsort(pairwise_dist.flatten())[0:num_results]
    pdists  = np.sort(pairwise_dist.flatten())[0:num_results]
    df_indices = list(data.index[indices])
    for i in range(len(indices)):
        rows = data[['medium_image_url','title']].loc[data['asin']==asins[indices[i]]]
        for indx, row in rows.iterrows():
            display(Image(url=row['medium_image_url'], embed=True))
            print('Product Title: ', row['title'])

            #print('Euclidean Distance from input image:', pdists[i])
            print('Amazon Url: www.amzon.com/dp/'+ asins[indices[i]])
            print('ASIN :',data['asin'].loc[df_indices[i]])
            print('Brand :',data['brand'].loc[df_indices[i]])
            print ('euclidean distance from the given image :', pdists[i])
            print('='*125)
```

Equal Weighted Similarity Results

In [95]:

```
get_similar_products_cnn(12566,1,1,1 ,20)
```

```
Product Title:  burnt umber tiger tshirt zebra stripes xl  xxl
Amazon Url: www.amzon.com/dp/B00JXQB5FQ
ASIN : B01M0IDUCV
Brand : Premise
euclidean distance from the given image : 0.014731391022602717
==========================================================================
===================================================
```

```
Product Title:  abaday multicolor cartoon cat print short sleeve longline
shirt large
Amazon Url: www.amzon.com/dp/B01CR57YY0
ASIN : B06ZYLKPRT
Brand : Xhilaration
euclidean distance from the given image : 22.367041073517697
==========================================================================
===================================================
```

```
Product Title:  cute pastel tops tees colorful butterfly design print size

Amazon Url: www.amzon.com/dp/B019E3TD10
ASIN : B01MTW6DJS
Brand : Utopiat
euclidean distance from the given image : 22.41887639360203
==========================================================================
===================================================
```

```
Product Title:  leona lauren leonard womens pippa top black 0
Amazon Url: www.amzon.com/dp/B0721VLBS6
ASIN : B016P80OKQ
Brand : Studio M
euclidean distance from the given image : 22.431992710381497
==========================================================================
===================================================
```

```
Product Title:  j america 8138 womens glitter tshirt forest green silver 3
xl
Amazon Url: www.amzon.com/dp/B0719NLWSL
ASIN : B007N3WV6I
Brand : Forgot My Souvenirs
euclidean distance from the given image : 22.466754531908787
==========================================================================
===================================================
```

```
Product Title:  womens tops tees cute cartoon owl graphic print size
Amazon Url: www.amzon.com/dp/B01NGZ4Y3K
ASIN : B01L2ZTKFM
Brand : One Clothing
euclidean distance from the given image : 22.53985225410862
===========================================================================
===================================================
```

```
Product Title:  woman casual cotton tees dream believe achieve short sleev
e tshirt
Amazon Url: www.amzon.com/dp/B07548GLPB
ASIN : B073JWSM1V
Brand : Fuming
euclidean distance from the given image : 22.55836144744003
===========================================================================
===================================================
```

```
Product Title:  mossimo supply co womens ribbed tank top xxl dark green sp
arkle
Amazon Url: www.amzon.com/dp/B071NS5FGG
ASIN : B071X6MSL8
Brand : General
euclidean distance from the given image : 22.711787281100722
===========================================================================
===================================================
```

```
Product Title:  kingde star pink flower dog stamp sleeveless vestbqn24
Amazon Url: www.amzon.com/dp/B015H3W9BM
ASIN : B074MJPLCB
Brand : BollyDoll
euclidean distance from the given image : 22.725876705577424
===========================================================================
====================================================
```

```
Product Title:  kawaii cotton pastel tops tees pink flower design
Amazon Url: www.amzon.com/dp/B071P4YKH5
ASIN : B071KG15YM
Brand : KENDALL + KYLIE
euclidean distance from the given image : 22.76512912417461
===========================================================================
====================================================
```

```
Product Title:  ya los angeles womens ya los angeles striped knit size sma
ll eggplantgray
Amazon Url: www.amzon.com/dp/B06XG2ZV5J
ASIN : B01J72N9QI
Brand : Stoosh
euclidean distance from the given image : 22.79422862761032
===========================================================================
====================================================
```

```
Product Title:  pink tiger tshirt zebra stripes xl  xxl
Amazon Url: www.amzon.com/dp/B00JXQASS6
ASIN : B01N4NQ7LX
Brand : CeCe by Cynthia Steffe
euclidean distance from the given image : 22.863902382701365
==========================================================================
===================================================
```

```
Product Title:  energie white tank top sleeveless size xs nwt  movaz
Amazon Url: www.amzon.com/dp/B00Z8RY6EG
ASIN : B01LWUIZYJ
Brand : Premise
euclidean distance from the given image : 22.87764977006472
==========================================================================
===================================================
```

```
Product Title:  mossimo supply co womens ribbed tank top xlarge olive char
coal
Amazon Url: www.amzon.com/dp/B072P5XQCK
ASIN : B0746RVF6K
Brand : Eileen Fisher
euclidean distance from the given image : 22.87920307939814
==========================================================================
===================================================
```

```
Product Title:  vertvie womens short sleeve crew neck shirt letter print t
ee tops xl
Amazon Url: www.amzon.com/dp/B0722971MM
ASIN : B074DL2HQ4
Brand : Beulah
euclidean distance from the given image : 22.92914684940852
==========================================================================
===================================================
```

```
Product Title:  new military arms poland womens black short sleeve tshirt
Amazon Url: www.amzon.com/dp/B01K76A2W2
ASIN : B01HT0LM5U
Brand : Lushfox
euclidean distance from the given image : 22.95786435102904
==========================================================================
===================================================
```

```
Product Title:  short sleeve crew neck tee slits
Amazon Url: www.amzon.com/dp/B01NAAIH0W
ASIN : B071DH39DL
Brand : Mossimo
euclidean distance from the given image : 23.025331801627505
==========================================================================
===================================================
```

```
Product Title:  fjallraven  womens ovik tshirt plum xxl
Amazon Url: www.amzon.com/dp/B06XC3CZF6
ASIN : B06VWD17JS
Brand : John Paul Richard
euclidean distance from the given image : 23.044122089461187
==========================================================================
===================================================
```

```
Product Title:  banana republic womens paisley printed floral dolman vee t
op green xl
Amazon Url: www.amzon.com/dp/B06XNXRK6K
ASIN : B072N5BBBK
Brand : Merona
euclidean distance from the given image : 23.04845816792449
==========================================================================
===================================================
```

```
Product Title:  cauau47 womens irregular black longline paillette tshirt
Amazon Url: www.amzon.com/dp/B01G8WU8DM
ASIN : B01J0L63K0
Brand : FAPIZI
euclidean distance from the given image : 23.054244113771293
==========================================================================
===================================================
```

**Weights with
idf feature Wights =1
Brand and Colour =10
image vector=5 results :**

In [86]:

```
get_similar_products_cnn(12566,1,10,5 ,20)
```

```
Product Title:  burnt umber tiger tshirt zebra stripes xl  xxl
Euclidean Distance from input image: 0.013810679316520691
Amazon Url: www.amzon.com/dp/B00JXQB5FQ
ASIN : B01M0IDUCV
Brand : Premise
euclidean distance from the given image : 0.013810679316520691
==========================================================================
===================================================
```

```
Product Title:  pink tiger tshirt zebra stripes xl  xxl
Euclidean Distance from input image: 13.48699533429923
Amazon Url: www.amzon.com/dp/B00JXQASS6
ASIN : B01N4NQ7LX
Brand : CeCe by Cynthia Steffe
euclidean distance from the given image : 13.48699533429923
==========================================================================
===================================================
```

```
Product Title:  yellow tiger tshirt tiger stripes  l 
Euclidean Distance from input image: 16.146923372847393
Amazon Url: www.amzon.com/dp/B00JXQCUIC
ASIN : B01IU645VU
Brand : Outback Red
euclidean distance from the given image : 16.146923372847393
==========================================================================
===================================================
```

```
Product Title:  brown  white tiger tshirt tiger stripes xl  xxl 
Euclidean Distance from input image: 17.161239092893577
Amazon Url: www.amzon.com/dp/B00JXQCWTO
ASIN : B01FQLKKMK
Brand : SLJD
euclidean distance from the given image : 17.161239092893577
==========================================================================
===================================================
```

```
Product Title:  cute pastel tops tees colorful butterfly design print size 

Euclidean Distance from input image: 17.475982985989393
Amazon Url: www.amzon.com/dp/B019E3TD10
ASIN : B01MTW6DJS
Brand : Utopiat
euclidean distance from the given image : 17.475982985989393
==========================================================================
===================================================
```

```
Product Title:  chicago chicago 18 shirt women pink
Euclidean Distance from input image: 17.54887571211394
Amazon Url: www.amzon.com/dp/B01GXAZTRY
ASIN : B071VZCT5W
Brand : Chloe K.
euclidean distance from the given image : 17.54887571211394
==========================================================================
===================================================
```

```
Product Title:  red  pink floral heel sleeveless shirt xl  xxl
Euclidean Distance from input image: 17.611854398698142
Amazon Url: www.amzon.com/dp/B00JV63QQE
ASIN : B00L8RE3PC
Brand : JSDY-Cloth
euclidean distance from the given image : 17.611854398698142
==========================================================================
===================================================
```

```
Product Title:  womens thin style tops tees pastel watermelon print
Euclidean Distance from input image: 17.622546945122217
Amazon Url: www.amzon.com/dp/B01JUNHBRM
ASIN : B00K77AN5S
Brand : Russell Collection
euclidean distance from the given image : 17.622546945122217
==========================================================================
===================================================
```

```
Product Title:  abaday multicolor cartoon cat print short sleeve longline
shirt large
Euclidean Distance from input image: 17.63806896155311
Amazon Url: www.amzon.com/dp/B01CR57YY0
ASIN : B06ZYLKPRT
Brand : Xhilaration
euclidean distance from the given image : 17.63806896155311
==========================================================================
===================================================
```

```
Product Title:  kawaii pastel tops tees baby blue flower design
Euclidean Distance from input image: 17.665412703536425
Amazon Url: www.amzon.com/dp/B071SBCY9W
ASIN : B01MG83UB4
Brand : MaxMara
euclidean distance from the given image : 17.665412703536425
==========================================================================
===================================================
```

```
Product Title:  mossimo supply co womens ribbed tank top xlarge olive char
coal
Euclidean Distance from input image: 17.672579869303043
Amazon Url: www.amzon.com/dp/B072P5XQCK
ASIN : B0746RVF6K
Brand : Eileen Fisher
euclidean distance from the given image : 17.672579869303043
==========================================================================
===================================================
```

```
Product Title:  kingde star pink flower dog stamp sleeveless vestbqn24
Euclidean Distance from input image: 17.68723890885449
Amazon Url: www.amzon.com/dp/B015H3W9BM
ASIN : B074MJPLCB
Brand : BollyDoll
euclidean distance from the given image : 17.68723890885449
==========================================================================
===================================================
```

```
Product Title:  adults cotton custom sesame street live family v neck shir
t black xxl
Euclidean Distance from input image: 17.706910192153202
Amazon Url: www.amzon.com/dp/B01LWTSLVC
ASIN : B01I2PK9GE
Brand : GRXBRS
euclidean distance from the given image : 17.706910192153202
==========================================================================
===================================================
```

```
Product Title:  ya los angeles womens ya los angeles striped knit size sma
ll eggplantgray
Euclidean Distance from input image: 17.73667677644798
Amazon Url: www.amzon.com/dp/B06XG2ZV5J
ASIN : B01J72N9QI
Brand : Stoosh
euclidean distance from the given image : 17.73667677644798
==========================================================================
===================================================
```

```
Product Title:  miss chievous juniors striped peplum tank top medium shado
wpeach
Euclidean Distance from input image: 17.741362274063157
Amazon Url: www.amzon.com/dp/B0177DM70S
ASIN : B01MXMG6KB
Brand : Mogul Interior
euclidean distance from the given image : 17.741362274063157
==========================================================================
===================================================
```

```
Product Title:  five finger death punch womens pink print 2014 tour girls
jr soft tee black
Euclidean Distance from input image: 17.809225514030604
Amazon Url: www.amzon.com/dp/B0148ROP3S
ASIN : B074337SFR
Brand : Sunhouse
euclidean distance from the given image : 17.809225514030604
==========================================================================
===================================================
```

```
Product Title:  cauau47 womens irregular black longline paillette tshirt
Euclidean Distance from input image: 17.81796904777894
Amazon Url: www.amzon.com/dp/B01G8WU8DM
ASIN : B01J0L63K0
Brand : FAPIZI
euclidean distance from the given image : 17.81796904777894
==========================================================================
===================================================
```

```
Product Title:  mossimo supply co womens ribbed tank top xxl dark green sp
arkle
Euclidean Distance from input image: 17.839375457082898
Amazon Url: www.amzon.com/dp/B071NS5FGG
ASIN : B071X6MSL8
Brand : General
euclidean distance from the given image : 17.839375457082898
===========================================================================
===================================================
```

```
Product Title:  fifth degree women short sleeve rhinestone printed tops ca
sual shirt
Euclidean Distance from input image: 17.840319600765362
Amazon Url: www.amzon.com/dp/B01M8I9VJJ
ASIN : B011TZQZ8K
Brand : ZEKO
euclidean distance from the given image : 17.840319600765362
===========================================================================
===================================================
```

```
Product Title:  kawaii cotton pastel tops tees pink flower design
Euclidean Distance from input image: 17.873181910412853
Amazon Url: www.amzon.com/dp/B071P4YKH5
ASIN : B071KG15YM
Brand : KENDALL + KYLIE
euclidean distance from the given image : 17.873181910412853
===========================================================================
===================================================
```

Trying out the same Weights for other items

In [87]:

```
get_similar_products_cnn(1256,1,10,5 ,20)
```

```
Product Title:  acting pro womens sassy since birth print racerback tank t
op medium pink
Euclidean Distance from input image: 0.013810679316520691
Amazon Url: www.amzon.com/dp/B01I2ZZ93C
ASIN : B06XYTF99Z
Brand : Genie
euclidean distance from the given image : 0.013810679316520691
==========================================================================
===================================================
```

```
Product Title:  nella fantasia womens owl print tank top small peach
Euclidean Distance from input image: 14.062512991287894
Amazon Url: www.amzon.com/dp/B01I2ZZC16
ASIN : B01BU802XA
Brand : Flores
euclidean distance from the given image : 14.062512991287894
==========================================================================
===================================================
```

```
Product Title:  women yabish print white sleeveless crop top
Euclidean Distance from input image: 15.228479989004835
Amazon Url: www.amzon.com/dp/B0748JNFL9
ASIN : B074KD6ZCP
Brand : ClothingLoves
euclidean distance from the given image : 15.228479989004835
==========================================================================
===================================================
```

```
Product Title:  crop tops women fashion sexy character vest casual tshirt
tank top
Euclidean Distance from input image: 15.549940567950664
Amazon Url: www.amzon.com/dp/B0107UEPVM
ASIN : B01GU92OPI
Brand : Brooks
euclidean distance from the given image : 15.549940567950664
==========================================================================
===================================================
```

```
Product Title:  women keep swimming print sleeveless crop top
Euclidean Distance from input image: 15.800180127832334
Amazon Url: www.amzon.com/dp/B0749CCCY4
ASIN : B074P9YR8S
Brand : Ramy Brook
euclidean distance from the given image : 15.800180127832334
==========================================================================
===================================================
```

```
Product Title:  drew womens beck racer back layered hem jersey top sz whit
e 230034f
Euclidean Distance from input image: 15.890111264908372
Amazon Url: www.amzon.com/dp/B01GSJZUGU
ASIN : B01HXCS9BO
Brand : Bigban
euclidean distance from the given image : 15.890111264908372
==========================================================================
===================================================
```

```
Product Title:  nella fantasia womens gypsy elephant racerback tank top me
dium black
Euclidean Distance from input image: 16.061468168868956
Amazon Url: www.amzon.com/dp/B01IJRD8OA
ASIN : B01MXG0FNQ
Brand : Tosangn
euclidean distance from the given image : 16.061468168868956
==========================================================================
===================================================
```

```
Product Title:  nella fantasia womens gypsy spirit anchor racerback tank t
op large black
Euclidean Distance from input image: 16.084611823561726
Amazon Url: www.amzon.com/dp/B01IJRE312
ASIN : B01I4A8T3M
Brand : Non Branded
euclidean distance from the given image : 16.084611823561726
==========================================================================
===================================================
```

```
Product Title:  bjorn borg womens solid wrestling tank top xlarge black
Euclidean Distance from input image: 16.101369591725685
Amazon Url: www.amzon.com/dp/B00W48DEA4
ASIN : B073WKFKLZ
Brand : Sanjoy
euclidean distance from the given image : 16.101369591725685
==========================================================================
===================================================
```

```
Product Title:  woman casual cotton tees dream believe achieve short sleev
e tshirt
Euclidean Distance from input image: 16.16409850625322
Amazon Url: www.amzon.com/dp/B07548GLPB
ASIN : B073JWSM1V
Brand : Fuming
euclidean distance from the given image : 16.16409850625322
==========================================================================
===================================================
```

```
Product Title:  women pattern 8 cute baby alien print sleeveless crop top
Euclidean Distance from input image: 16.190974287183575
Amazon Url: www.amzon.com/dp/B074BNJM8S
ASIN : B01JLSSCRY
Brand : LEEMASTER
euclidean distance from the given image : 16.190974287183575
==========================================================================
===================================================
```

```
Product Title:  couthclothing womens wolf racerback junior tank top charco
al black
Euclidean Distance from input image: 16.217578481193737
Amazon Url: www.amzon.com/dp/B06XVGH2VW
ASIN : B01G8N82KW
Brand : BRMWs
euclidean distance from the given image : 16.217578481193737
==========================================================================
===================================================
```

```
Product Title:  women three wise monkeys emoji print sleeveless crop top
Euclidean Distance from input image: 16.23024439049667
Amazon Url: www.amzon.com/dp/B074VPC98H
ASIN : B06Y3CKDML
Brand : Eileen Fisher
euclidean distance from the given image : 16.23024439049667
==========================================================================
===================================================
```

```
Product Title:  jm collection womens plus ombre shutter pleat casual top w
hite 1x
Euclidean Distance from input image: 16.234948123456803
Amazon Url: www.amzon.com/dp/B01H456MU0
ASIN : B074MHV9GX
Brand : MSK
euclidean distance from the given image : 16.234948123456803
==========================================================================
===================================================
```

```
Product Title:  baomabao women tank tops letter print sleeveless blouse sm
all white
Euclidean Distance from input image: 16.263431948715393
Amazon Url: www.amzon.com/dp/B01EW93U7O
ASIN : B00W3MMKS8
Brand : HEYFAIR
euclidean distance from the given image : 16.263431948715393
==========================================================================
===================================================
```

```
Product Title:  women quotes boys print white sleeveless crop top
Euclidean Distance from input image: 16.303613399159314
Amazon Url: www.amzon.com/dp/B0748CKWF3
ASIN : B01L79BFYC
Brand : Namnoi Clothing Store
euclidean distance from the given image : 16.303613399159314
==========================================================================
===================================================
```

```
Product Title:  fashion crop tops women casual summer emoji sexy lady girl
shirt hipster tank top
Euclidean Distance from input image: 16.3243421179051
Amazon Url: www.amzon.com/dp/B010V3B44G
ASIN : B071W8XRB2
Brand : Olivia Moon
euclidean distance from the given image : 16.3243421179051
==========================================================================
===================================================
```

```
Product Title:  fornarina womens manu bis sequin accent halter top sz smal
l black
Euclidean Distance from input image: 16.3266764109029
Amazon Url: www.amzon.com/dp/B00BKB3VT0
ASIN : B01AFL5WTW
Brand : Absolutely
euclidean distance from the given image : 16.3266764109029
==========================================================================
===================================================
```

```
Product Title:  women pattern six three aliens printed white sleeveless cr
op top
Euclidean Distance from input image: 16.367408617081797
Amazon Url: www.amzon.com/dp/B01MRFOU3R
ASIN : B074TVZB9L
Brand : Bobeau
euclidean distance from the given image : 16.367408617081797
==========================================================================
===================================================
```

```
Product Title:  grab life joystick gray cami tank top shirt small
Euclidean Distance from input image: 16.36818105275248
Amazon Url: www.amzon.com/dp/B01MRF2LPP
ASIN : B06VSDV771
Brand : Soprano
euclidean distance from the given image : 16.36818105275248
==========================================================================
===================================================
```

Similar items for item 500

In [97]:

```
get_similar_products_cnn(500,1,10,5 ,20)
```

```
Product Title:  alo sport ladies bamboo racerback tank w2006leafslatexl
Amazon Url: www.amzon.com/dp/B0023UNW7I
ASIN : B01G4OEW1S
Brand : FOCUST
euclidean distance from the given image : 0.009765625
==========================================================================
===================================================
```

```
Product Title:  alo sport  ladies racerback bamboo tank
Amazon Url: www.amzon.com/dp/B003IWOLYS
ASIN : B073ZC75WJ
Brand : Focal20
euclidean distance from the given image : 7.8089468854694175
==========================================================================
===================================================
```

```
Product Title:  alo sport ladies bamboo racerback tank  pinkwhite  xs
Amazon Url: www.amzon.com/dp/B004J8LKP8
ASIN : B01CH48FVC
Brand : FIFTEEN TWENTY
euclidean distance from the given image : 9.825102403542804
==========================================================================
===================================================
```

```
Product Title:  alo sport womens 3button mesh polo shirt sport crlna blue
medium
Amazon Url: www.amzon.com/dp/B00IM7XQ40
ASIN : B072FTMQ3S
Brand : Alfani
euclidean distance from the given image : 15.478399945074756
===========================================================================
===================================================
```

```
Product Title:  alo ladies junior fit performance mesh polo shirt w1709 la
rge sport athletic gold
Amazon Url: www.amzon.com/dp/B00PH3DJC6
ASIN : B06XXWPSMC
Brand : 10 Crosby Derek Lam
euclidean distance from the given image : 15.690350461575507
===========================================================================
===================================================
```

```
Product Title:  fruit loom ladies 100 heavy cotton hd tshirt xl purple
Amazon Url: www.amzon.com/dp/B014WBV6E6
ASIN : B00VZD9W46
Brand : New Balance
euclidean distance from the given image : 15.713161303211912
===========================================================================
===================================================
```

```
Product Title:  district made  ladies modal blend tank dm481 white 2xl
Amazon Url: www.amzon.com/dp/B00KC6OZQC
ASIN : B0719R5YZ5
Brand : Krisa
euclidean distance from the given image : 15.964950177740706
==========================================================================
===================================================
```

```
Product Title:  sugarlips womens relaxed fit seamless ribbed tank skin nud
e
Amazon Url: www.amzon.com/dp/B00IJHSY54
ASIN : B071RQKPFK
Brand : BCX
euclidean distance from the given image : 16.026504782073108
==========================================================================
===================================================
```

```
Product Title:  lole womens rhea tank top black tank top
Amazon Url: www.amzon.com/dp/B01N4ATA6H
ASIN : B01INUM5U6
Brand : Current / Elliott
euclidean distance from the given image : 16.058412032598742
==========================================================================
===================================================
```

```
Product Title:  comfort colors ladies 54 oz ringspun longsleeve tshirts bu
tter c3014
Amazon Url: www.amzon.com/dp/B00390D6FY
ASIN : B01N5OKGNQ
Brand : Fjällräven
euclidean distance from the given image : 16.169521094865587
==========================================================================
===================================================
```

```
Product Title:  district juniors vintage wash vneck tee4xl deep turquoise
dt4501
Amazon Url: www.amzon.com/dp/B00TSNVHZC
ASIN : B071CMN66J
Brand : A.L.C.
euclidean distance from the given image : 16.30048627798388
==========================================================================
===================================================
```

```
Product Title:  district juniors vintage wash vneck teem black dt4501
Amazon Url: www.amzon.com/dp/B00TSNTQI2
ASIN : B01M0IHJJE
Brand : West Kei
euclidean distance from the given image : 16.4019352040565
==========================================================================
===================================================
```

```
Product Title:  authentic pigment ladies true spirit raglan tshirt smoke x
xlarge
Amazon Url: www.amzon.com/dp/B01GESXYTU
ASIN : B074QV3HFZ
Brand : Chloe K.
euclidean distance from the given image : 16.453012964667067
===========================================================================
===================================================
```

```
Product Title:  miraclebody womens jersey slimming tunic top black
Amazon Url: www.amzon.com/dp/B0059GPDDE
ASIN : B01GESXRTC
Brand : Authentic Pigment
euclidean distance from the given image : 16.466493225904806
===========================================================================
===================================================
```

```
Product Title:  comfort colors ladies 54 oz ringspun longsleeve tshirtl la
goon blue c3014
Amazon Url: www.amzon.com/dp/B00390KELS
ASIN : B007C0HVRQ
Brand : FeatherLite
euclidean distance from the given image : 16.52349008675032
===========================================================================
===================================================
```

```
Product Title:  comfort colors womens ribbed collar longsleeve tshirt lago
on blue xlarge
Amazon Url: www.amzon.com/dp/B00390IRLM
ASIN : B01J9DRTDO
Brand : FIG Clothing
euclidean distance from the given image : 16.544279049828088
==========================================================================
===================================================
```

```
Product Title:  alo ladies performance threebutton polo shirt  sport royal
medium
Amazon Url: www.amzon.com/dp/B01GESYBOM
ASIN : B0758ZB8WP
Brand : Dantelle
euclidean distance from the given image : 16.546161013361985
==========================================================================
===================================================
```

```
Product Title:  lat apparel womens combed ringspun jersey longsleeve tshir
t3588royal2xl
Amazon Url: www.amzon.com/dp/B019MT215Q
ASIN : B019JKKPRO
Brand : Namnoi Cute Tee Top
euclidean distance from the given image : 16.599378874430567
==========================================================================
===================================================
```

```
Product Title:  lat ladies combed ringspun jersey longsleeve tshirt heathe
r xxxlarge
Amazon Url: www.amzon.com/dp/B007C3JXXS
ASIN : B0745J9HNS
Brand : Almost Famous
euclidean distance from the given image : 16.68157110042954
===========================================================================
===================================================
```

```
Product Title:  nili lotan womens normandy blouse black xsmall
Amazon Url: www.amzon.com/dp/B0736D8BVV
ASIN : B01G9RV9D4
Brand : CAUAU47
euclidean distance from the given image : 16.70386295735334
===========================================================================
===================================================
```

Report

Procedure

Product recommondation to users based on amazon's API data

Claeaning the data and removing near duplicate items using the words in the text for each item

Featurizing data using BagOfWords, Tfidf, Idf and computing Pairwise simalirities

Using Text Semantics (Word to vec, Tfidf Word to vec, Idf Word to vec) based product similarity

Using More Features Such as Color, Brand, Type along with BOW,TFIDF,IDF and Text Semantics to out Pairwise simalirities

Vectorizing the image using(CNN) with bottleneck features of pretrained VGG-16

Finally using IDF, Color, Brand, Image Vector computing Pairwise Weightedsimalirities

Conclusion

Idf feature distance Weights =1
Brand and Colour distance Weights =10
Image vector distance Weights=5

In [ ]: