Amazon Apparel Recommendations

[4.2] Data and Code:

https://drive.google.com/open?id=0BwNkduBnePt2VWhCYXhMV3p4dTg

[4.3] Overview of the data

In [1]:

```
#import all the necessary packages.

from PIL import Image
import requests
from io import BytesIO
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import warnings
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
import math
import time
import re
import os
import seaborn as sns
from collections import Counter
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics import pairwise_distances
from matplotlib import gridspec
from scipy.sparse import hstack
import plotly
import plotly.figure_factory as ff
from plotly.graph_objs import Scatter, Layout

plotly.offline.init_notebook_mode(connected=True)
warnings.filterwarnings("ignore")
```

In [14]:

```
# we have give a json file which consists of all information about
# the products
# loading the data using pandas' read_json file.
data = pd.read_json('tops_fashion.json')
```

In [15]:

```
print ('Number of data points : ', data.shape[0], \
       'Number of features/variables:', data.shape[1])
```

```
Number of data points :  183138 Number of features/variables: 19
```

Terminology:

What is a dataset?
Rows and columns
Data-point
Feature/variable

In [16]:

```
# each product/item has 19 features in the raw dataset.
data.columns # prints column-names or feature-names.
```

Out[16]:

```
Index(['asin', 'author', 'availability', 'availability_type', 'brand', 'color',
       'editorial_reivew', 'editorial_review', 'formatted_price',
       'large_image_url', 'manufacturer', 'medium_image_url', 'model',
       'product_type_name', 'publisher', 'reviews', 'sku', 'small_image_url',
       'title'],
      dtype='object')
```

Of these 19 features, we will be using only 6 features in this workshop.

1. asin (Amazon standard identification number)
2. brand (brand to which the product belongs to)
3. color (Color information of apparel, it can contain many colors as a value ex: red and black stripes)
4. product_type_name (type of the apperal, ex: SHIRT/TSHIRT)
5. medium_image_url (url of the image)
6. title (title of the product.)
7. formatted_price (price of the product)

In [0]:

```
data = data[['asin', 'brand', 'color', 'medium_image_url', 'product_type_name', 'title', 'formatted_price']]
```

In [0]:

```
print ('Number of data points : ', data.shape[0], \
       'Number of features:', data.shape[1])
data.head() # prints the top rows in the table.
```

```
Number of data points :  183138 Number of features: 7
```

Out[0]:

	asin	brand	color	medium_image_url	product_type_name	title	formatted_price
0	B016I2TS4W	FNC7C	None	https://images-na.ssl-images-amazon.com/images...	SHIRT	Minions Como Superheroes Ironman Long Sleeve R...	None
1	B01N49AI08	FIG Clothing	None	https://images-na.ssl-images-amazon.com/images...	SHIRT	FIG Clothing Womens Izo Tunic	None
2	B01JDPCOHO	FIG Clothing	None	https://images-na.ssl-images-amazon.com/images...	SHIRT	FIG Clothing Womens Won Top	None
3	B01N19U5H5	Focal18	None	https://images-na.ssl-images-amazon.com/images...	SHIRT	Focal18 Sailor Collar Bubble Sleeve Blouse Shi...	None
4	B004GSI2OS	FeatherLite	Onyx Black/ Stone	https://images-na.ssl-images-amazon.com/images...	SHIRT	Featherlite Ladies' Long Sleeve Stain Resistan...	$26.26

[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 index, row in
images.iterrows():\n        url = row['large_image_url']\n        response = requests.get(url)\n
img = Image.open(BytesIO(response.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

[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	formatted_price
4	B004GSI2OS	FeatherLite	Onyx Black/ Stone	https://images-na.ssl-images-amazon.com/images...	SHIRT	Featherlite Ladies' Long Sleeve Stain Resistan...	$26.26
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...	$9.99
11	B001LOUGE4	Fitness Etc.	Black	https://images-na.ssl-images-amazon.com/images...	SHIRT	Ladies Cotton Tank 2x1 Ribbed Tank Top	$11.99
15	B003BSRPB0	FeatherLite	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	FeatherLite Ladies' Moisture Free Mesh Sport S...	$20.54
21	B014ICEDNA	FNC7C	Purple	https://images-na.ssl-images-amazon.com/images...	SHIRT	Supernatural Chibis Sam Dean And Castiel Short...	$7.50

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	title	formatted_price
61973	B06Y1KZ2WB	Éclair	Black/Pink	https://images-na.ssl-images-amazon.com/images...	SHIRT	Éclair Women's Printed Thin Strap Blouse Black...	$24.99
133820	B010RV33VE	xiaoming	Pink	https://images-na.ssl-images-amazon.com/images...	SHIRT	xiaoming Womens Sleeveless Loose Long T-shirts...	$18.19
81461	B01DDSDLNS	xiaoming	White	https://images-na.ssl-images-	SHIRT	xiaoming Women's White	$21.58

	asin	brand	color	medium_image_url	product_type_name	title	formatted_price
75995	B00X5LYO9Y	xiaoming	Red Anchors	https://images-na.ssl-images-amazon.com/images...	SHIRT	xiaoming Stripes Tank Patch/Bear Sleeve Anchor...	$15.91
151570	B00WPJG35K	xiaoming	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	xiaoming Sleeve Sheer Loose Tassel Kimono Woma...	$14.32

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 Print Head
Shirt for woman Neon Wolf t-shirt
62. psychedelic colorful Howling Galaxy Wolf T-shirt/Colorful Rainbow Animal Print Head
Shirt for woman Neon Wolf t-shirt
63. psychedelic colorful Howling Galaxy Wolf T-shirt/Colorful Rainbow Animal Print Head
Shirt for woman Neon Wolf t-shirt
64. psychedelic colorful Howling Galaxy Wolf T-shirt/Colorful Rainbow Animal Print 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', 'Diam
onds', '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', 'Queen', '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 considering it as t
hose two apperals are different
        # if the number of words in which both strings differ are < 2 , we are considering it as t
hose 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]])


            # 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.Then, 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, XX-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', 'Diam
onds', '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', 'Queen', 'of', '
Diamonds', 'Women's', 'Shirt', 'X-Large']

        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 < 3 , we are considering it as t
hose 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.
```

6. Text pre-processing

In [30]:

```
data = pd.read_pickle('pickels/16k_apperal_data')

# NLTK download stop words. [RUN ONLY ONCE]
# goto Terminal (Linux/Mac) or Command-Prompt (Window)
# In the temrinal, type these commands
# $import nltk
# $nltk.download()
```

In [31]:

```
# we use the list of stop words that are downloaded from nltk lib.
stop_words = set(stopwords.words('english'))
print ('list of stop words:', stop_words)

def nlp_preprocessing(total_text, index, column):
    if type(total_text) is not int:
        string = ""
        for words in total_text.split():
            # remove the special chars in review like '"#$@!%^&*()_+-~?>< etc.
            word = ("".join(e for e in words if e.isalnum()))
            # Conver all letters to lower-case
            word = word.lower()
            # stop-word removal
            if not word in stop_words:
                string += word + " "
        data[column][index] = string
```

```
list of stop words: {'between', 'all', 'having', 'd', 'are', 'am', "hadn't", "wasn't", "wouldn't",
'yourself', 'than', 'has', 'shouldn', 'she', 'an', 'to', 'both', "that'll", 'my', 'herself',
'ain', 'against', 'doesn', 'their', "should've", 'needn', "you'd", 'too', 'doing', 'and', 'y', 're
', "aren't", 'me', "couldn't", 'shan', "didn't", "weren't", 'by', 'very', 'll', 'was', 'isn',
'as', 'you', 'can', 'when', 'our', 'i', 'who', 'below', 'the', 'not', 'should', 'ma', 'weren', 'of
', 'were', 'at', "don't", 'aren', 'any', 'is', 'its', "it's", 't', 'couldn', "she's", 'there', 'ou
t', 'more', 'didn', 'haven', 'mightn', 'in', 'few', 'his', 'above', 'nor', 'whom', 'own', 'up', 'm
', 'it', 'we', 'over', 'they', 'being', 'so', 'hadn', 'then', 'most', "haven't", 'with', "you've",
'because', 'do', 'for', 'don', "isn't", 'had', 'hers', 'which', 'myself', 'no', 'be', 'after', 've
', 'same', 'into', "hasn't", 'under', 'further', "needn't", 'them', 'wasn', 'ours', "won't", 'on',
'did', 'a', 'about', 'will', 'these', "shouldn't", 'off', 'if', 'theirs', 'this', 'such', 'have',
'during', 'o', 'that', 'how', 'itself', 'he', 'some', "shan't", 's', 'wouldn', 'why', 'only', 'fro
m', 'your', 'through', 'hasn', 'themselves', 'mustn', 'been', 'her', 'just', 'now', "mightn't", 'w
hat', 'those', 'or', 'won', "you'll", 'does', 'down', 'here', 'but', 'again', 'himself', 'other',
"doesn't", 'where', 'ourselves', 'yours', 'until', 'once', 'each', 'yourselves', "mustn't",
'while', 'him', "you're", 'before'}
```

In [32]:

```
start_time = time.clock()
# we take each title and we text-preprocess it.
for index, row in data.iterrows():
    nlp_preprocessing(row['title'], index, 'title')
# we print the time it took to preprocess whole titles
print(time.clock() - start_time, "seconds")
```

```
8.066374004052705 seconds
```

In [33]:

```
data.head()
```

Out[33]:

	asin	brand	color	medium_image_url	product_type_name	title	formatted_price
4	B004GSI2OS	FeatherLite	Onyx Black/ Stone	https://images-na.ssl-images-amazon.com/images...	SHIRT	featherlite ladies long sleeve stain resistant...	$26.26
6	B012YX2ZPI	HX-Kingdom Fashion T-	White	https://images-na.ssl-images-	SHIRT	womens unique 100 cotton	$9.99

	asin	shirts brand	color	amazon medium_image_url	product_type_name	special olympics wor... title	formatted_price
15	B003BSRPB0	FeatherLite	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	featherlite ladies moisture free mesh sport sh...	$20.54
27	B014ICEJ1Q	FNC7C	Purple	https://images-na.ssl-images-amazon.com/images...	SHIRT	supernatural chibis sam dean castiel neck tshi...	$7.39
46	B01NACPBG2	Fifth Degree	Black	https://images-na.ssl-images-amazon.com/images...	SHIRT	fifth degree womens gold foil graphic tees jun...	$6.95

In [34]:

```
data.to_pickle('pickels/16k_apperal_data_preprocessed')
```

Stemming

In [35]:

```
from nltk.stem.porter import *
stemmer = PorterStemmer()
print(stemmer.stem('arguing'))
print(stemmer.stem('fishing'))


# We tried using stemming on our titles and it didnot work very well.
```

```
argu
fish
```

[8] Text based product similarity

In [15]:

```
#load the features and corresponding ASINS info.
bottleneck_features_train = np.load('16k_data_cnn_features.npy')
asins = np.load('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'])

df_asin=pd.DataFrame(asins,columns=['asin'])
df1 = data.set_index('asin')
df1 = df1.reindex(index=df_asin['asin'])
data = df1.reset_index()
data.head()
```

Out[15]:

	asin	brand	color	medium_image_url	product_type_name	title	formatted_price
0	0000000060	Jofit	White	https://images-na.ssl-images-amazon.com/images...	BUILDING_MATERIAL	jofit womens jo tech polo white extra large	$25.99
1	6042589113	Dress on Trend	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	dress trend womens lion chain print baggy over...	$16.00
						crazy womens floral	

	asin	brand	color	medium_image_url	product_type_name	title	formatted_price
2	6342521018	Crazy	Beige	https://images-na.ssl-images-amazon.com/images...	SHIRT	sleeveless vintage crochet...	$9.99
3	B00029I0Z6	Roper	Red 2	https://images-na.ssl-images-amazon.com/images...	SHIRT	roper womens ls stars stripes pieced flag red ...	$48.78
4	B0006LTFTK	Ideology	Noir	https://images-na.ssl-images-amazon.com/images...	PANTS	ideology womens graphic training leggings noir...	$49.50

In [16]:

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

#Display an image
def display_img(url,ax,fig):
    # we get the url of the apparel and download it
    response = requests.get(url)
    img =PIL.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 we 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 the 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 difference 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 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))

    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 corpus
            # tfidf_title_features[doc_id, index_of_word_in_corpus] will give the tfidf value of w
rd in given document (doc_id)
            if x in  tfidf_title_vectorizer.vocabulary_:
                labels.append(tfidf_title_features[doc_id, tfidf_title_vectorizer.vocabulary_[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 value of word
in given document (doc_id)
            if x in  idf_title_vectorizer.vocabulary_:
                labels.append(idf_title_features[doc_id, idf_title_vectorizer.vocabulary_[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 returns 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)
```

[8.2] Bag of Words (BoW) on product titles.

In [17]:

```
from sklearn.feature_extraction.text import CountVectorizer
title_vectorizer = CountVectorizer()
title_features   = title_vectorizer.fit_transform(data['title'])
```

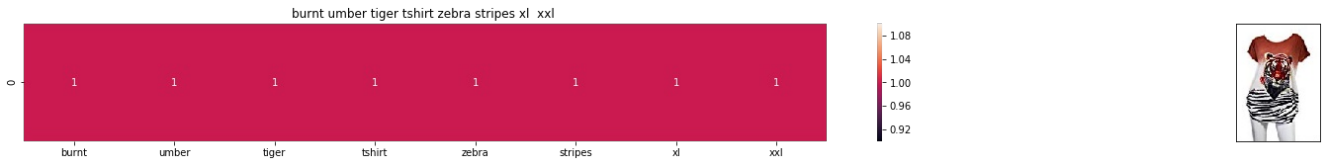
```
title_features = title_vectorizer.fit_transform(data['title'])
title_features.get_shape() # get number of rows and columns in feature matrix.
# title_features.shape = #data_points * #words_in_corpus
# CountVectorizer().fit_transform(corpus) returns
# the a sparase matrix of dimensions #data_points * #words_in_corpus

# What is a sparse vector?

# title_features[doc_id, index_of_word_in_corpus] = number of times the word occured in that doc
```

Out[17]:

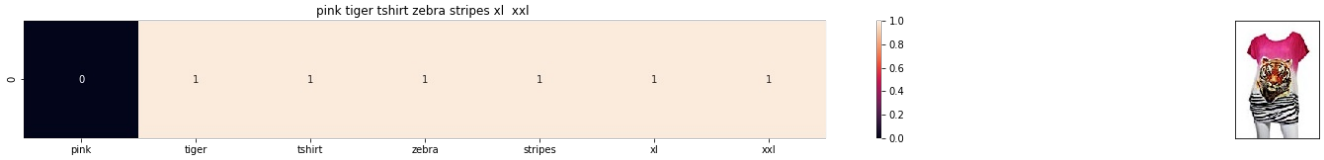
```
(16042, 12609)
```

In [24]:

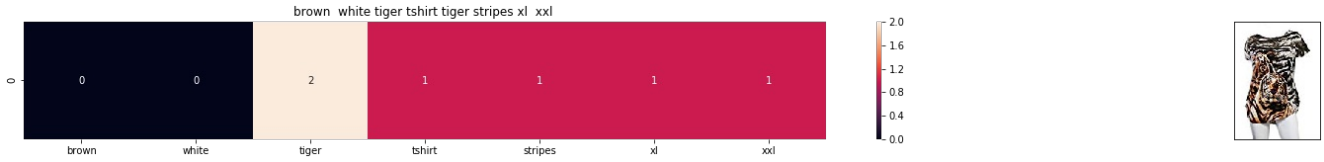
```
# Utility Functions which we will use through the rest of the workshop.
import PIL.Image

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

    # np.argsort will return indices of the smallest distances
    indices = np.argsort(pairwise_dist.flatten())[0:num_results]
    #pdists will store the 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)):
        # we will pass 1. doc_id, 2. title1, 3. title2, url, model
        get_result(indices[i],data['title'].loc[df_indices[0]], data['title'].loc[df_indices[i]], d
ata['medium_image_url'].loc[df_indices[i]], 'bag_of_words')
        print('ASIN :',data['asin'].loc[df_indices[i]])
        print ('Brand:', data['brand'].loc[df_indices[i]])
        print ('Title:', data['title'].loc[df_indices[i]])
        print ('Euclidean similarity with the query image :', pdists[i])
        print('='*60)

#call the bag-of-words model for a product to get similar products.
bag_of_words_model(1416, 20) # change the index if you want to.
# In the output heat map each value represents the count value
# of the label word, the color represents the intersection
# with inputs title.

#try 12566
#try 931
```




```
ASIN : B00JXQB5FQ
Brand: Si Row
Title: burnt umber tiger tshirt zebra stripes xl  xxl
Euclidean similarity with the query image : 0.0
============================================================
```

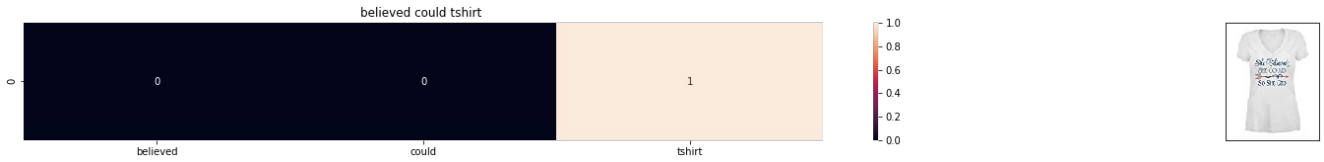


ASIN : B00JXQB5FQ
Brand: Si Row
Title: pink tiger tshirt zebra stripes xl xxl
Euclidean similarity with the query image : 1.7320508075688772
==

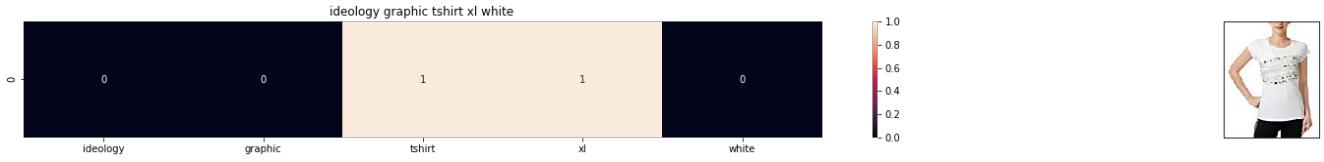


ASIN : B00JXQB5FQ
Brand: Si Row
Title: brown white tiger tshirt tiger stripes xl xxl
Euclidean similarity with the query image : 2.449489742783178
==

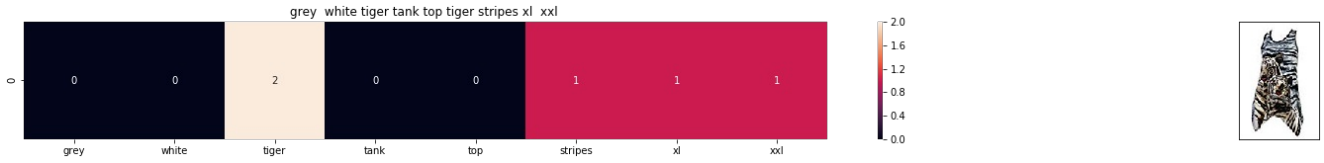
ASIN : B00JXQB5FQ
Brand: Si Row
Title: yellow tiger tshirt tiger stripes l
Euclidean similarity with the query image : 2.6457513110645907
==



ASIN : B00JXQB5FQ
Brand: Rustic Grace
Title: believed could tshirt
Euclidean similarity with the query image : 3.0
==



ASIN : B00JXQB5FQ
Brand: Ideology
Title: ideology graphic tshirt xl white
Euclidean similarity with the query image : 3.0
==

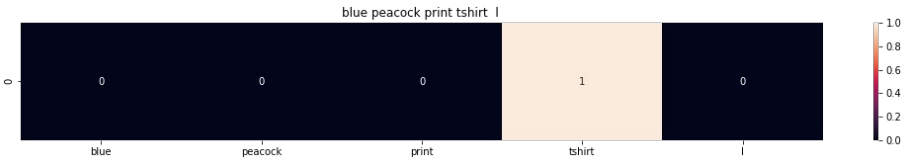


ASIN : B00JXQB5FQ
Brand: Si Row
Title: grey white tiger tank top tiger stripes xl xxl
Euclidean similarity with the query image : 3.0
===

ASIN : B00JXQB5FQ
Brand: BLVD
Title: blvd womens graphic tshirt l
Euclidean similarity with the query image : 3.1622776601683795
===

ASIN : B00JXQB5FQ
Brand: I Love Lucy
Title: juniors love lucywaaaaahhhh tshirt size xl
Euclidean similarity with the query image : 3.1622776601683795
===

ASIN : B00JXQB5FQ
Brand: simple
Title: womens fashion beautiful tshirt
Euclidean similarity with the query image : 3.1622776601683795
===



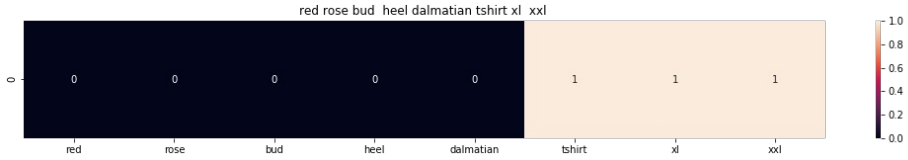
ASIN : B00JXQB5FQ
Brand: Si Row
Title: blue peacock print tshirt l
Euclidean similarity with the query image : 3.1622776601683795
===

ASIN : B00JXQB5FQ
Brand: PUMA
Title: ferrari ladies tshirt white
Euclidean similarity with the query image : 3.1622776601683795
===

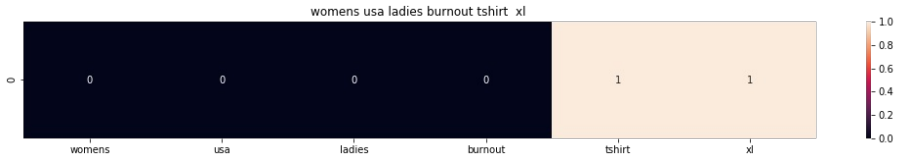
ASIN : B00JXQB5FQ
Brand: Hetalia
Title: hetalia us girl tshirt
Euclidean similarity with the query image : 3.1622776601683795
==========

ASIN : B00JXQB5FQ
Brand: KM T-shirt
Title: km tiger printed sleeveless vest tshirt
Euclidean similarity with the query image : 3.1622776601683795
==========

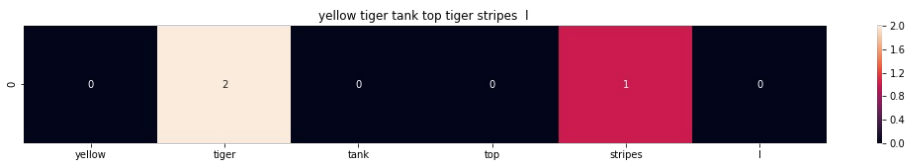
ASIN : B00JXQB5FQ
Brand: Awake
Title: morning person tshirt troll picture xl
Euclidean similarity with the query image : 3.1622776601683795
==========



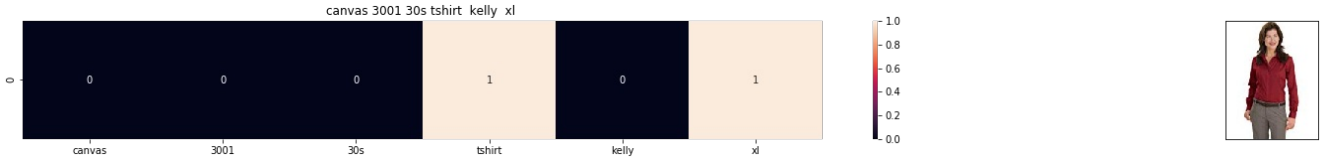
ASIN : B00JXQB5FQ
Brand: Si Row
Title: red rose bud heel dalmatian tshirt xl xxl
Euclidean similarity with the query image : 3.1622776601683795
==========



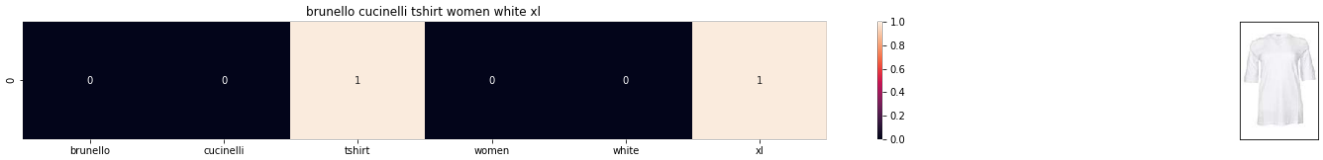
ASIN : B00JXQB5FQ
Brand: Calhoun Sportswear
Title: womens usa ladies burnout tshirt xl
Euclidean similarity with the query image : 3.1622776601683795
==========



```
ASIN : B00JXQB5FQ
Brand: Si Row
Title: yellow tiger tank top tiger stripes  l
Euclidean similarity with the query image : 3.1622776601683795
============================================================
```




```
ASIN : B00JXQB5FQ
Brand: Red House
Title: canvas 3001 30s tshirt  kelly  xl
Euclidean similarity with the query image : 3.1622776601683795
============================================================
```




```
ASIN : B00JXQB5FQ
Brand: Brunello Cucinelli
Title: brunello cucinelli tshirt women white xl
Euclidean similarity with the query image : 3.1622776601683795
============================================================
```

[8.5] TF-IDF based product similarity

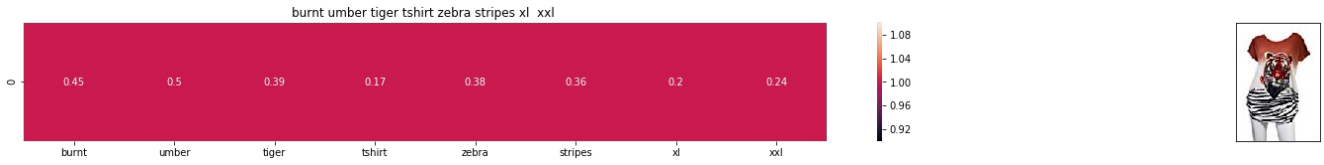
In [25]:

```
tfidf_title_vectorizer = TfidfVectorizer(min_df = 0)
tfidf_title_features = tfidf_title_vectorizer.fit_transform(data['title'])
# tfidf_title_features.shape = #data_points * #words_in_corpus
# CountVectorizer().fit_transform(courpus) returns the a sparase matrix of dimensions #data_points
* #words_in_corpus
# tfidf_title_features[doc_id, index_of_word_in_corpus] = tfidf values of the word in given doc
```

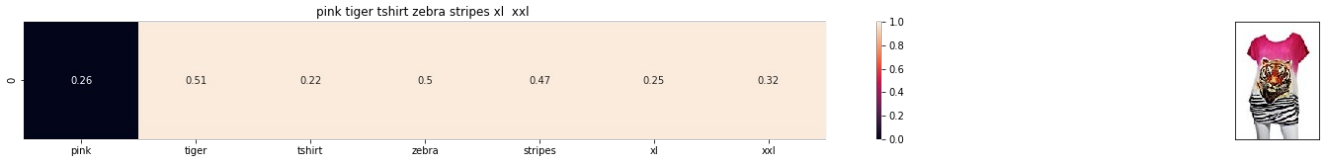
In [26]:

```
def tfidf_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 apparels
    # 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(tfidf_title_features,tfidf_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)):
        # we will pass 1. doc_id, 2. title1, 3. title2, url, model
        get_result(indices[i], data['title'].loc[df_indices[0]], data['title'].loc[df_indices[i]],
data['medium_image_url'].loc[df_indices[i]], 'tfidf')
        print('ASIN :',data['asin'].loc[df_indices[i]])
        print('BRAND :',data['brand'].loc[df_indices[i]])
        print ('Eucliden distance from the given image :', pdists[i])
        print('='*125)
```

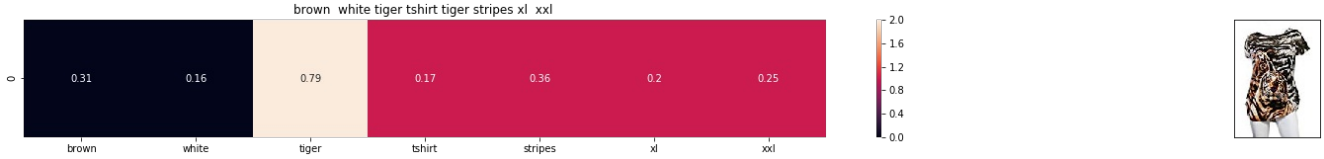
```
tfidf_model(1416, 20)
# in the output heat map each value represents the tfidf values of the label word, the color repre
sents the intersection with inputs title
```



ASIN : B00JXQB5FQ
BRAND : Si Row
Eucliden distance from the given image : 0.0



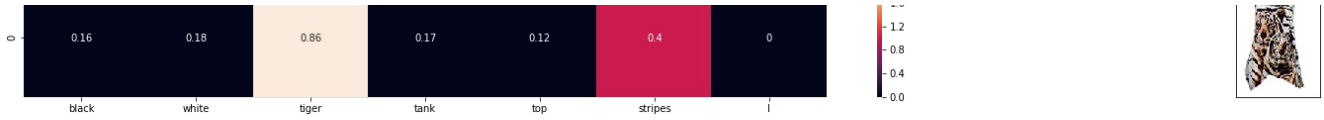
ASIN : B00JXQB5FQ
BRAND : Si Row
Eucliden distance from the given image : 0.7536331912451361



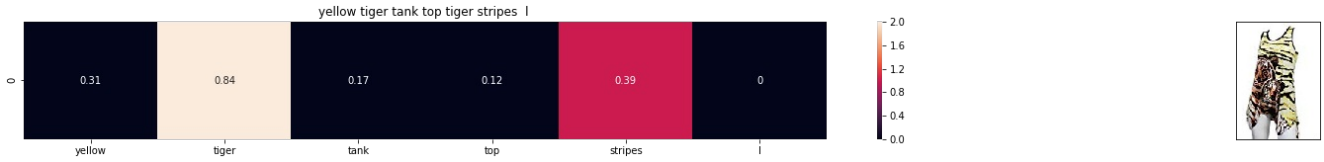
ASIN : B00JXQB5FQ
BRAND : Si Row
Eucliden distance from the given image : 0.9357643943769645

ASIN : B00JXQB5FQ
BRAND : Si Row
Eucliden distance from the given image : 0.9586153524200749

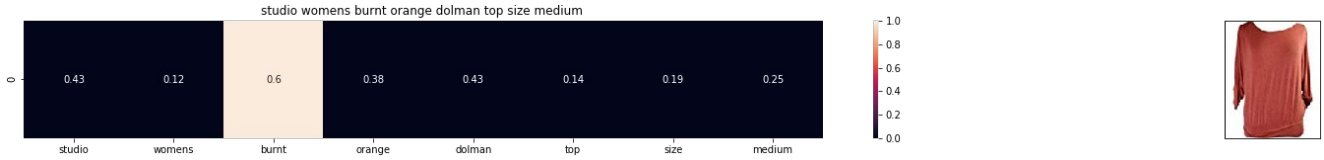
ASIN : B00JXQB5FQ
BRAND : Si Row
Eucliden distance from the given image : 1.000074961446881



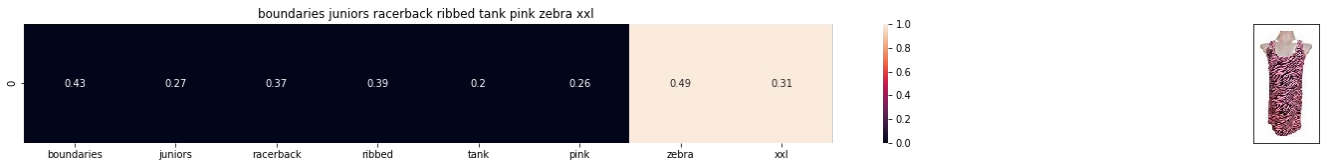
ASIN : B00JXQB5FQ
BRAND : Si Row
Eucliden distance from the given image : 1.023215552457452
==========



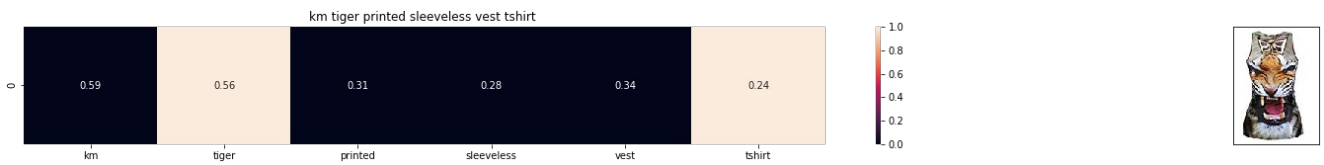
ASIN : B00JXQB5FQ
BRAND : Si Row
Eucliden distance from the given image : 1.031991846303421
==========



ASIN : B00JXQB5FQ
BRAND : Studio M
Eucliden distance from the given image : 1.2106843670424716
==========



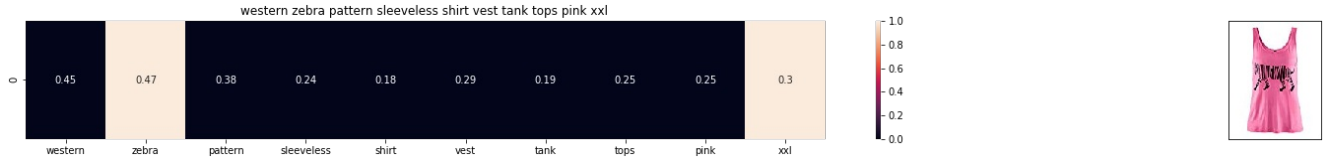
ASIN : B00JXQB5FQ
BRAND : No Boundaries
Eucliden distance from the given image : 1.212168381072083
==========



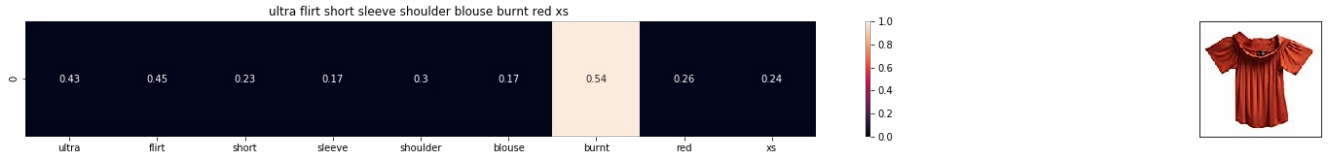
ASIN : B00JXQB5FQ
BRAND : KM T-shirt
Eucliden distance from the given image : 1.219790640280982
==========

ASIN : B00JXQB5FQ

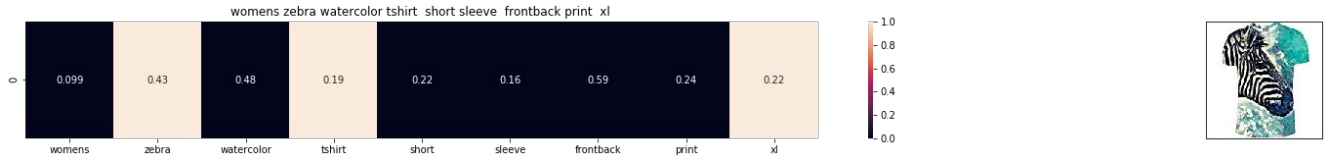
BRAND : Black Swan
Eucliden distance from the given image : 1.2206849659998316
==========



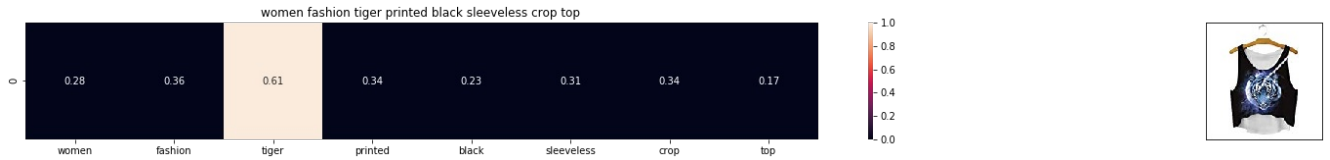
ASIN : B00JXQB5FQ
BRAND : Black Temptation
Eucliden distance from the given image : 1.221281392120943
==========



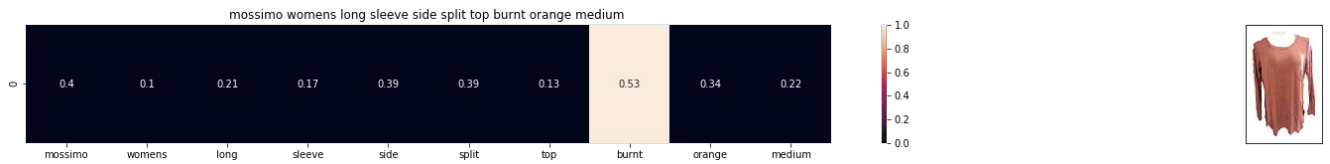
ASIN : B00JXQB5FQ
BRAND : Ultra Flirt
Eucliden distance from the given image : 1.2313364094597743
==========



ASIN : B00JXQB5FQ
BRAND : WHAT ON EARTH
Eucliden distance from the given image : 1.2318451972624518
==========



ASIN : B00JXQB5FQ
BRAND : MKP Crop Top
Eucliden distance from the given image : 1.2340607457359425
==========

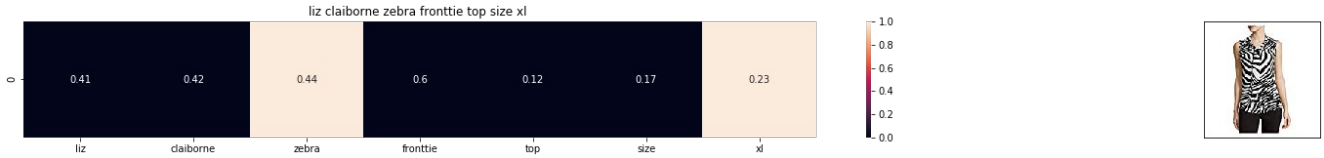


ASIN : B00JXQB5FQ
BRAND : Mossimo
Eucliden distance from the given image : 1.2352785577664824
==========


```
ASIN : B00JXQB5FQ
BRAND : Tultex
Eucliden distance from the given image : 1.2364572988127 82
==========================================================================================================
======================
```


```
ASIN : B00JXQB5FQ
BRAND : Vivian's Fashions
Eucliden distance from the given image : 1.24996155052848
==========================================================================================================
======================
```


```
ASIN : B00JXQB5FQ
BRAND : YICHUN
Eucliden distance from the given image : 1.2535461420856102
==========================================================================================================
======================
```




```
ASIN : B00JXQB5FQ
BRAND : Liz Claiborne
Eucliden distance from the given image : 1.2538832938357722
==========================================================================================================
======================
```

[8.5] IDF based product similarity

In [27]:

```
idf_title_vectorizer = CountVectorizer()
idf_title_features = idf_title_vectorizer.fit_transform(data['title'])

# idf_title_features.shape = #data_points * #words_in_corpus
# CountVectorizer().fit_transform(courpus) returns the a sparase matrix of dimensions #data_points
* #words_in_corpus
# idf_title_features[doc_id, index_of_word_in_corpus] = number of times the word occured in that d
oc
```

In [28]:

```
def n_containing(word):
    # return the number of documents which had the given word
    return sum(1 for blob in data['title'] if word in blob.split())
```

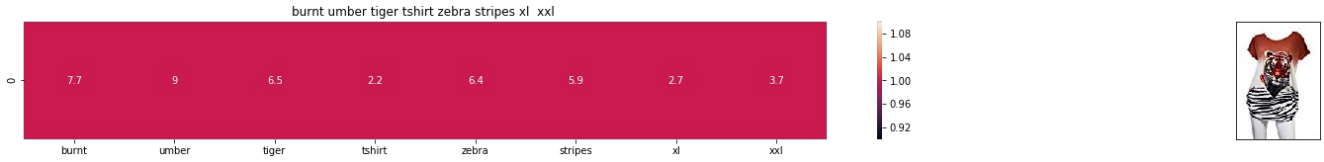
```
def idf(word):
    # idf = log(#number of docs / #number of docs which had the given word)
    return math.log(data.shape[0] / (n_containing(word)))
```

In [29]:

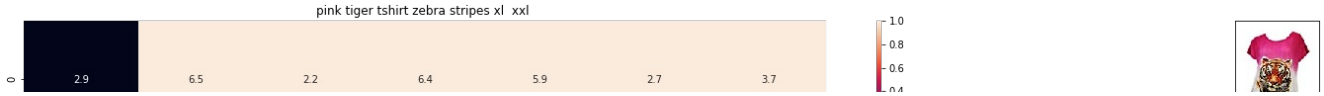
```
# we need to convert the values into float
idf_title_features  = idf_title_features.astype(np.float)

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 t
he word
    # idf_title_features[:, idf_title_vectorizer.vocabulary_[i]].nonzero()[0] will return all docu
ments 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
```

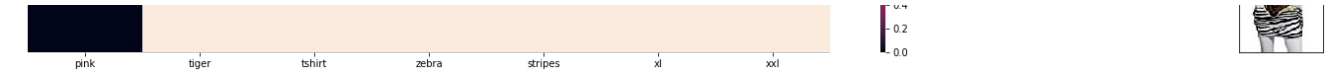
In [31]:

```
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 apparels
    # 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_indices[i]], d
ata['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(1416,20)
# in the output heat map each value represents the idf values of the label word, the color represe
nts the intersection with inputs title
```

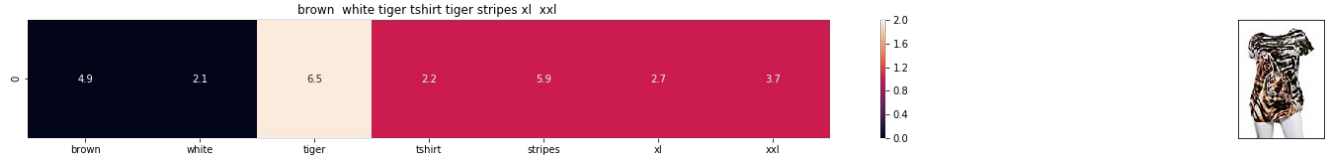



```
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 0.0
=============================================================================================================
======================
```

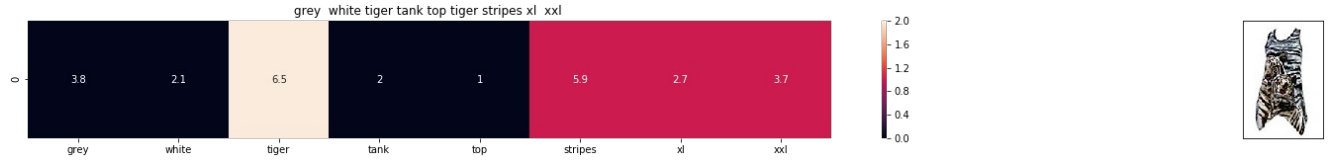




ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 12.205071311221 77
==========



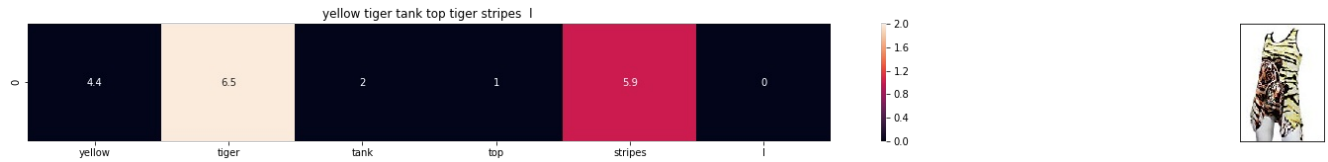
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 14.468362685603465
==========



ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 14.486832924778964
==========

ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 14.833392966672909
==========

ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 14.898744516719225
==========

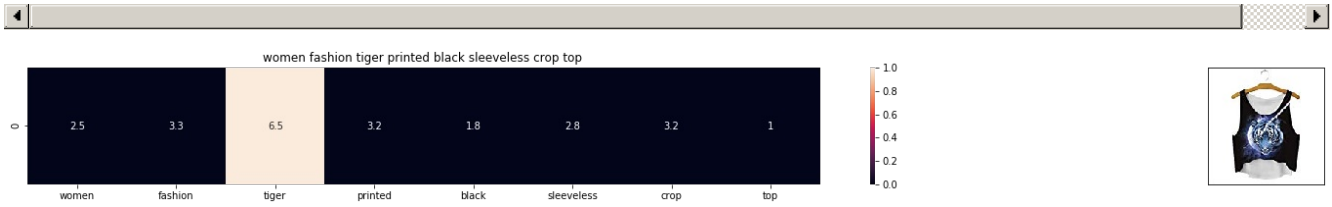


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from the given image : 15.224458287343769

euclidean distance from the given image : 15.224502675457697

==========

==========



ASIN : B00JXQB5FQ
Brand : MKP Crop Top
euclidean distance from the given image : 17.080812955631995

==========

==========

ASIN : B00JXQB5FQ
Brand : Vietsbay
euclidean distance from the given image : 17.090168125645416

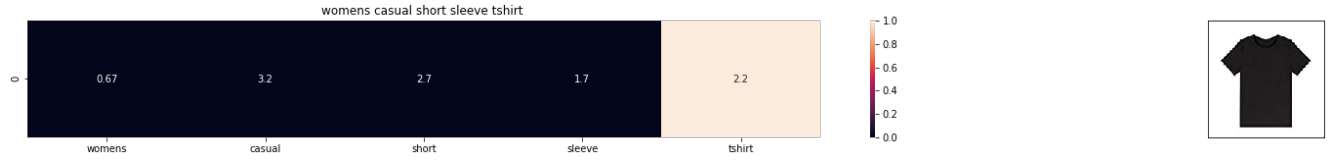
==========

==========

ASIN : B00JXQB5FQ
Brand : Sofra
euclidean distance from the given image : 17.153215337562703

==========

==========



ASIN : B00JXQB5FQ
Brand : Rain
euclidean distance from the given image : 17.33671523874989

==========

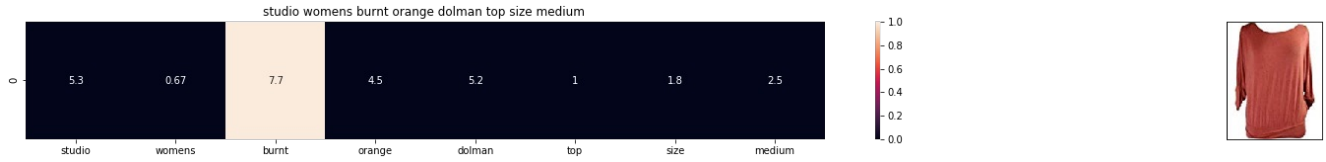
==========

ASIN : B00JXQB5FQ
Brand : Vivian's Fashions
euclidean distance from the given image : 17.410075941001253

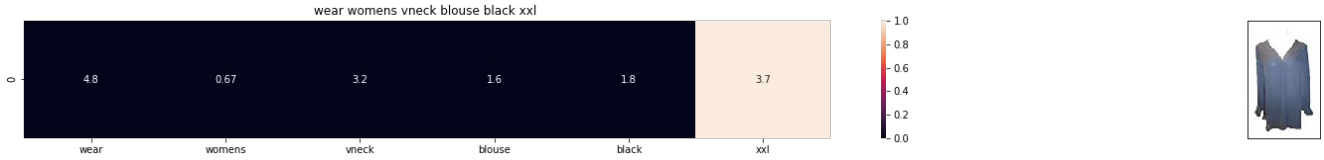
==========

==========

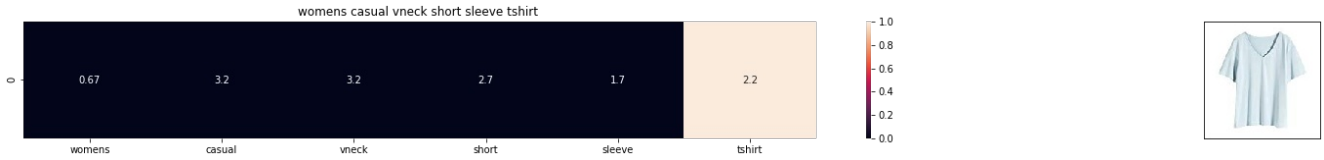
ASIN : B00JXQB5FQ
Brand : ERMANNO SCERVINO
euclidean distance from the given image : 17.539921335459557
==========
==========



ASIN : B00JXQB5FQ
Brand : Studio M
euclidean distance from the given image : 17.61275854366134
==========
==========

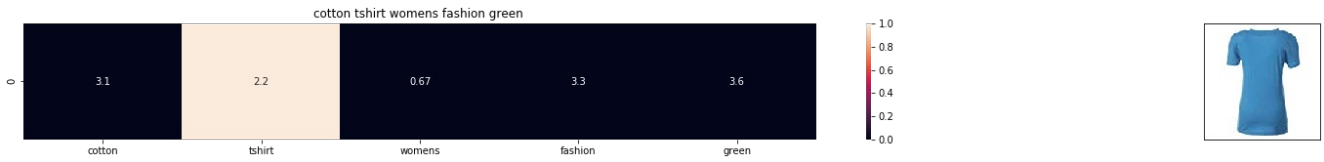


ASIN : B00JXQB5FQ
Brand : Who What Wear
euclidean distance from the given image : 17.623745282500135
==========
==========



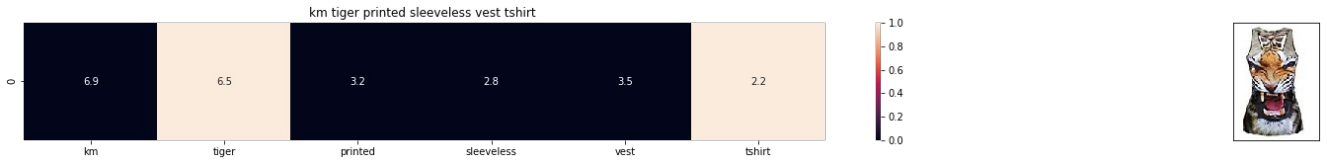
ASIN : B00JXQB5FQ
Brand : Rain
euclidean distance from the given image : 17.634342496835046
==========
==========

ASIN : B00JXQB5FQ
Brand : Very J
euclidean distance from the given image : 17.63753712743611
==========
==========



ASIN : B00JXQB5FQ
Brand : Ivan Levi

```
euclidean distance from the given image : 17.7230738913371
=========================================================================================================
======================
```




```
ASIN : B00JXQB5FQ
Brand : KM T-shirt
euclidean distance from the given image : 17.762588561202364
=========================================================================================================
======================
```


```
ASIN : B00JXQB5FQ
Brand : HP-LEISURE
euclidean distance from the given image : 17.779536864674238
=========================================================================================================
======================
```

[9] Text Semantics based product similarity

In [32]:

```
# credits: https://www.kaggle.com/c/word2vec-nlp-tutorial#part-2-word-vectors
# Custom Word2Vec using your own text data.
# Do NOT RUN this code.
# It is meant as a reference to build your own Word2Vec when you have
# lots of data.

'''
# Set values for various parameters
num_features = 300    # Word vector dimensionality
min_word_count = 1    # Minimum word count
num_workers = 4       # Number of threads to run in parallel
context = 10          # Context window size

downsampling = 1e-3   # Downsample setting for frequent words

# Initialize and train the model (this will take some time)
from gensim.models import word2vec
print ("Training model...")
model = word2vec.Word2Vec(sen_corpus, workers=num_workers, \
            size=num_features, min_count = min_word_count, \
            window = context)

'''
```

Out[32]:

```
'\n# Set values for various parameters\nnum_features = 300    # Word vector dimensionality
\nmin_word_count = 1    # Minimum word count                          \nnum_workers = 4       # Numbe
of threads to run in parallel\ncontext = 10          # Context window size
\ndownsampling = 1e-3   # Downsample setting for frequent words\n\n# Initialize and train the
model (this will take some time)\nfrom gensim.models import word2vec\nprint ("Training
model...")\nmodel = word2vec.Word2Vec(sen_corpus, workers=num_workers,
size=num_features, min_count = min_word_count,             window = context)\n    \n'
```

In [33]:

```
from gensim.models import Word2Vec
```

```
from gensim.models import KeyedVectors
import pickle

# in this project we are using a pretrained model by google
# its 3.3G file, once you load this into your memory
# it occupies ~9Gb, so please do this step only if you have >12G of ram
# we will provide a pickle file wich contains a dict ,
# and it contains all our courpus words as keys and  model[word] as values
# To use this code-snippet, download "GoogleNews-vectors-negative300.bin"
# from https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
# it's 1.9GB in size.

'''
model = KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin', binary=True)
'''

#if you do NOT have RAM >= 12GB, use the code below.
with open('word2vec_model', 'rb') as handle:
    model = pickle.load(handle)
```

In [34]:

```
# Utility functions

def get_word_vec(sentence, doc_id, m_name):
    # sentence : title of the apparel
    # doc_id: document id in our corpus
    # m_name: model information it will take two values
        # if  m_name == 'avg', we will append the model[i], w2v representation of word i
        # if m_name == 'weighted', we will multiply each w2v[word] with the idf(word)
    vec = []
    for i in sentence.split():
        if i in vocab:
            if m_name == 'weighted' and i in  idf_title_vectorizer.vocabulary_:
                vec.append(idf_title_features[doc_id, idf_title_vectorizer.vocabulary_[i]] * model[
i])
            elif m_name == 'avg':
                vec.append(model[i])
        else:
            # if the word in our courpus is not there in the google word2vec corpus, we are just
ignoring it
            vec.append(np.zeros(shape=(300,)))
    # we will return a numpy array of shape (#number of words in title * 300 ) 300 =
len(w2v_model[word])
    # each row represents the word2vec representation of each word (weighted/avg) in given
sentance
    return  np.array(vec)

def get_distance(vec1, vec2):
    # vec1 = np.array(#number_of_words_title1 * 300), each row is a vector of length 300
corresponds to each word in give title
    # vec2 = np.array(#number_of_words_title2 * 300), each row is a vector of length 300
corresponds to each word in give title

    final_dist = []
    # for each vector in vec1 we caluclate the distance(euclidean) to all vectors in vec2
    for i in vec1:
        dist = []
        for j in vec2:
            # np.linalg.norm(i-j) will result the euclidean distance between vectors i, j
            dist.append(np.linalg.norm(i-j))
        final_dist.append(np.array(dist))
    # final_dist = np.array(#number of words in title1 * #number of words in title2)
    # final_dist[i,j] = euclidean distance between vectors i, j
    return np.array(final_dist)


def heat_map_w2v(sentence1, sentence2, url, doc_id1, doc_id2, model):
    # sentance1 : title1, input apparel
    # sentance2 : title2, recommended apparel
    # url: apparel image url
    # doc_id1: document id of input apparel
    # doc_id2: document id of recommended apparel
    # model: it can have two values, 1. avg 2. weighted

    #s1_vec = np.array(#number_of_words_title1 * 300), each row is a vector(weighted/avg) of lengt
```

```
h 300 corresponds to each word in give title
    s1_vec = get_word_vec(sentence1, doc_id1, model)
    #s2_vec = np.array(#number_of_words_title1 * 300), each row is a vector(weighted/avg) of lengt
h 300 corresponds to each word in give title
    s2_vec = get_word_vec(sentence2, doc_id2, model)

    # s1_s2_dist = np.array(#number of words in title1 * #number of words in title2)
    # s1_s2_dist[i,j] = euclidean distance between words i, j
    s1_s2_dist = get_distance(s1_vec, s2_vec)



    # devide whole figure into 2 parts 1st part displays heatmap 2nd part displays image of appare
l
    gs = gridspec.GridSpec(2, 2, width_ratios=[4,1],height_ratios=[2,1])
    fig = plt.figure(figsize=(15,15))

    ax = plt.subplot(gs[0])
    # ploting the heap map based on the pairwise distances
    ax = sns.heatmap(np.round(s1_s2_dist,4), annot=True)
    # set the x axis labels as recommended apparels title
    ax.set_xticklabels(sentence2.split())
    # set the y axis labels as input apparels title
    ax.set_yticklabels(sentence1.split())
    # set title as recommended apparels title
    ax.set_title(sentence2)

    ax = plt.subplot(gs[1])
    # we remove all grids and axis labels for image
    ax.grid(False)
    ax.set_xticks([])
    ax.set_yticks([])
    display_img(url, ax, fig)

    plt.show()
```

In [35]:

```
# vocab = stores all the words that are there in google w2v model
# vocab = model.wv.vocab.keys() # if you are using Google word2Vec

vocab = model.keys()
# this function will add the vectors of each word and returns the avg vector of given sentance
def build_avg_vec(sentence, num_features, doc_id, m_name):
    # sentace: its title of the apparel
    # num_features: the lenght of word2vec vector, its values = 300
    # m_name: model information it will take two values
        # if  m_name == 'avg', we will append the model[i], w2v representation of word i
        # if m_name == 'weighted', we will multiply each w2v[word] with the idf(word)

    featureVec = np.zeros((num_features,), dtype="float32")
    # we will intialize a vector of size 300 with all zeros
    # we add each word2vec(wordi) to this fetureVec
    nwords = 0

    for word in sentence.split():
        nwords += 1
        if word in vocab:
            if m_name == 'weighted' and word in  idf_title_vectorizer.vocabulary_:
                featureVec = np.add(featureVec, idf_title_features[doc_id, idf_title_vectorizer.voc
abulary_[word]] * model[word])
            elif m_name == 'avg':
                featureVec = np.add(featureVec, model[word])
    if(nwords>0):
        featureVec = np.divide(featureVec, nwords)
    # returns the avg vector of given sentance, its of shape (1, 300)
    return featureVec
```

[9.2] Average Word2Vec product similarity.

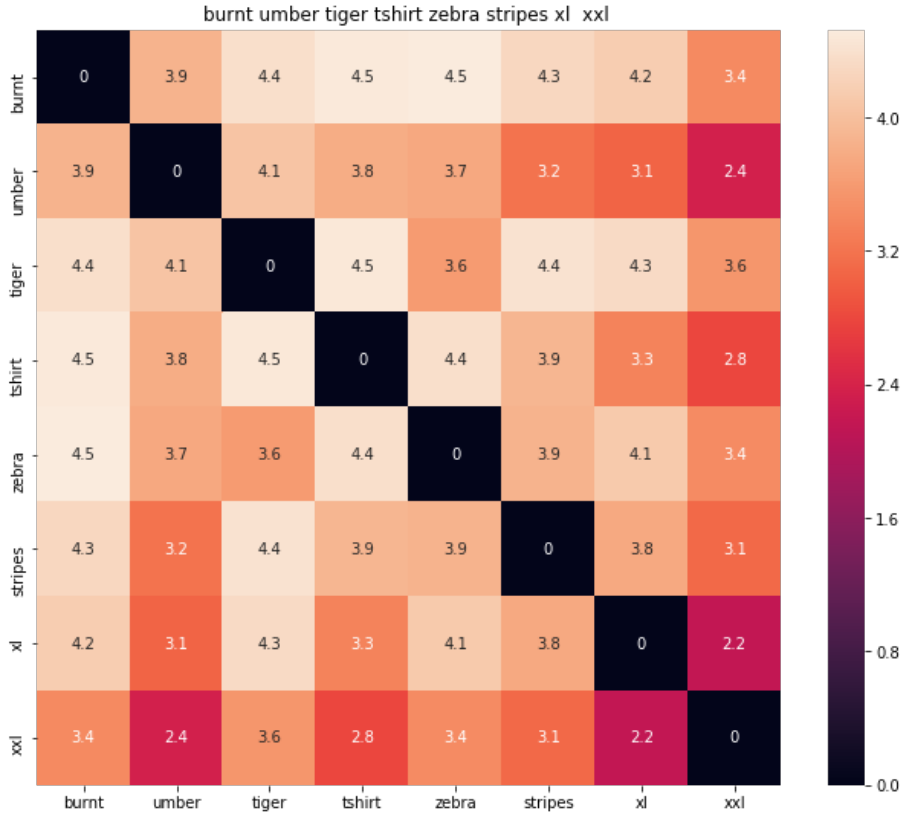
In [36]:

```
doc_id = 0
w2v_title = []
```

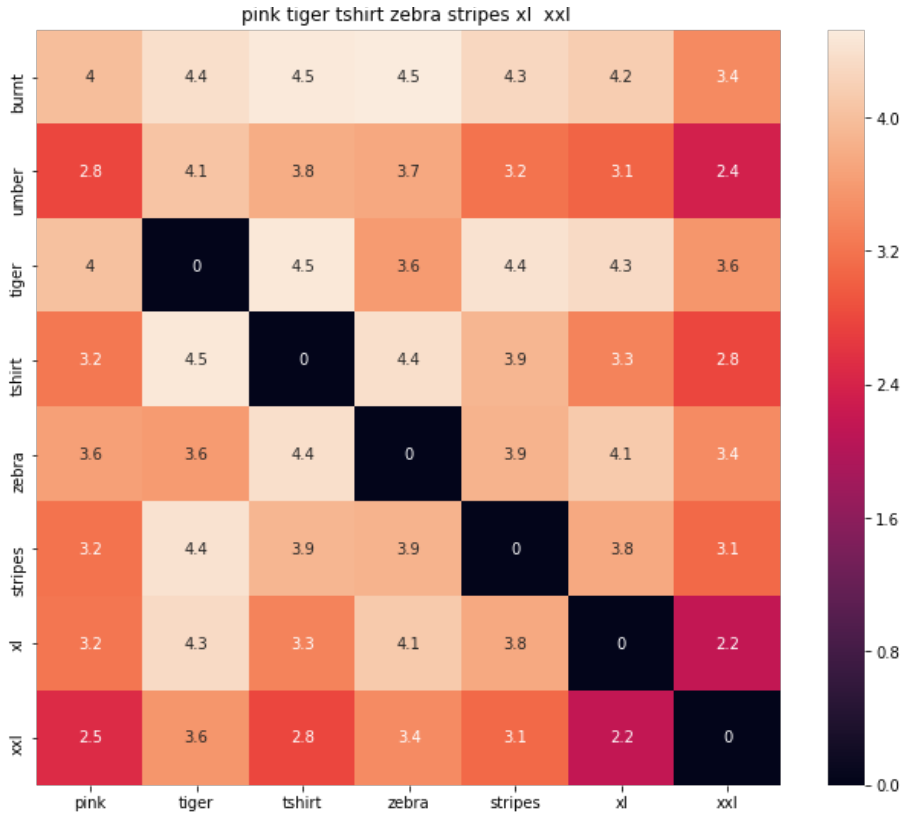
```
w2v_title = []
# for every title we build a avg vector representation
for i in data['title']:
    w2v_title.append(build_avg_vec(i, 300, doc_id,'avg'))
    doc_id += 1

# w2v_title = np.array(# number of doc in courpus * 300), each row corresponds to a doc
w2v_title = np.array(w2v_title)
```

In [37]:

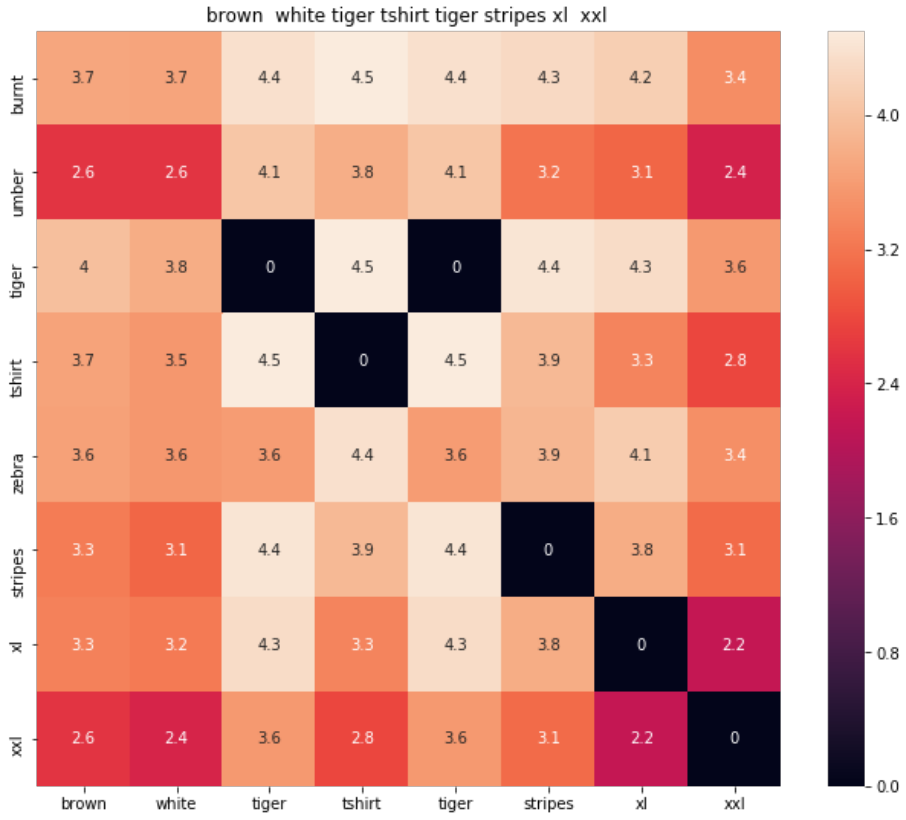
```
def avg_w2v_model(doc_id, num_results):
    # doc_id: apparel's id in given corpus

    # dist(x, y) = sqrt(dot(x, x) - 2 * dot(x, y) + dot(y, y))
    pairwise_dist = pairwise_distances(w2v_title, w2v_title[doc_id].reshape(1,-1))

    # 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)):
        heat_map_w2v(data['title'].loc[df_indices[0]],data['title'].loc[df_indices[i]], data['mediu
m_image_url'].loc[df_indices[i]], indices[0], indices[i], 'avg')
        print('ASIN :',data['asin'].loc[df_indices[i]])
        print('BRAND :',data['brand'].loc[df_indices[i]])
        print ('euclidean distance from given input image :', pdists[i])
        print('='*125)


avg_w2v_model(1416, 20)
# in the give heat map, each cell contains the euclidean distance between words i, j
```



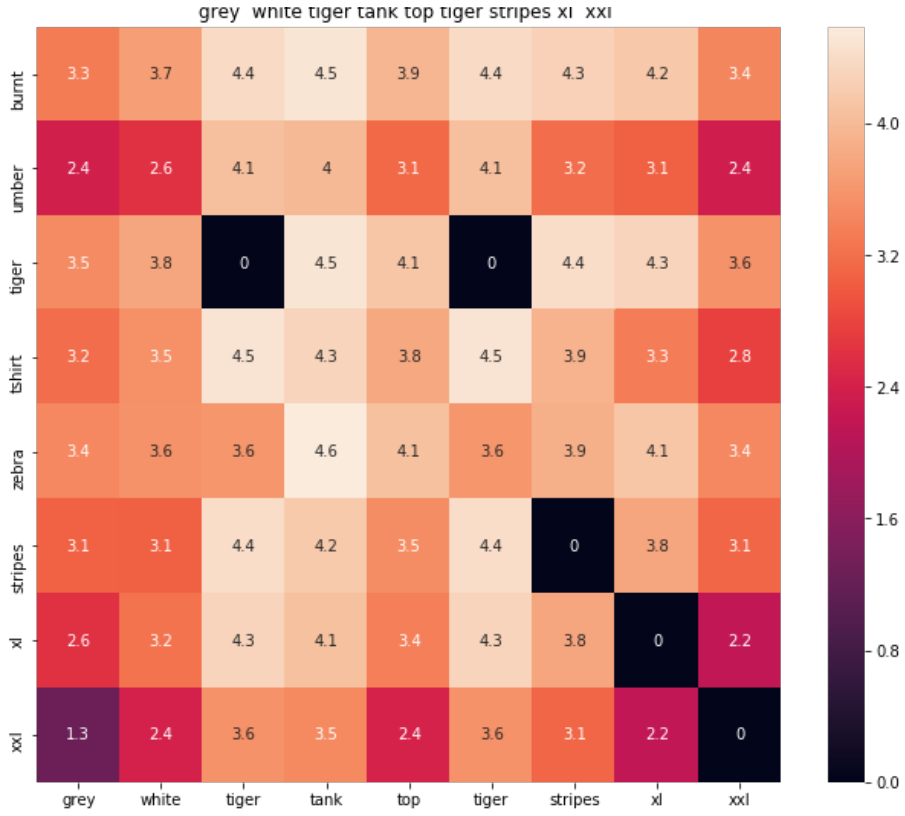
```
ASIN : B00JXQB5FQ
BRAND : Si Row
euclidean distance from given input image : 0.0
=============================================================================================================
======================
```



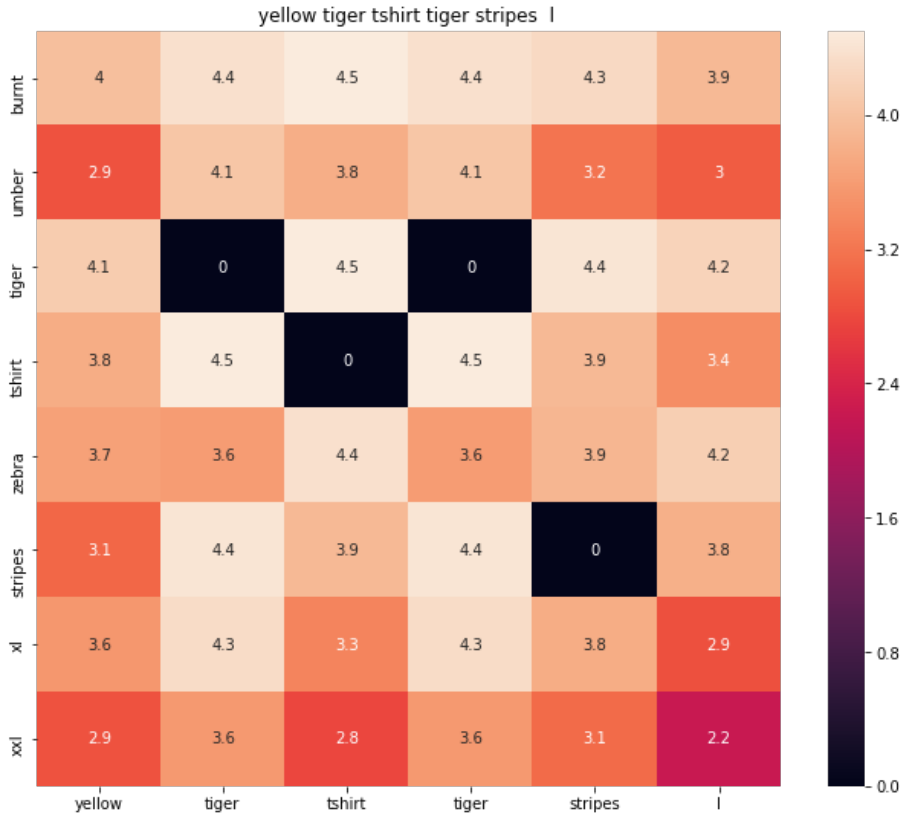
ASIN : B00JXQB5FQ
BRAND : Si Row
euclidean distance from given input image : 0.5891926
==========



ASIN : B00JXQB5FQ
BRAND : Si Row
euclidean distance from given input image : 0.7003438
==========

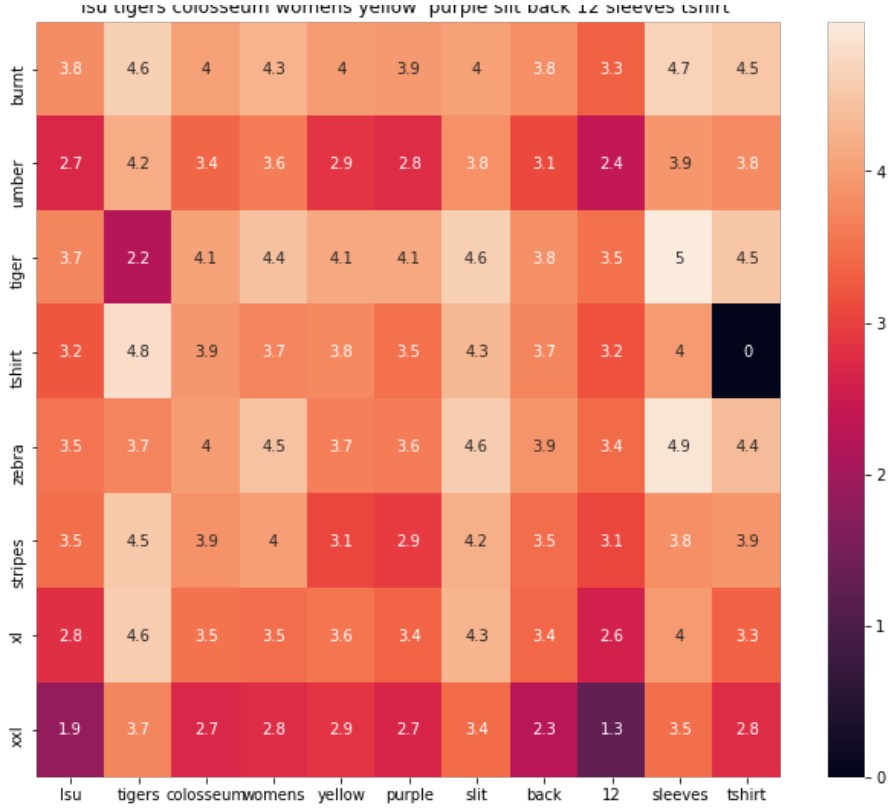


ASIN : B00JXQB5FQ
BRAND : Si Row
euclidean distance from given input image : 0.89283955
==========

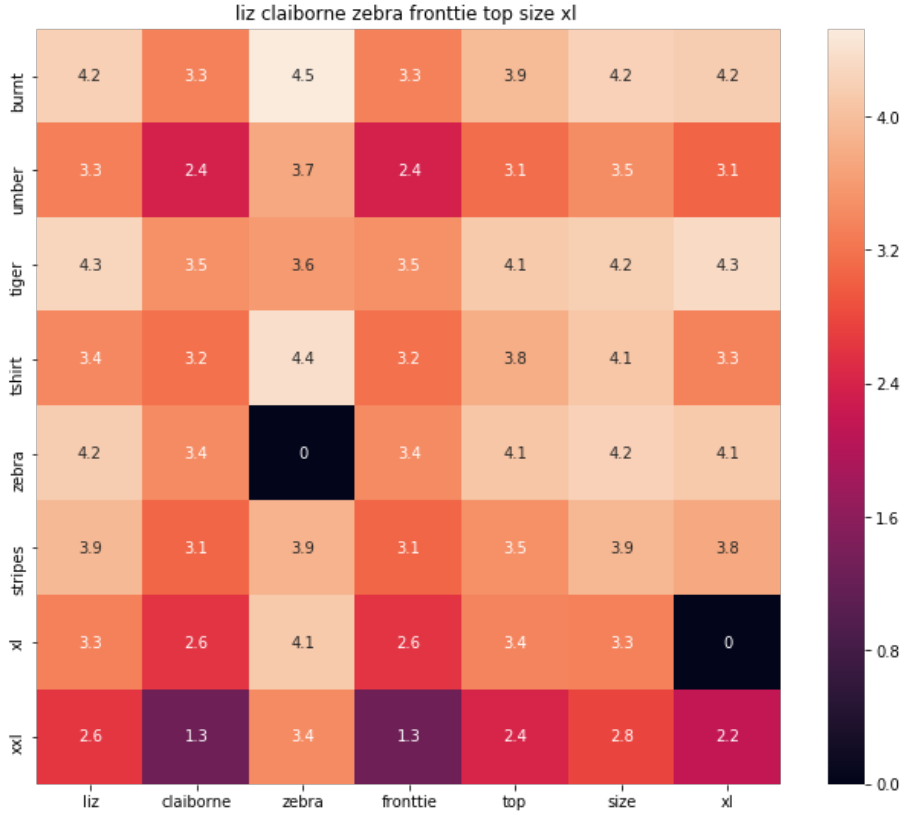


ASIN : B00JXQB5FQ
BRAND : Si Row
euclidean distance from given input image : 0.95601255
==========

lsu tigers colosseum womens yellow purple slit back 12 sleeves tshirt

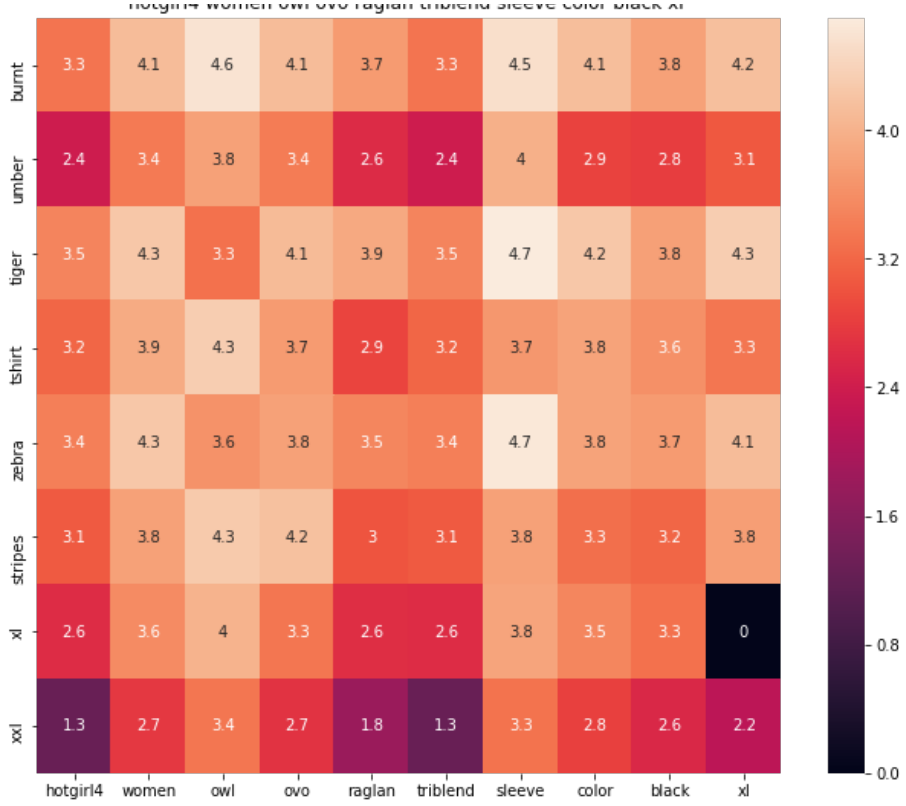


ASIN : B00JXQB5FQ
BRAND : Colosseum
euclidean distance from given input image : 1.022969
==========

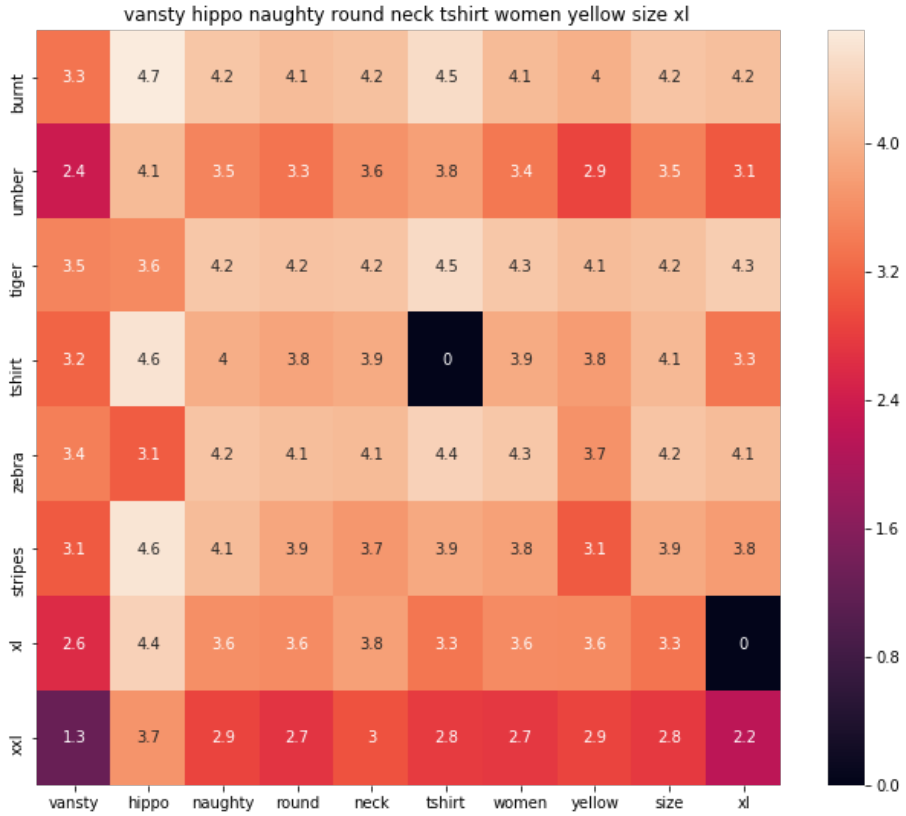


ASIN : B00JXQB5FQ
BRAND : Liz Claiborne
euclidean distance from given input image : 1.0669324
==========

hotgirl4 women owl ovo raglan triblend sleeve color black xl

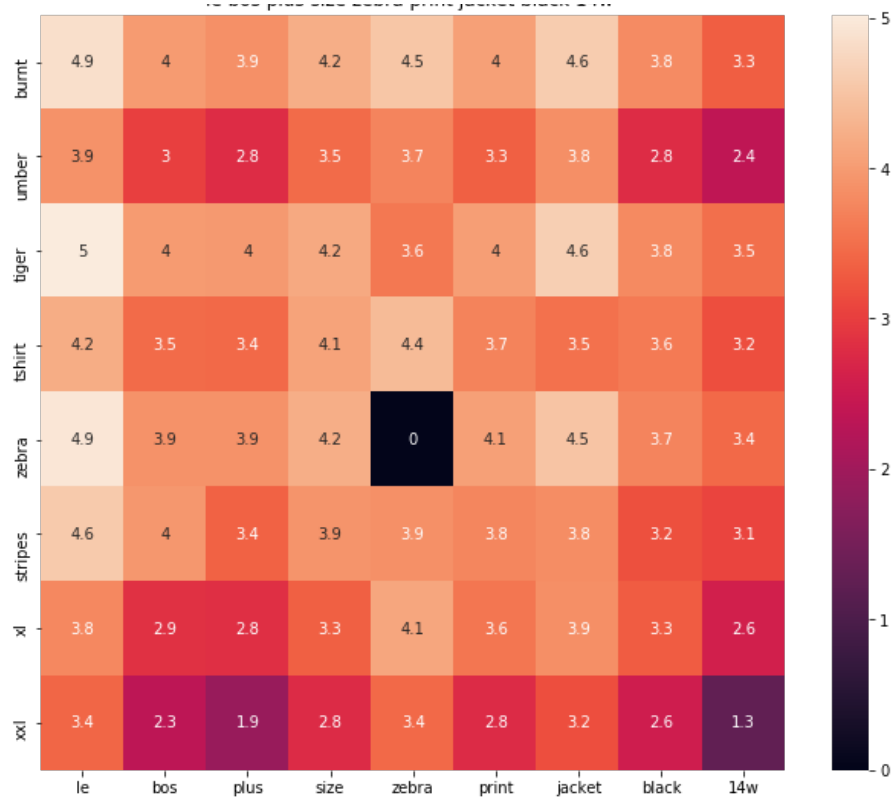


ASIN : B00JXQB5FQ
BRAND : Hotgirl4 Raglan Design
euclidean distance from given input image : 1.0731405
==========



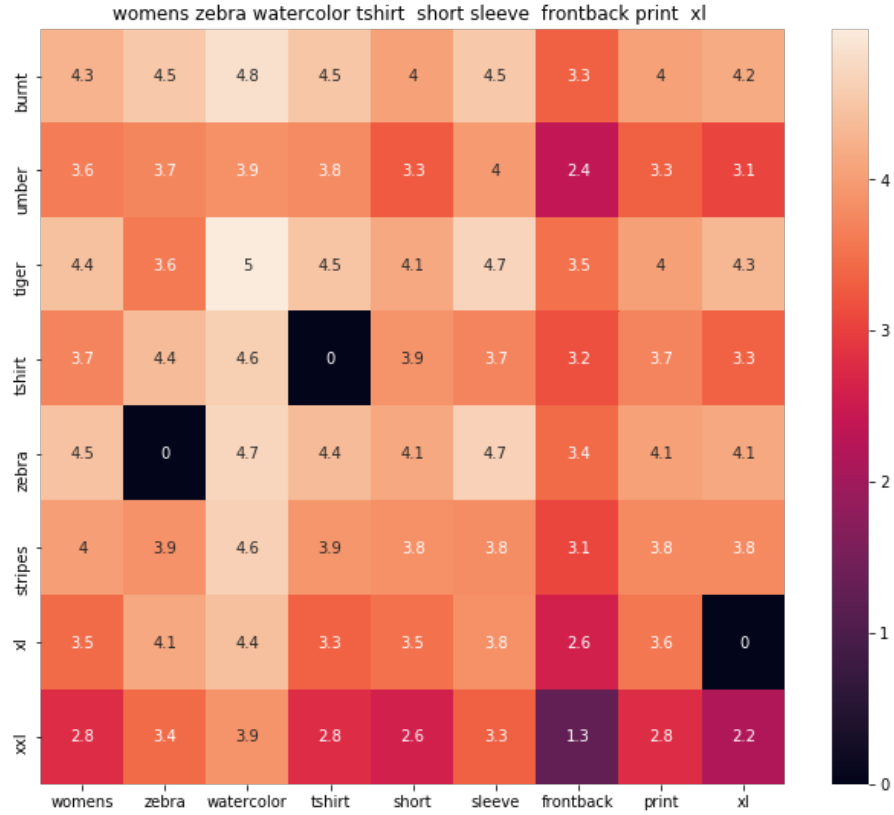
ASIN : B00JXQB5FQ
BRAND : Vansty
euclidean distance from given input image : 1.075719
==========

le bos plus size zebra print jacket black 14w



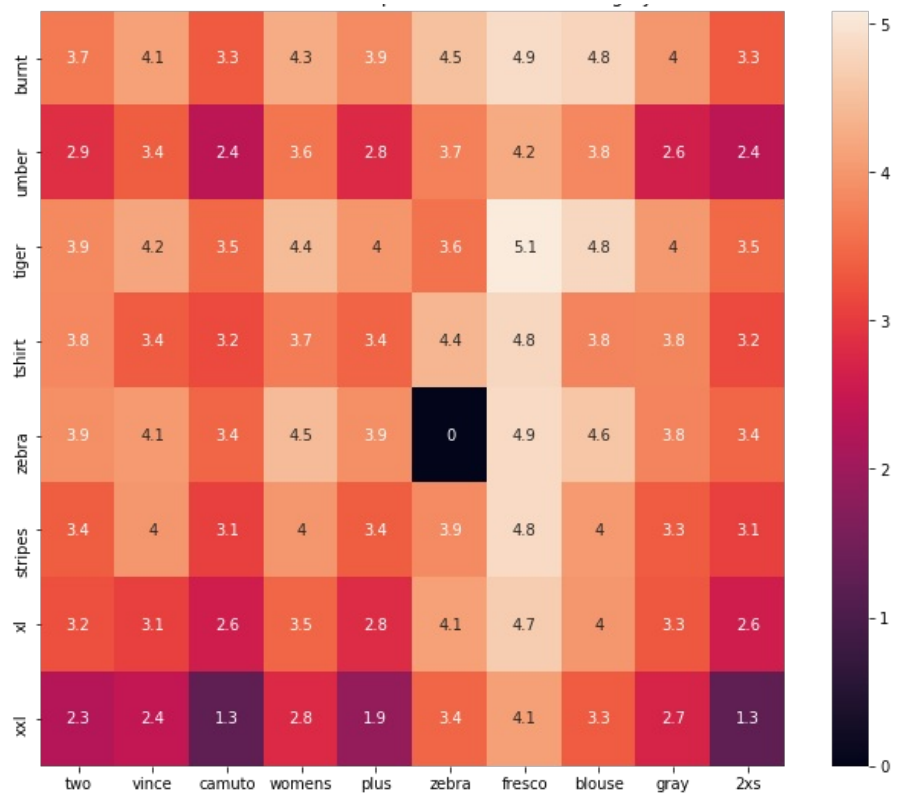
ASIN : B00JXQB5FQ
BRAND : Le Bos
euclidean distance from given input image : 1.0839964
==========

womens zebra watercolor tshirt short sleeve frontback print xl

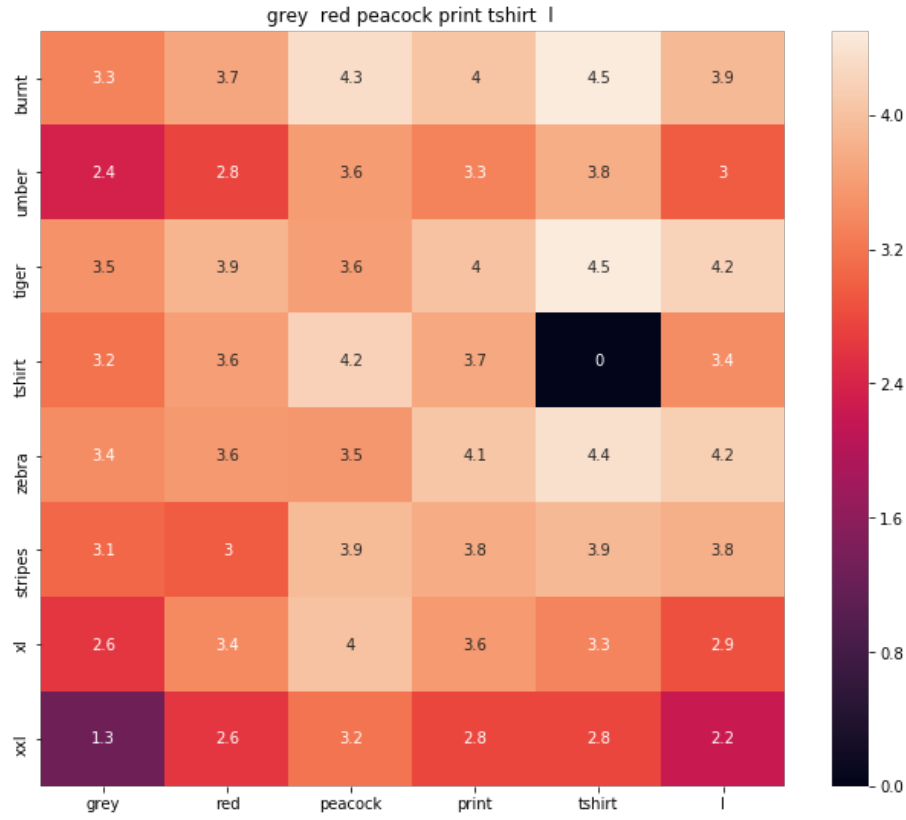


ASIN : B00JXQB5FQ
BRAND : WHAT ON EARTH
euclidean distance from given input image : 1.0842218
==========

two vince camuto womens plus zebra fresco blouse gray 2xs

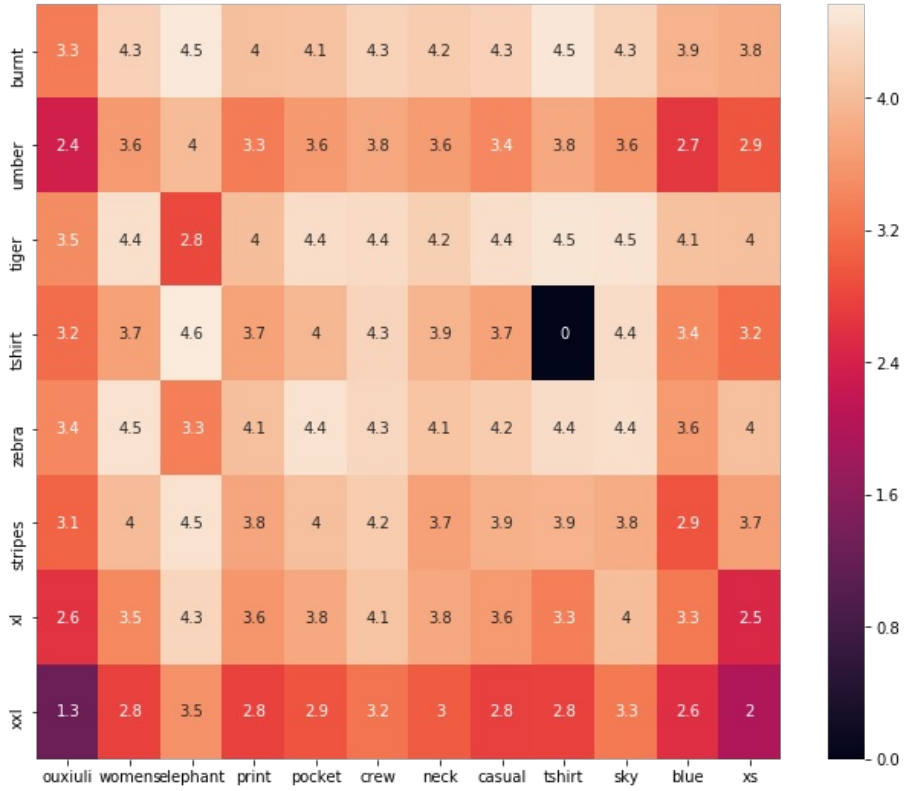


ASIN : B00JXQB5FQ
BRAND : Two by Vince Camuto
euclidean distance from given input image : 1.0895038
==========



ASIN : B00JXQB5FQ
BRAND : Si Row
euclidean distance from given input image : 1.0900588
==========

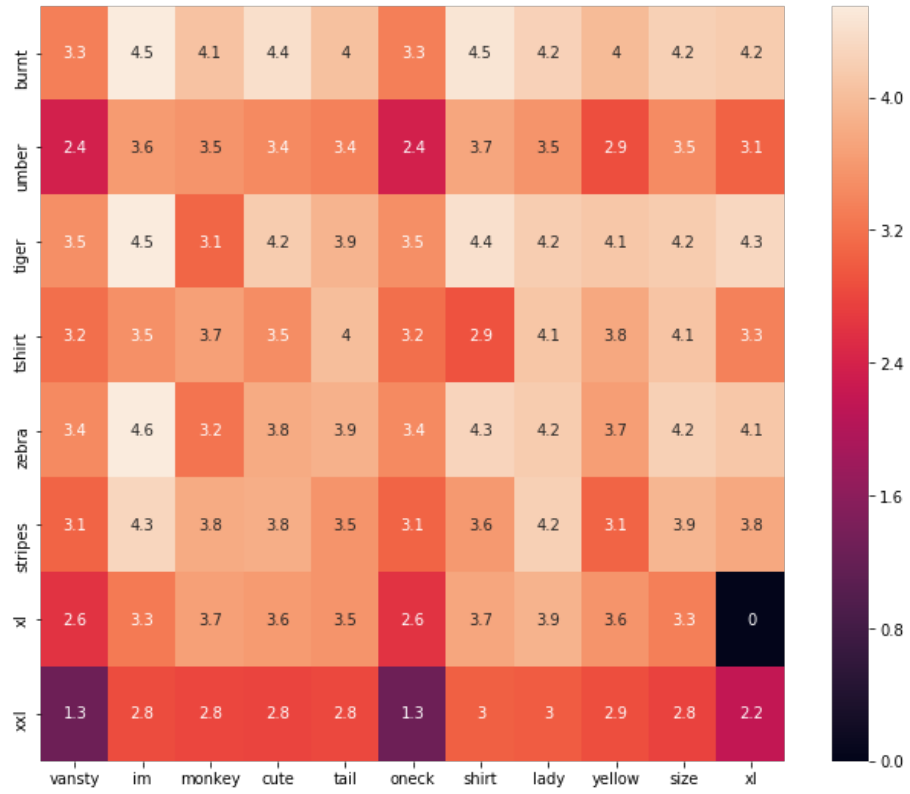
ouxiuli womens elephant print pocket crew neck casual tshirt sky blue xs



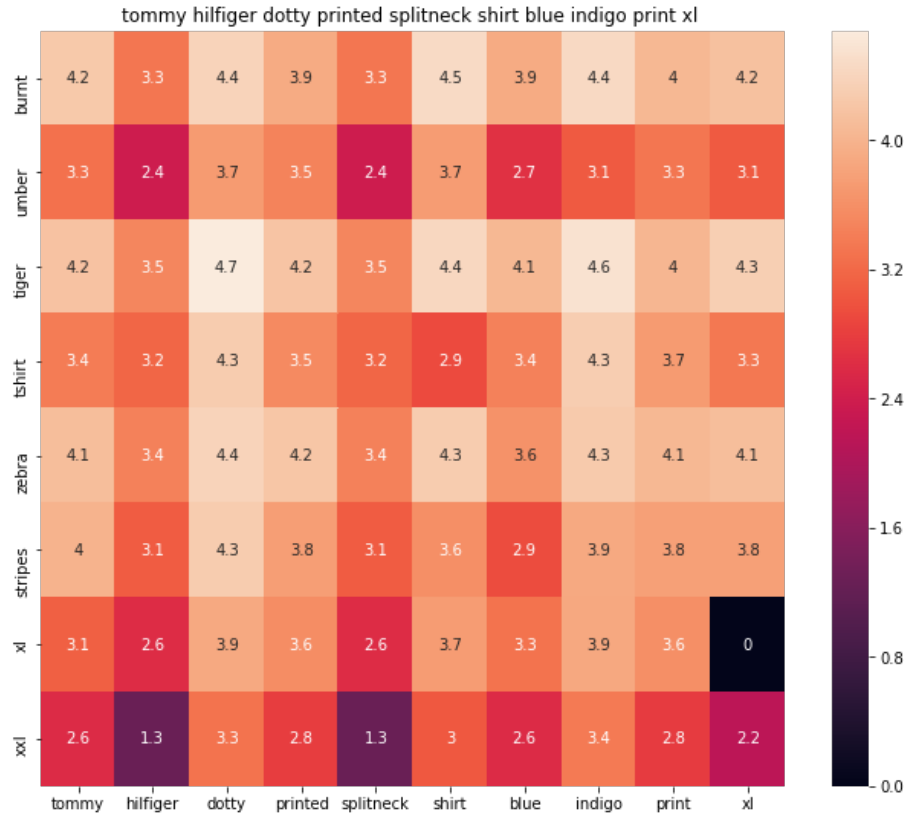
ASIN : B00JXQB5FQ
BRAND : ouxiuli
euclidean distance from given input image : 1.0920111
==========

ASIN : B00JXQB5FQ
BRAND : THILFIGER RTW
euclidean distance from given input image : 1.0923415
==========

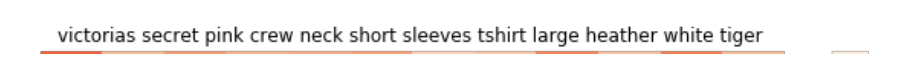
vansty im monkey cute tail oneck shirt lady yellow size xl

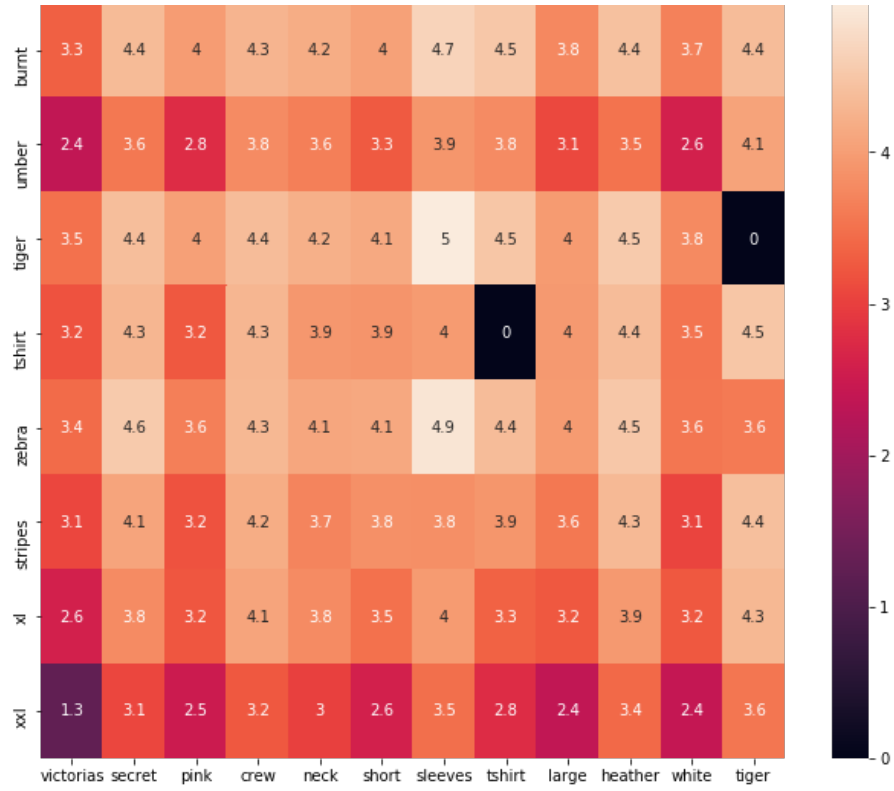


ASIN : B00JXQB5FQ
BRAND : Vansty
euclidean distance from given input image : 1.0934004
==========

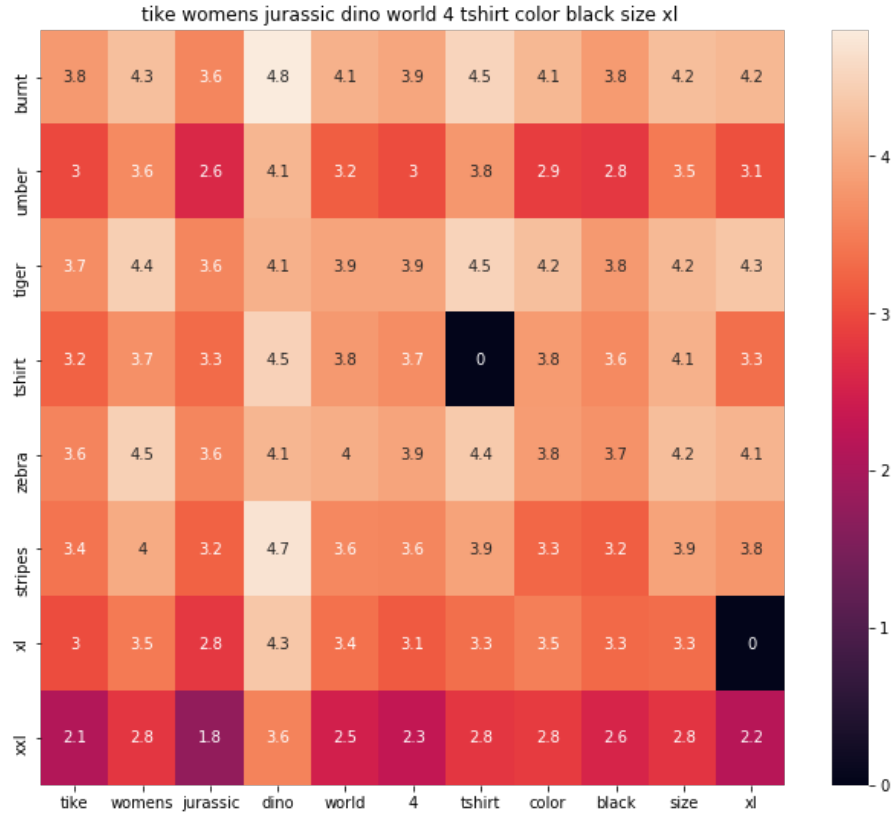


ASIN : B00JXQB5FQ
BRAND : THILFIGER RTW
euclidean distance from given input image : 1.0942024
==========

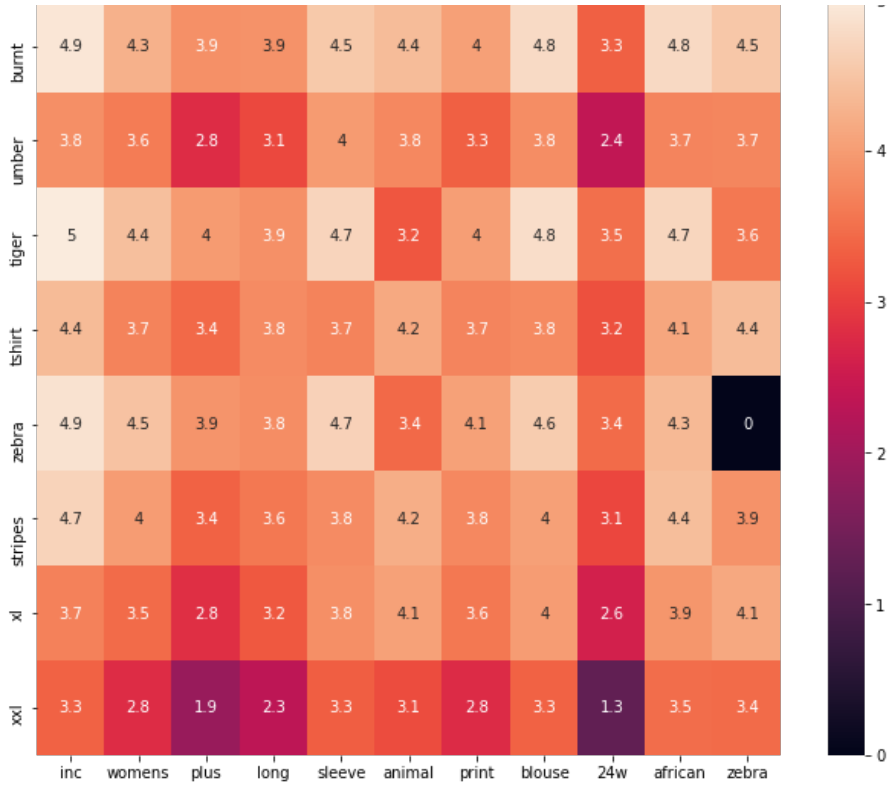




ASIN : B00JXQB5FQ
BRAND : V.Secret
euclidean distance from given input image : 1.0948304
==========



ASIN : B00JXQB5FQ
BRAND : TIKE Fashions
euclidean distance from given input image : 1.0951275
==========



```
ASIN : B00JXQB5FQ
BRAND : INC - International Concepts Woman
euclidean distance from given input image : 1.0966892
================================================================================================================
```

[9.4] IDF weighted Word2Vec for product similarity

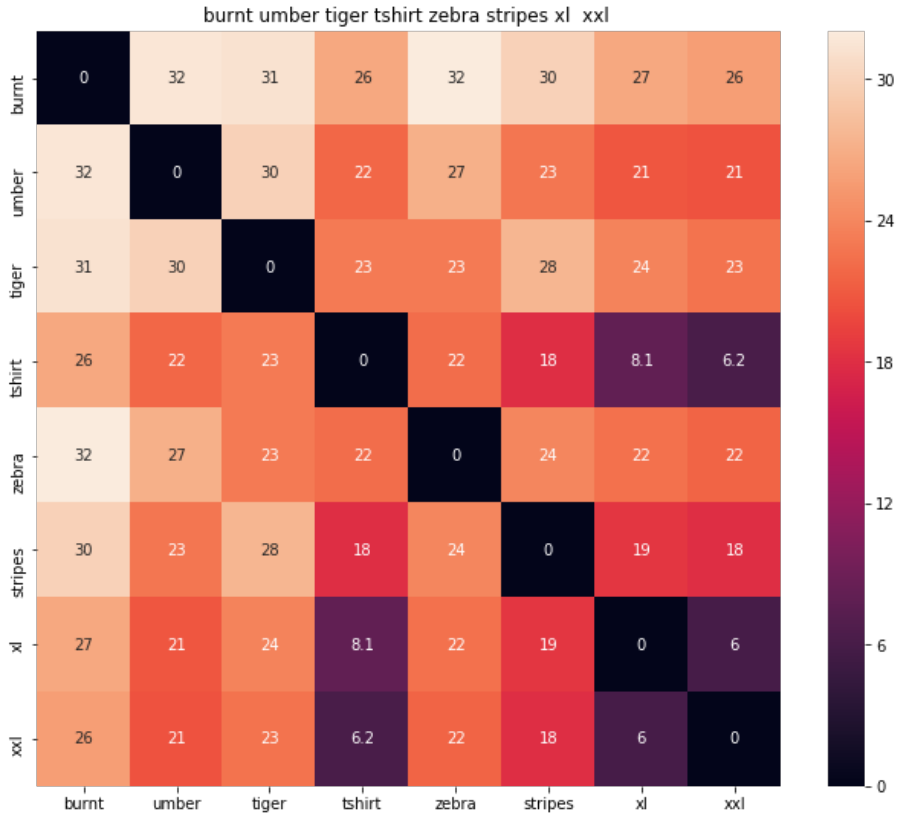
In [38]:

```
doc_id = 0
w2v_title_weight = []
# for every title we build a weighted vector representation
for i in data['title']:
    w2v_title_weight.append(build_avg_vec(i, 300, doc_id,'weighted'))
    doc_id += 1
# w2v_title = np.array(# number of doc in courpus * 300), each row corresponds to a doc
w2v_title_weight = np.array(w2v_title_weight)
```

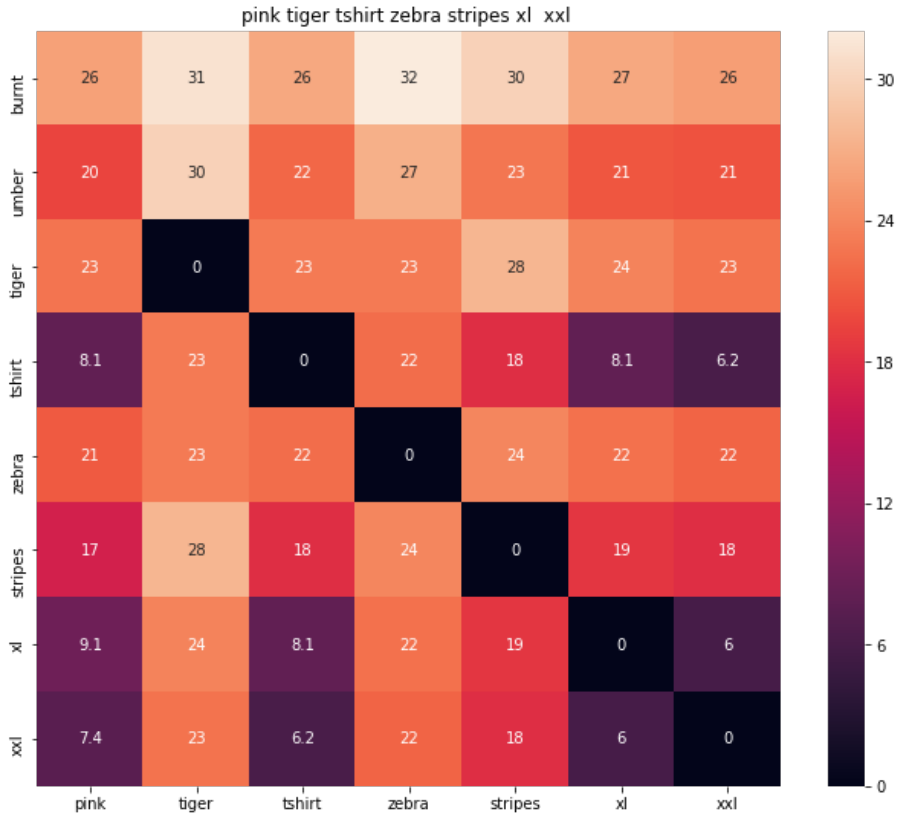
In [39]:

```
def weighted_w2v_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 apparels
    # 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(w2v_title_weight, w2v_title_weight[doc_id].reshape(1,-1))

    # 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)):
        heat_map_w2v(data['title'].loc[df_indices[0]],data['title'].loc[df_indices[i]], data['mediu
m_image_url'].loc[df_indices[i]], indices[0], indices[i], 'weighted')
        print('ASIN :',data['asin'].loc[df_indices[i]])
        print('Brand :',data['brand'].loc[df_indices[i]])
```

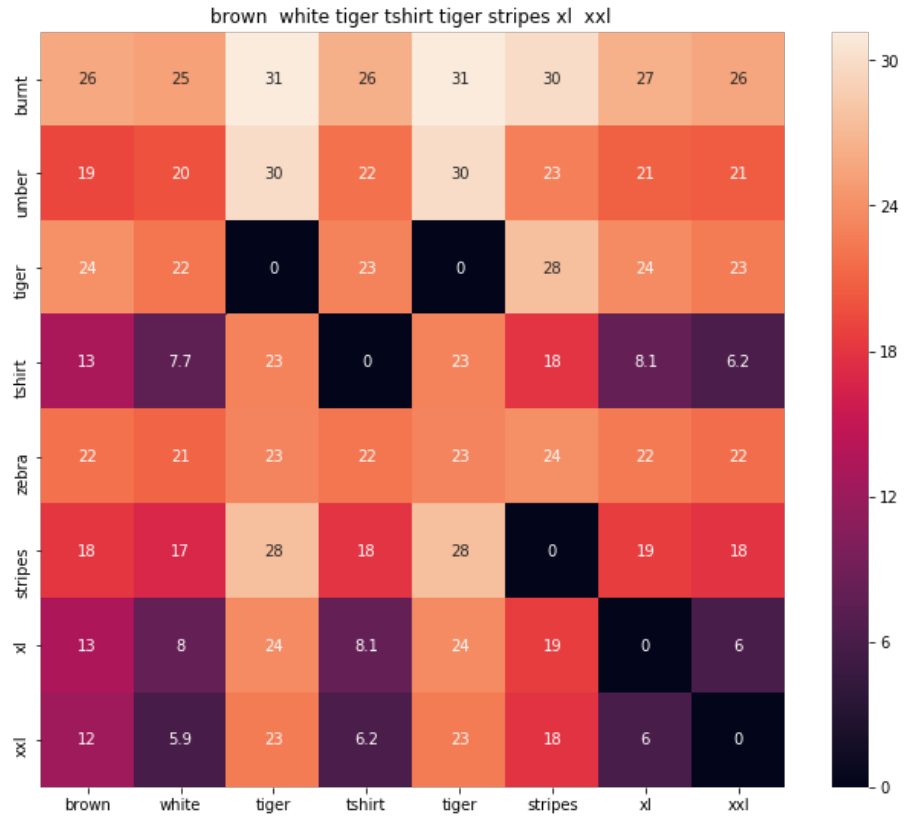
```
        print('euclidean distance from input :', pdists[i])
        print('='*125)

weighted_w2v_model(1416, 20)
#931
#12566
# in the give heat map, each cell contains the euclidean distance between words i, j
```



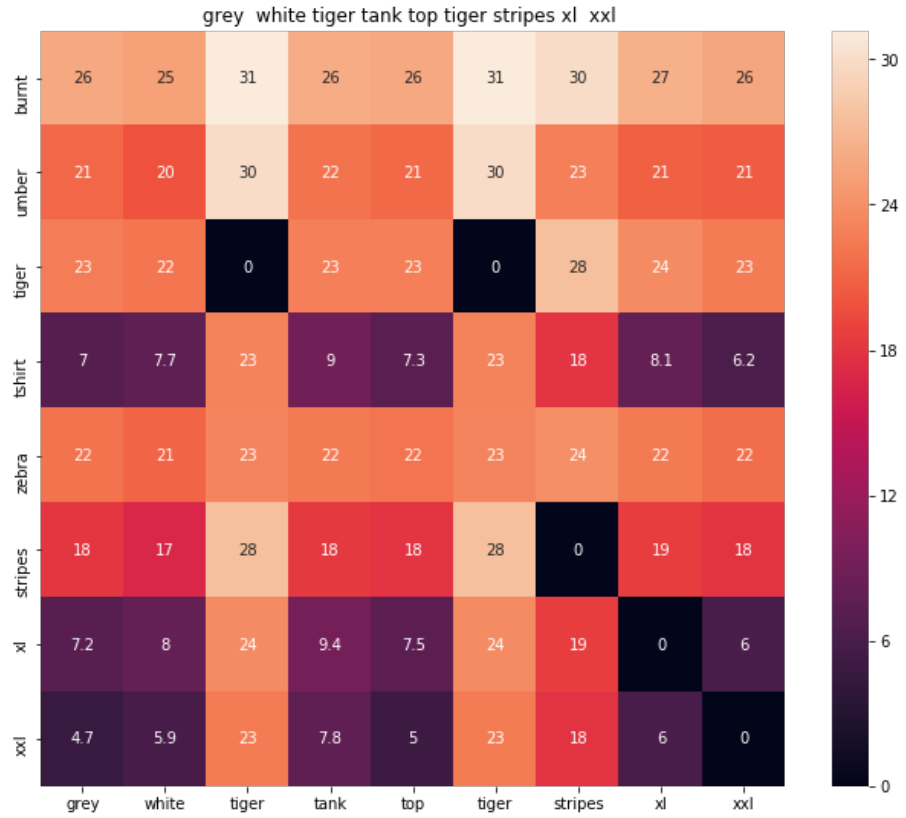
```
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 0.0
```



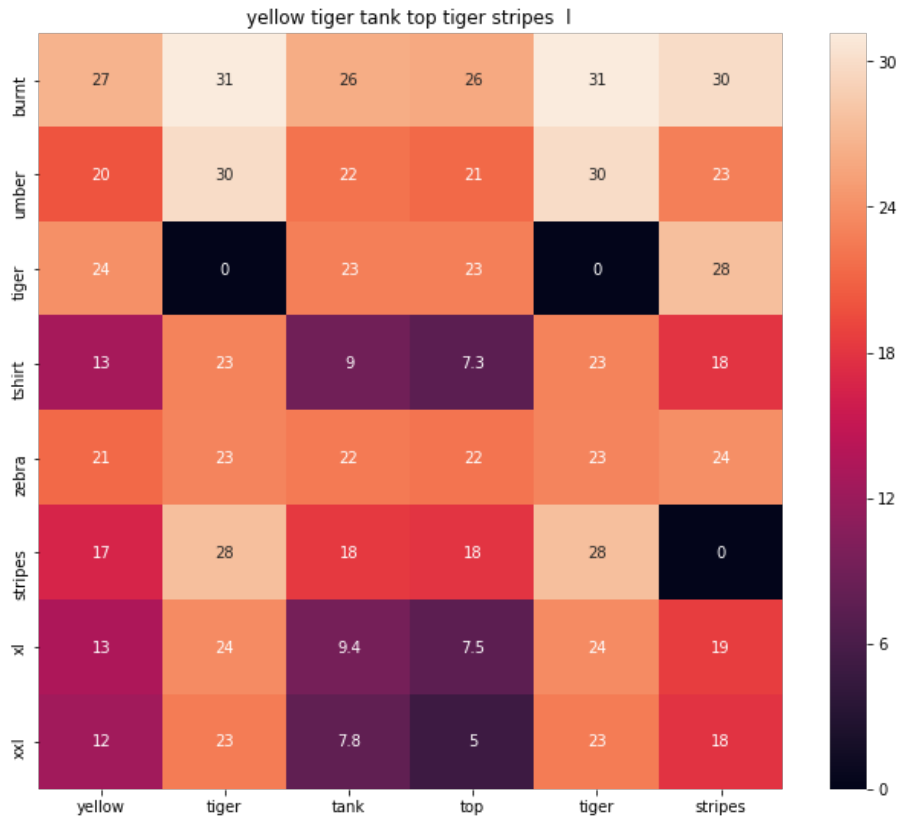
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 4.0638866



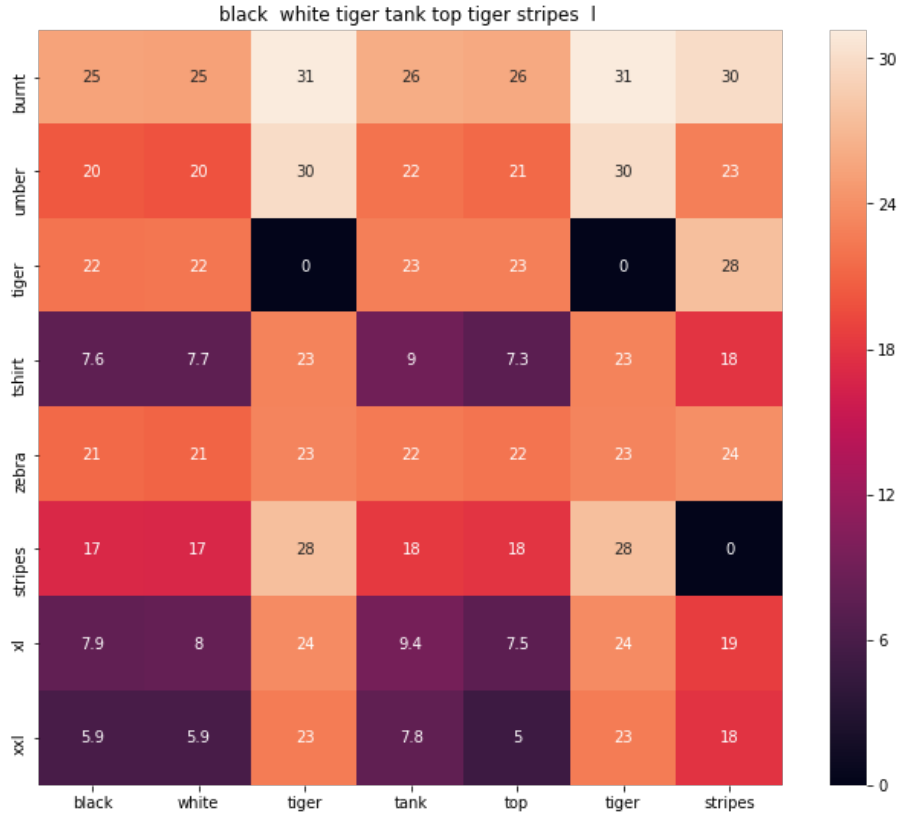
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 4.7709413



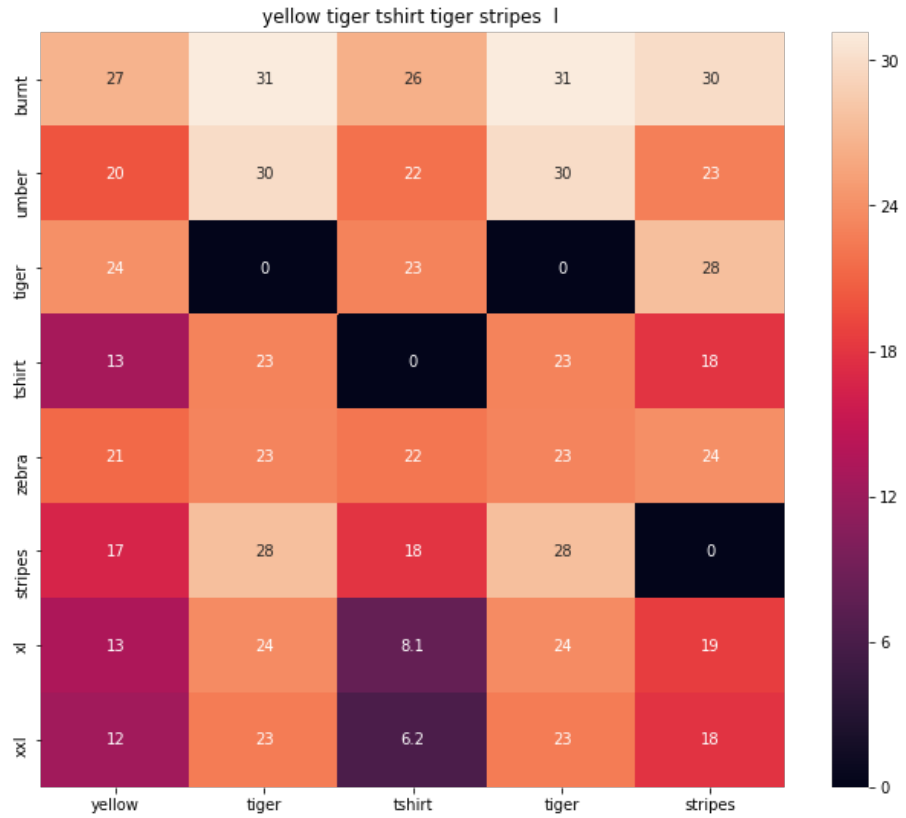
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 5.3601604



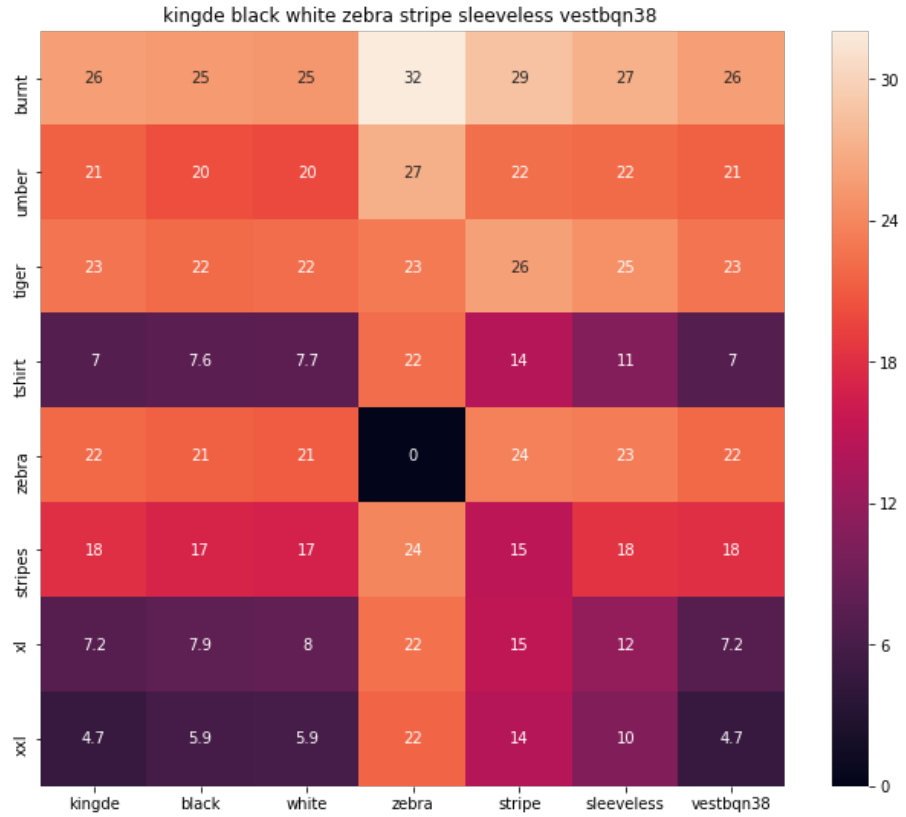
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 5.6895227



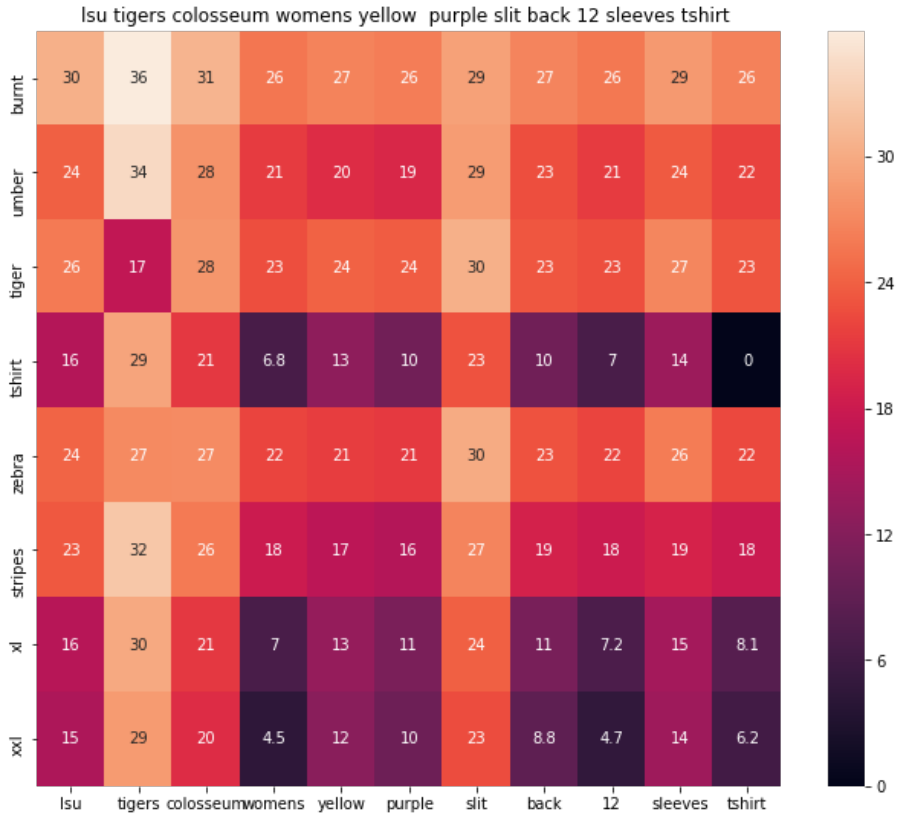
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 5.693021



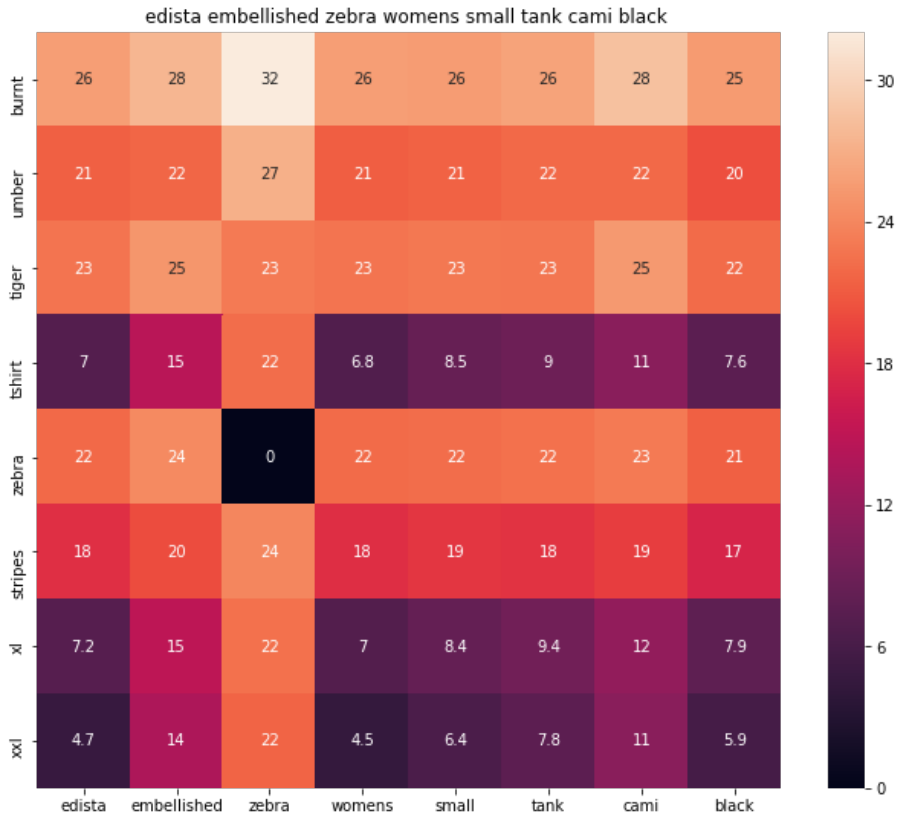
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 5.893442



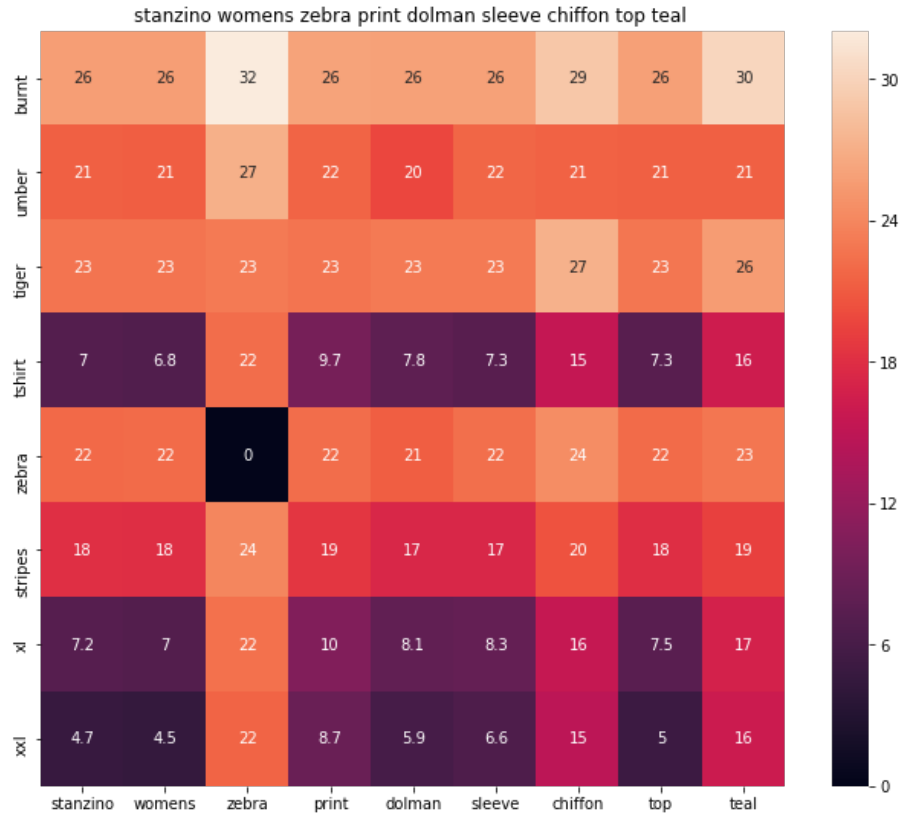
ASIN : B00JXQB5FQ
Brand : KINGDE
euclidean distance from input : 6.1329894
==========
==========



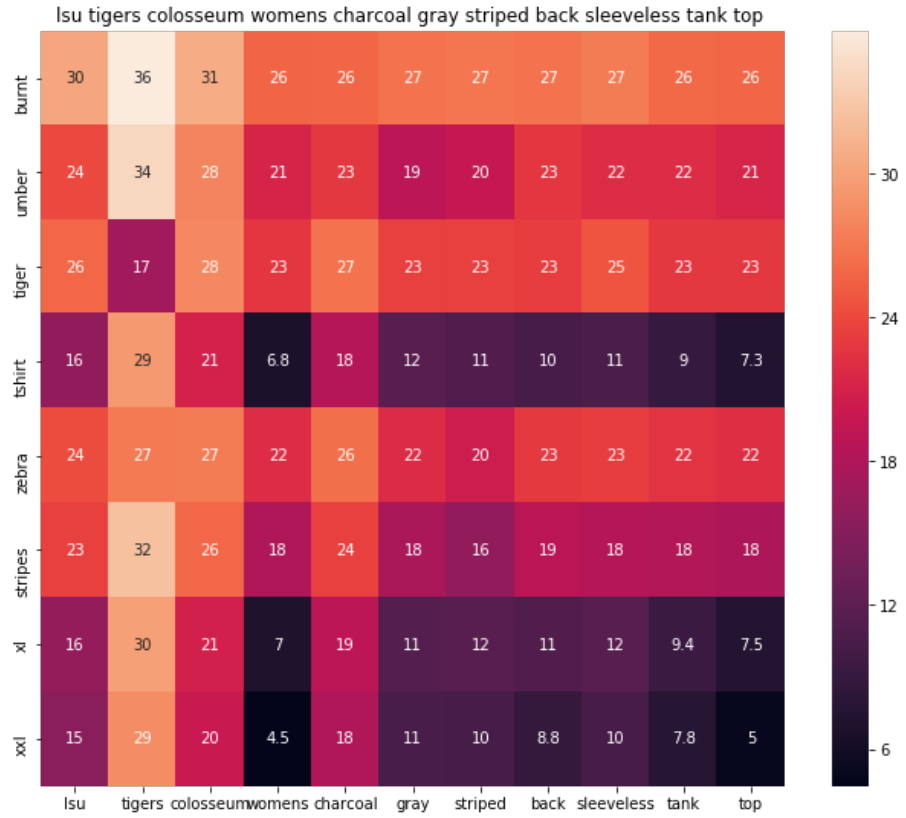
ASIN : B00JXQB5FQ
Brand : Colosseum
euclidean distance from input : 6.2567053
==========
==========



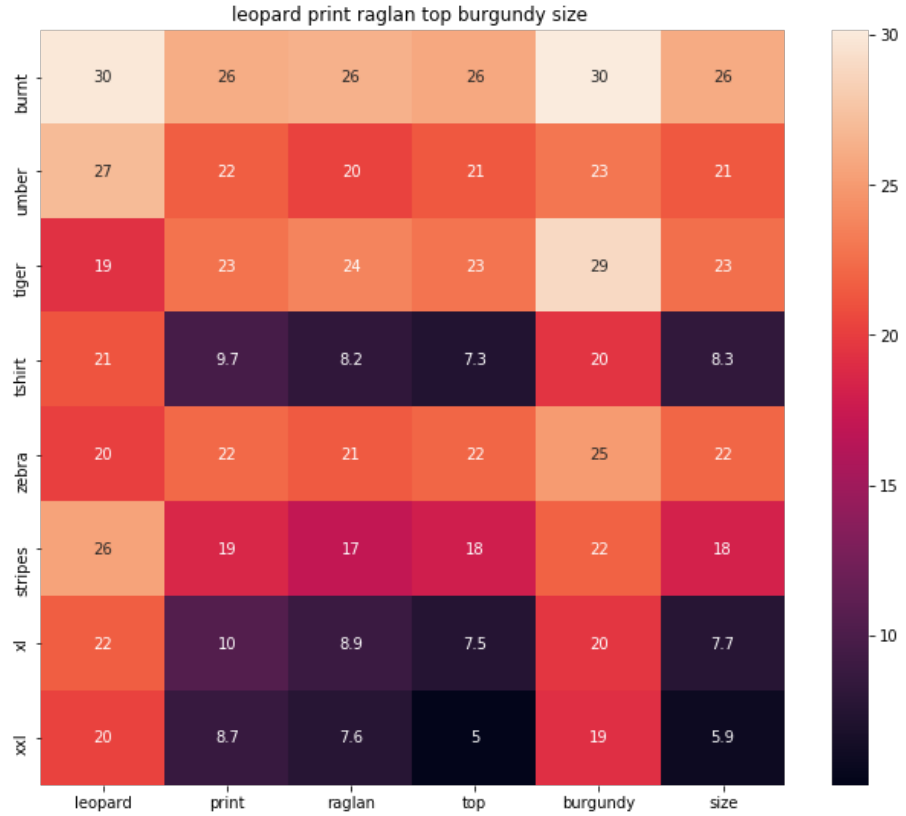
ASIN : B00JXQB5FQ
Brand : Edista
euclidean distance from input : 6.3922033
==========



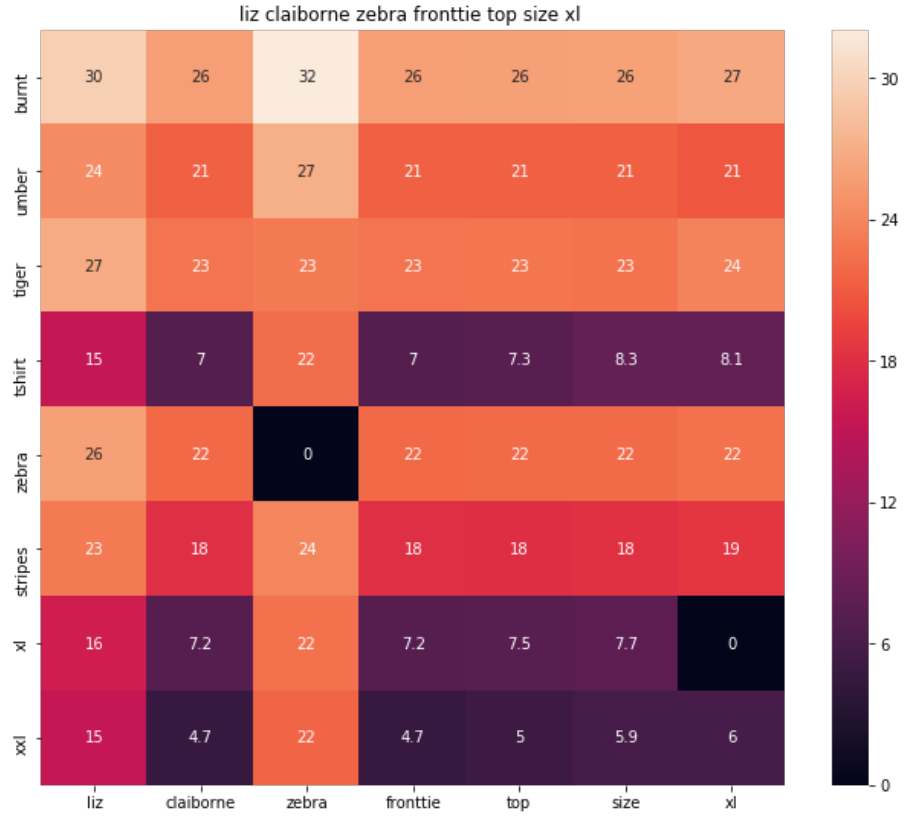
ASIN : B00JXQB5FQ
Brand : Stanzino
euclidean distance from input : 6.4149003
==========



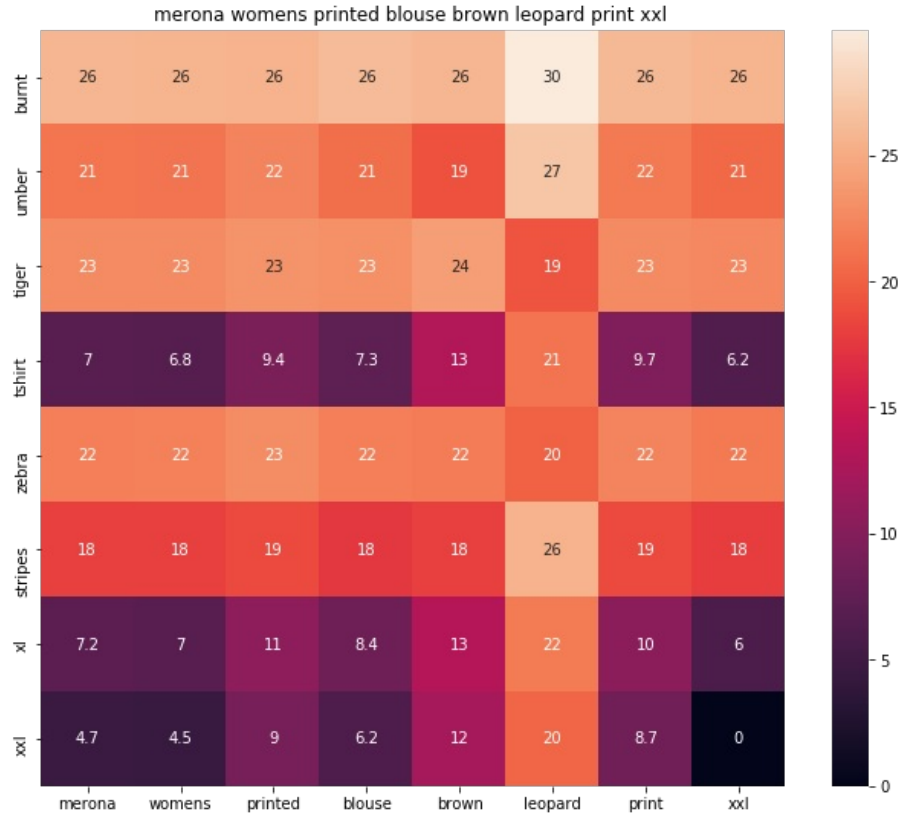
ASIN : B00JXQB5FQ
Brand : Colosseum
euclidean distance from input : 6.450959
==========



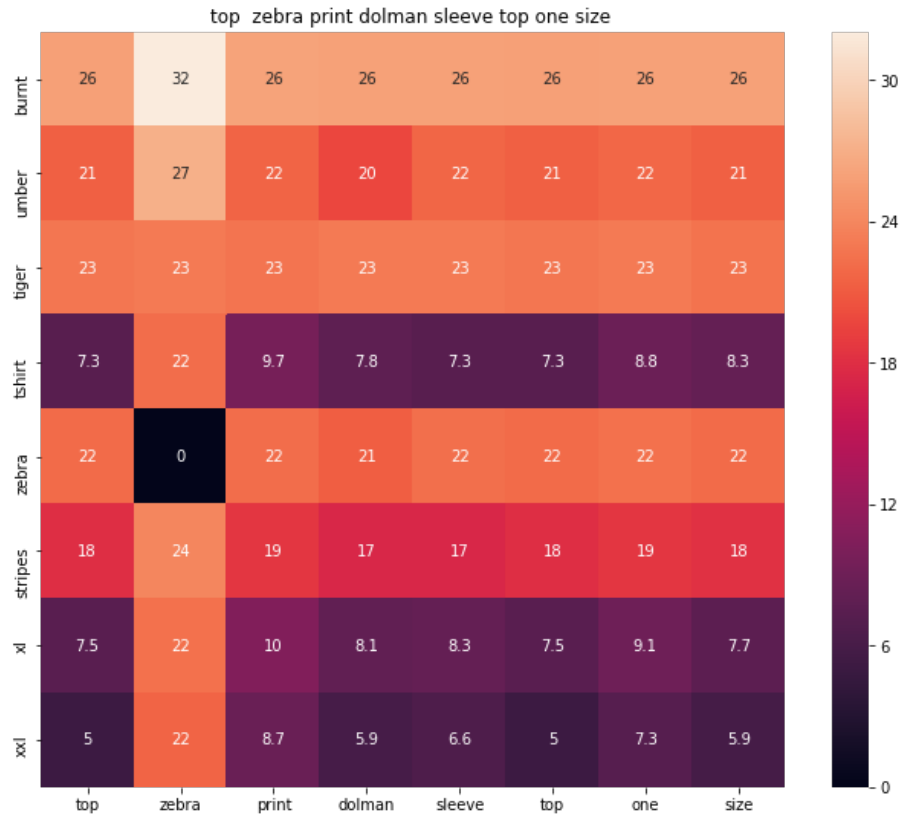
ASIN : B00JXQB5FQ
Brand : 1 Mad Fit
euclidean distance from input : 6.463409
==========



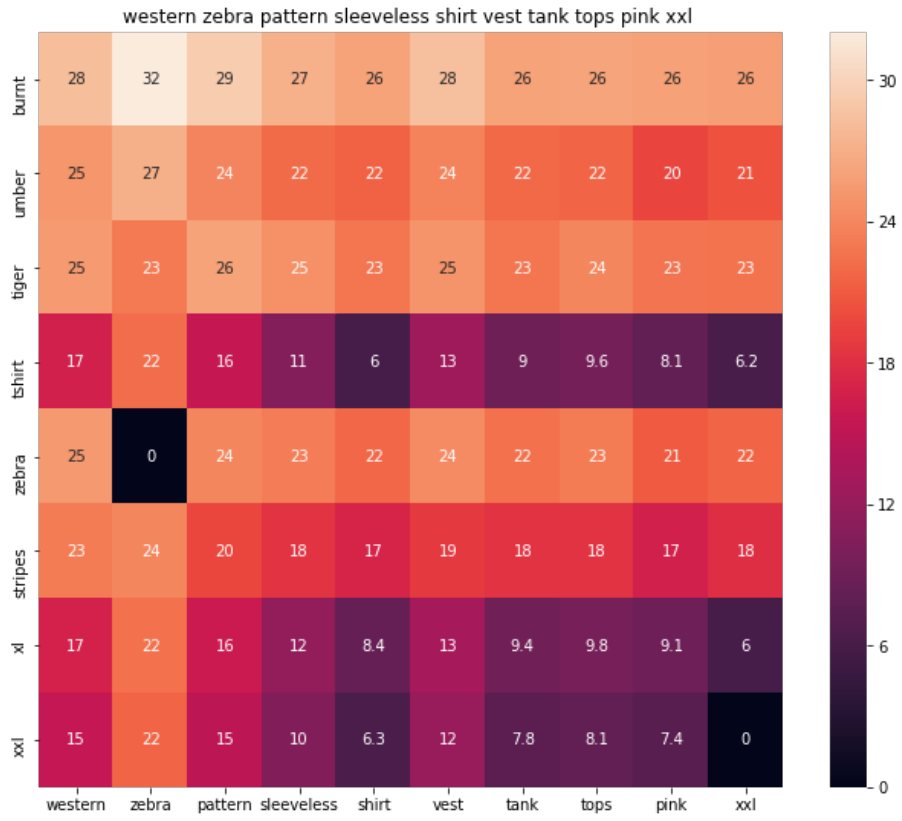
ASIN : B00JXQB5FQ
Brand : Liz Claiborne
euclidean distance from input : 6.5392227
==========



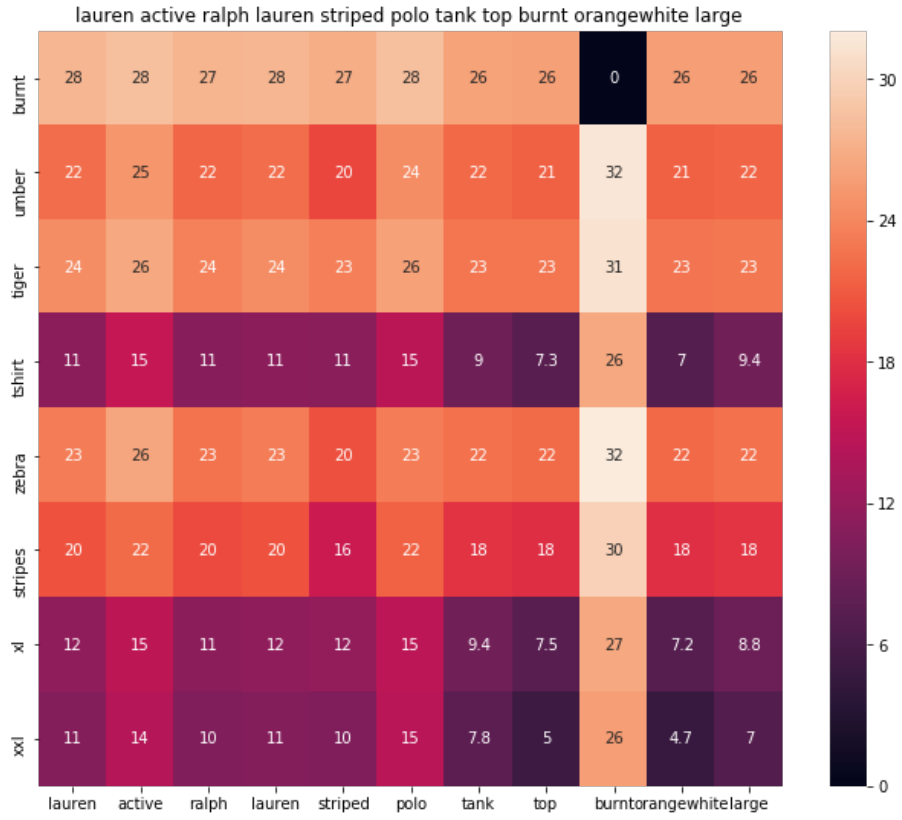
ASIN : B00JXQB5FQ
Brand : Merona
euclidean distance from input : 6.5755024
==========



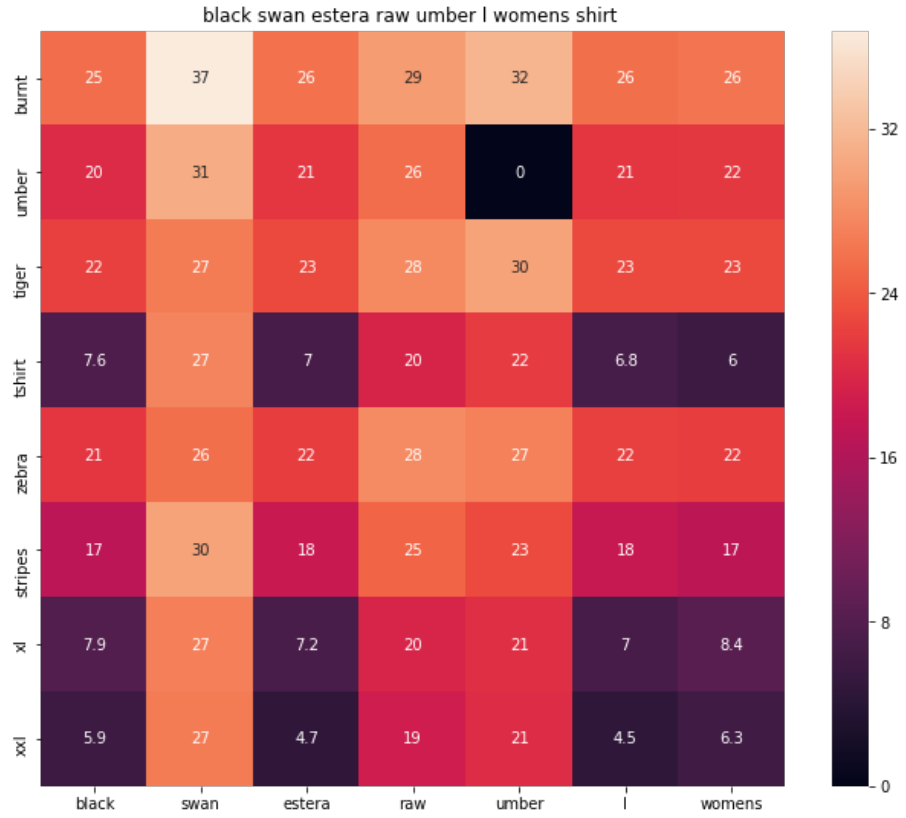
ASIN : B00JXQB5FQ
Brand : Vivian's Fashions
euclidean distance from input : 6.6382146



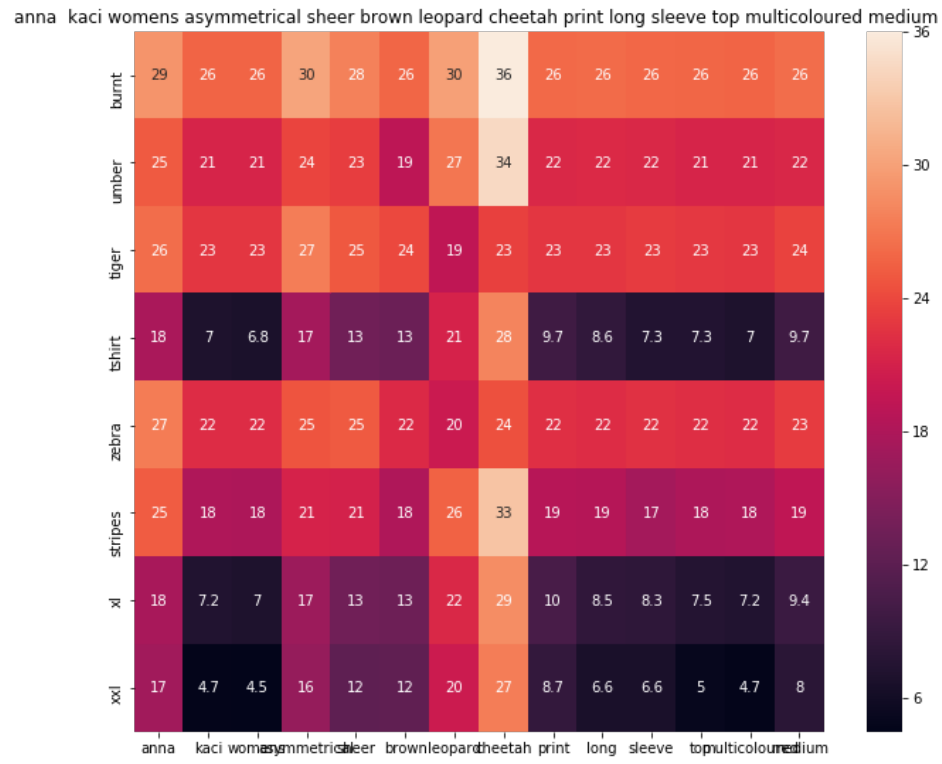
ASIN : B00JXQB5FQ
Brand : Black Temptation
euclidean distance from input : 6.6607366



ASIN : B00JXQB5FQ
Brand : Ralph Lauren Active
euclidean distance from input : 6.6839046
==========



ASIN : B00JXQB5FQ
Brand : Black Swan
euclidean distance from input : 6.705763
==========



ASIN : B00JXQB5FQ
Brand : Anna-Kaci

```
Brand : Allen Solly
euclidean distance from input : 6.7061243
================================================================================================================
======================
```

[9.6] Weighted similarity using brand and color.

In [40]:

```
# some of the brand values are empty.
# Need to replace Null with string "NULL"
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()
```

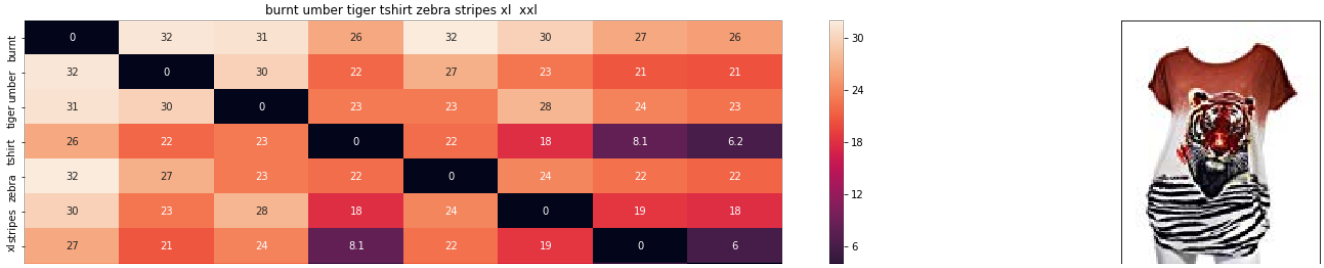
In [41]:

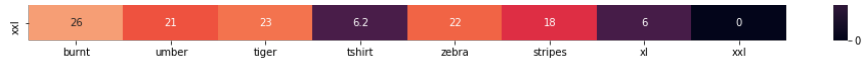
```
def heat_map_w2v_brand(sentance1, sentance2, url, doc_id1, doc_id2, df_id1, df_id2, model):

    # sentance1 : title1, input apparel
    # sentance2 : title2, recommended apparel
    # url: apparel image url
    # doc_id1: document id of input apparel
    # doc_id2: document id of recommended apparel
    # df_id1: index of document1 in the data frame
    # df_id2: index of document2 in the data frame
    # model: it can have two values, 1. avg 2. weighted

    #s1_vec = np.array(#number_of_words_title1 * 300), each row is a vector(weighted/avg) of lengt
h 300 corresponds to each word in give title
    s1_vec = get_word_vec(sentance1, doc_id1, model)
    #s2_vec = np.array(#number_of_words_title2 * 300), each row is a vector(weighted/avg) of lengt
h 300 corresponds to each word in give title
    s2_vec = get_word_vec(sentance2, doc_id2, model)

    # s1_s2_dist = np.array(#number of words in title1 * #number of words in title2)
    # s1_s2_dist[i,j] = euclidean distance between words i, j
    s1_s2_dist = get_distance(s1_vec, s2_vec)

    data_matrix = [['Asin','Brand', 'Color', 'Product type'],
                [data['asin'].loc[df_id1],brands[doc_id1], colors[doc_id1], types[doc_id1]], # input
apparel's features
                [data['asin'].loc[df_id2],brands[doc_id2], colors[doc_id2], types[doc_id2]]] #
recommonded apparel's features

    colorscale = [[0, '#1d004d'],[.5, '#f2e5ff'],[1, '#f2e5d1']] # to color the headings of each cc
lumn

    # we create a table with the data_matrix
    table = ff.create_table(data_matrix, index=True, colorscale=colorscale)
    # plot it with plotly
    plotly.offline.iplot(table, filename='simple_table')

    # devide whole figure space into 25 * 1:10 grids
    gs = gridspec.GridSpec(25, 15)
    fig = plt.figure(figsize=(25,5))

    # in first 25*10 grids we plot heatmap
    ax1 = plt.subplot(gs[:, :-5])
    # ploting the heap map based on the pairwise distances
```

```
    ax1 = sns.heatmap(np.round(s1_s2_dist,6), annot=True)
    # set the x axis labels as recommended apparels title
    ax1.set_xticklabels(sentance2.split())
    # set the y axis labels as input apparels title
    ax1.set_yticklabels(sentance1.split())
    # set title as recommended apparels title
    ax1.set_title(sentance2)

    # in last 25 * 10:15 grids we display image
    ax2 = plt.subplot(gs[:, 10:16])
    # we dont display grid lins and axis labels to images
    ax2.grid(False)
    ax2.set_xticks([])
    ax2.set_yticks([])

    # pass the url it display it
    display_img(url, ax2, fig)

    plt.show()
```

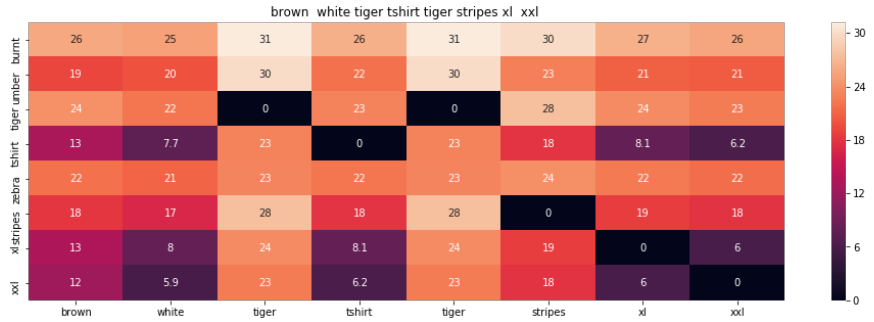
In [43]:

```
def idf_w2v_brand(doc_id, w1, w2, num_results):
    # doc_id: apparel's id in given corpus
    # w1: weight for  w2v features
    # w2: weight for brand and color features

    # pairwise_dist will store the distance from given input apparel to all remaining apparels
    # 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
    idf_w2v_dist  = pairwise_distances(w2v_title_weight, w2v_title_weight[doc_id].reshape(1,-1))
    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)):
        heat_map_w2v_brand(data['title'].loc[df_indices[0]],data['title'].loc[df_indices[i]], data[
'medium_image_url'].loc[df_indices[i]], indices[0], indices[i],df_indices[0], df_indices[i], 'weigh
ted')
        print('ASIN :',data['asin'].loc[df_indices[i]])
        print('Brand :',data['brand'].loc[df_indices[i]])
        print('euclidean distance from input :', pdists[i])
        print('='*125)

idf_w2v_brand(1416, 5, 5, 20)
# in the give heat map, each cell contains the euclidean distance between words i, j
```

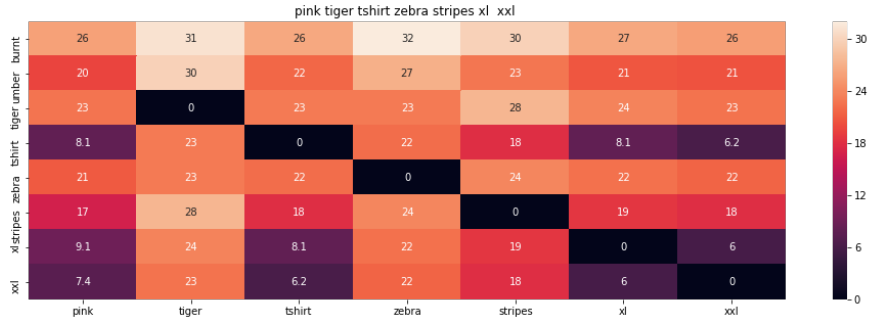




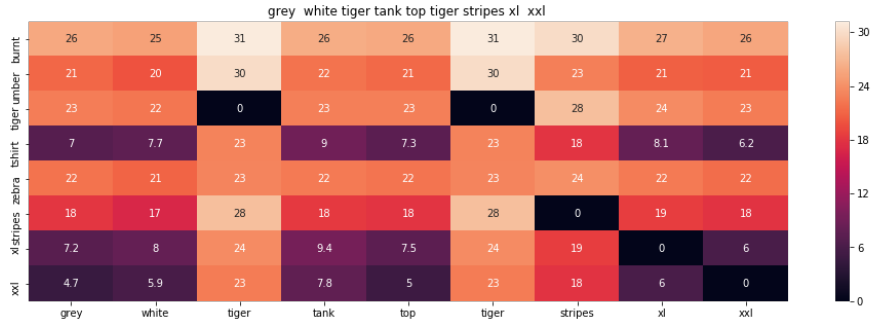
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 0.0



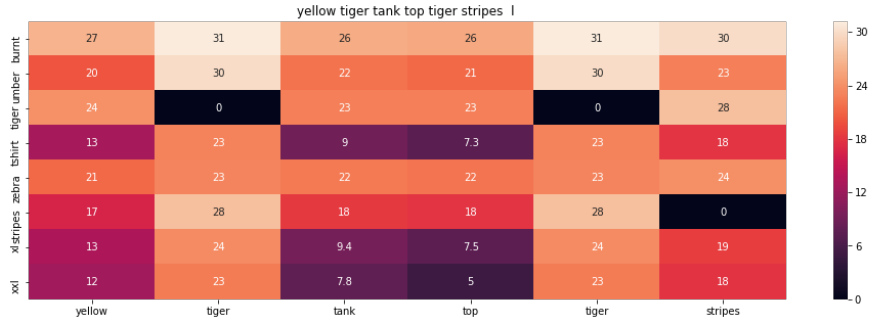
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.3854705810546877



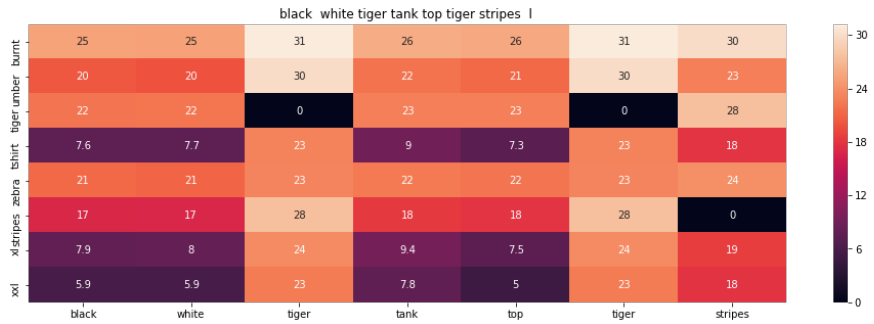
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.739050102414575



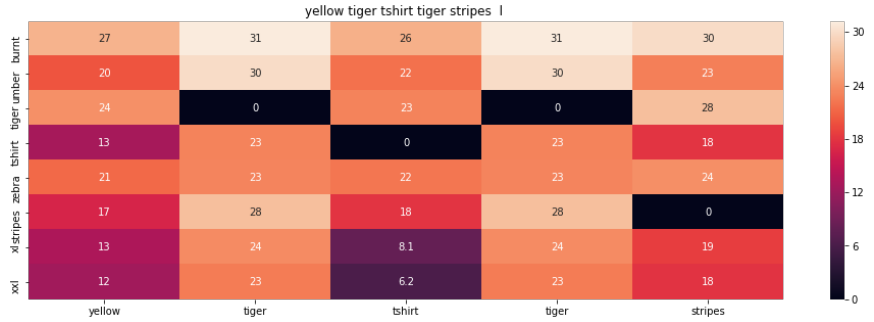
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 3.387187004270044
==========
==========



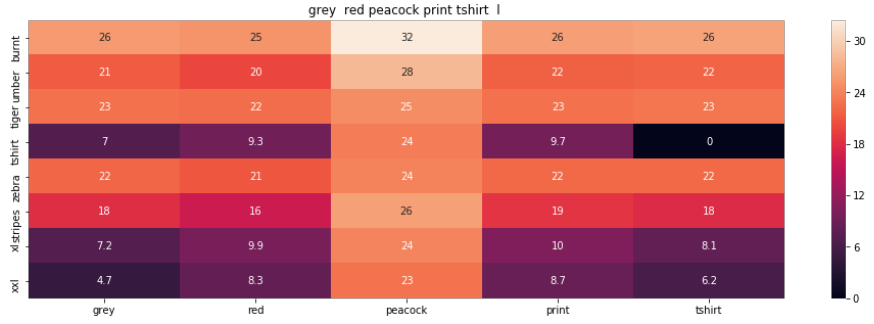
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 3.5518680574316646
==========
==========



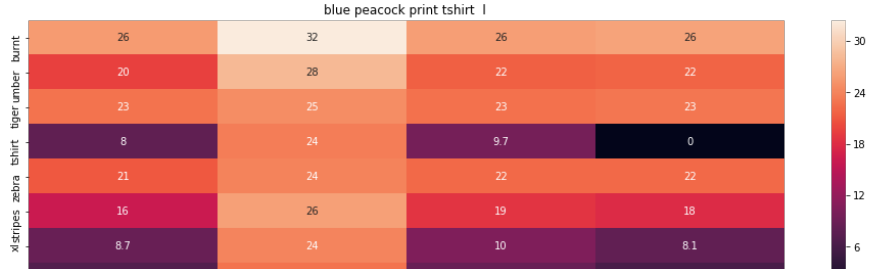
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 3.5536170961279536
==========
==========



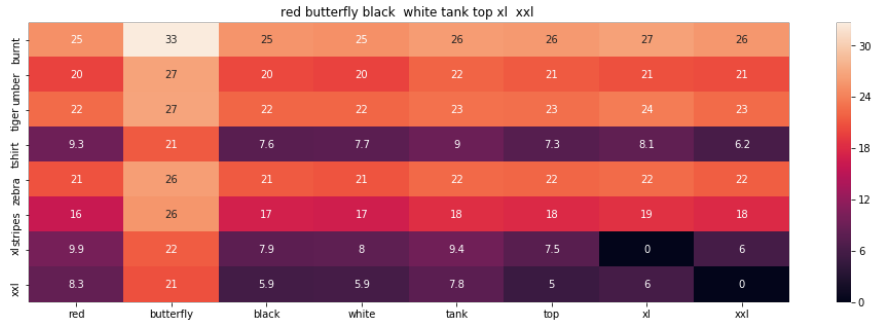
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 3.6538278581518795



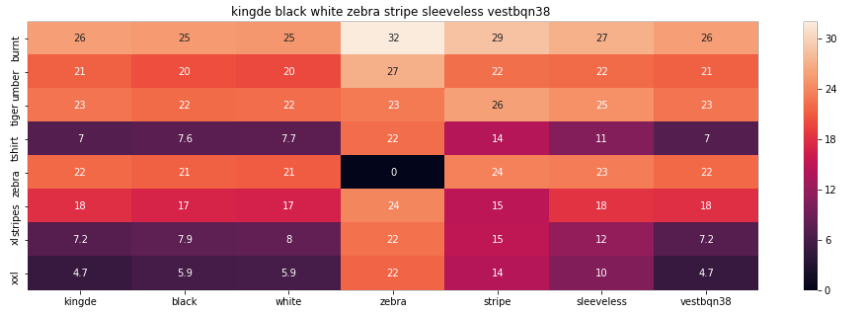
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 4.128811264218774



ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 4.203900146665063
==========
==========

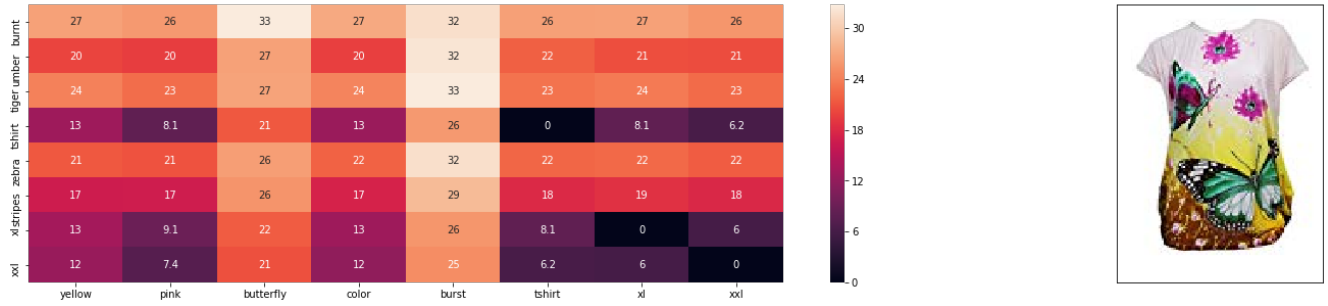


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 4.286586380185571
==========
==========

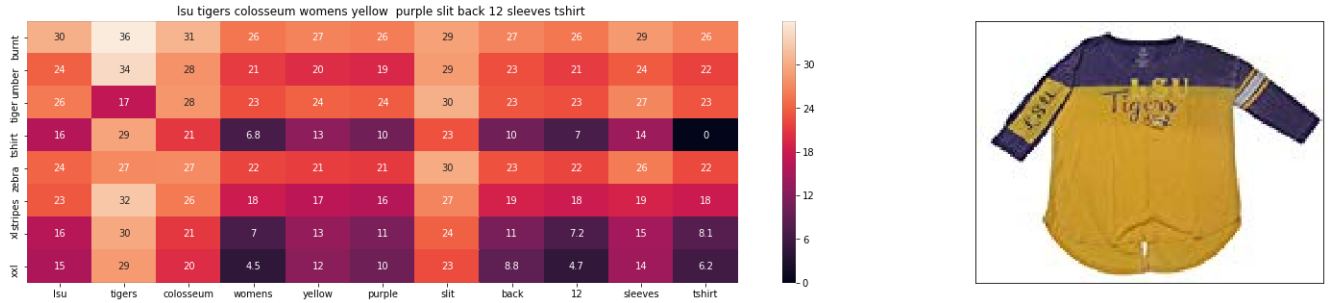


ASIN : B00JXQB5FQ
Brand : KINGDE
euclidean distance from input : 4.389370406508858
==========
==========

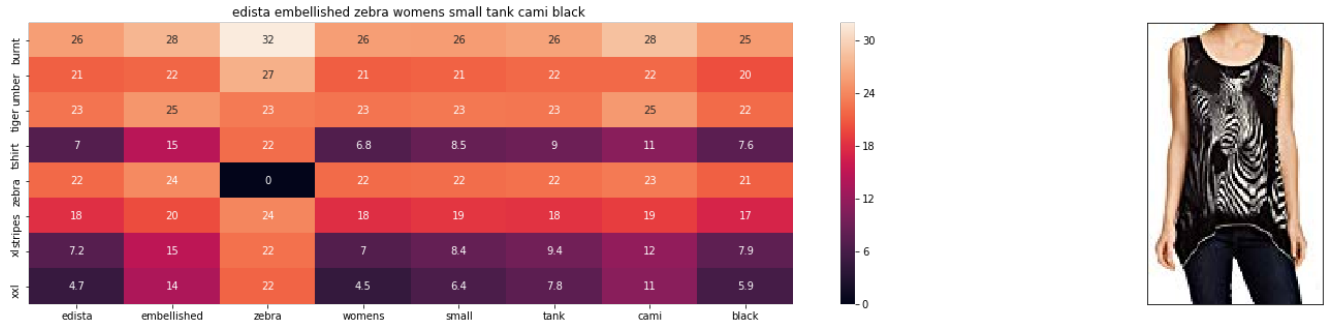
yellow pink butterfly color burst tshirt xl xxl



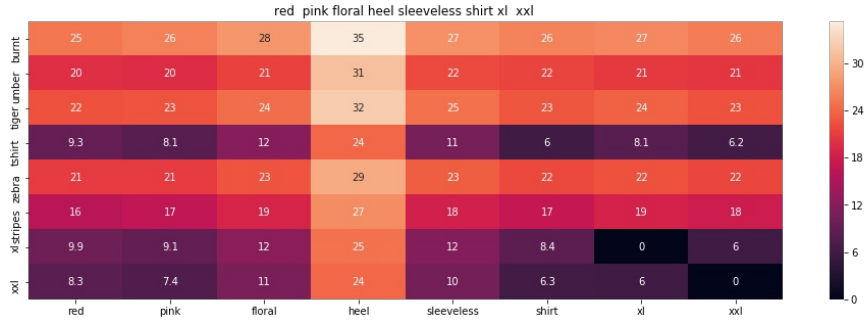
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 4.397909927548852
==========



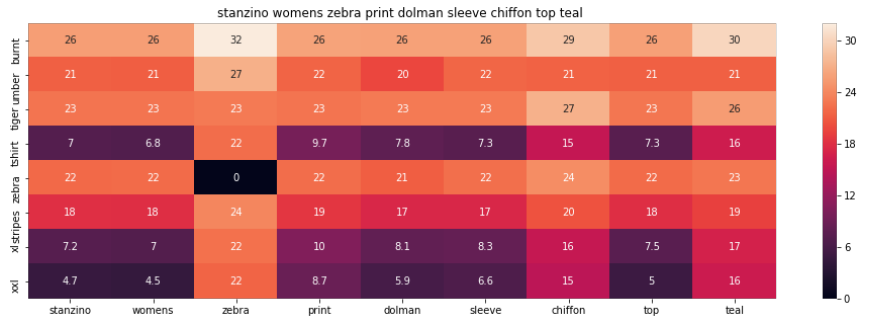
ASIN : B00JXQB5FQ
Brand : Colosseum
euclidean distance from input : 4.451228392959053
==========



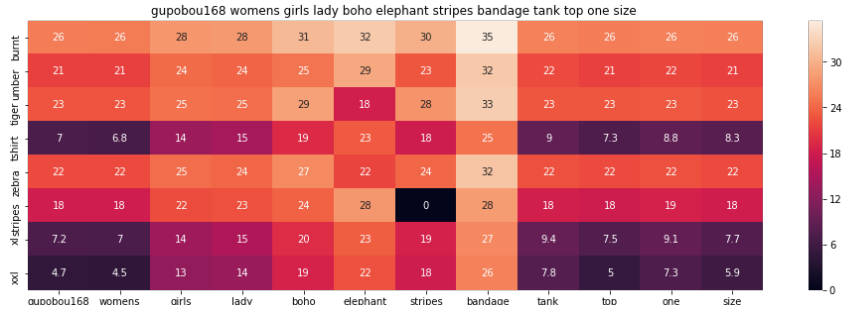
ASIN : B00JXQB5FQ
Brand : Edista
euclidean distance from input : 4.518977416396553
==========



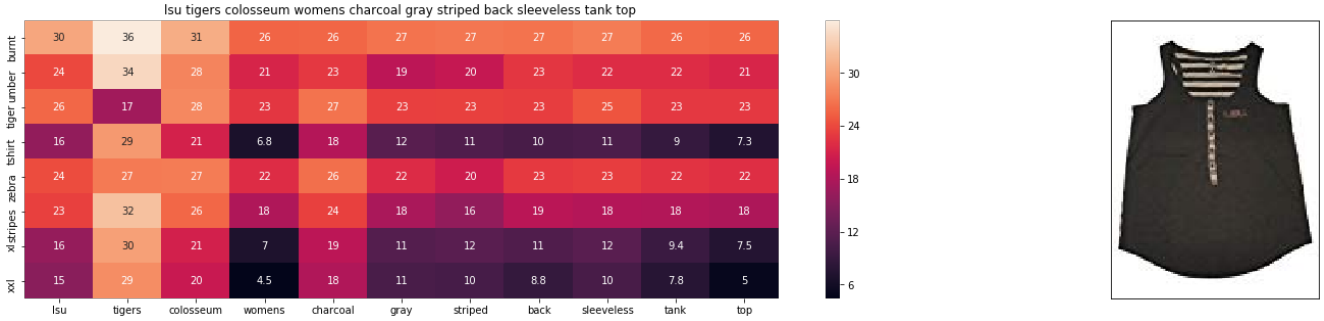
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 4.529374695004907



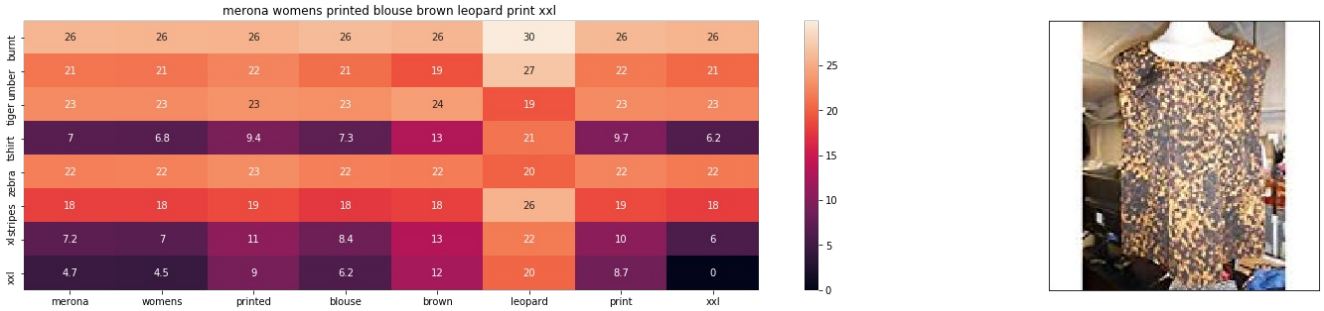
ASIN : B00JXQB5FQ
Brand : Stanzino
euclidean distance from input : 4.530325759292061



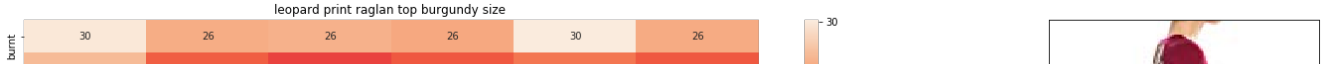
ASIN : B00JXQB5FQ
Brand : GuPoBoU168
euclidean distance from input : 4.546816642558488
==========

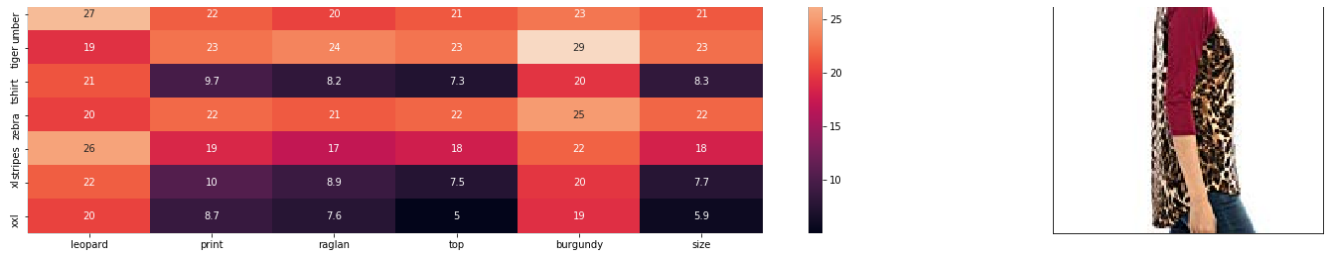


ASIN : B00JXQB5FQ
Brand : Colosseum
euclidean distance from input : 4.548355162978584
==========



ASIN : B00JXQB5FQ
Brand : Merona
euclidean distance from input : 4.610626662612374
==========

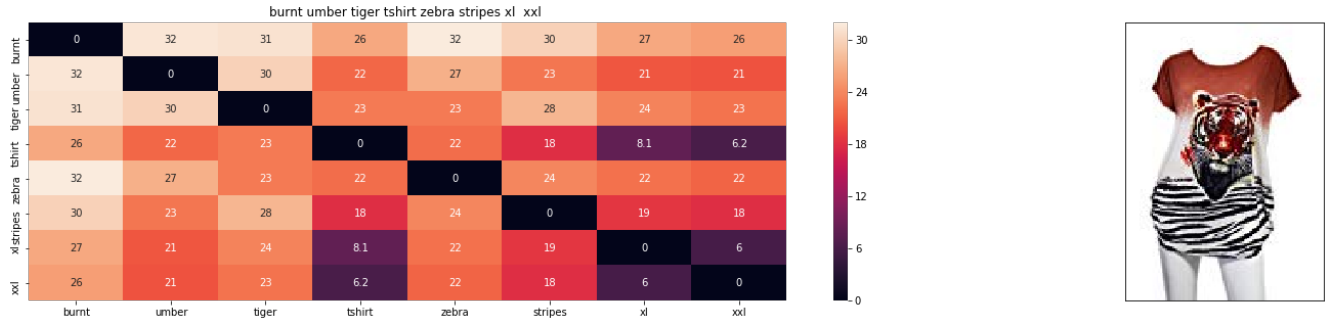




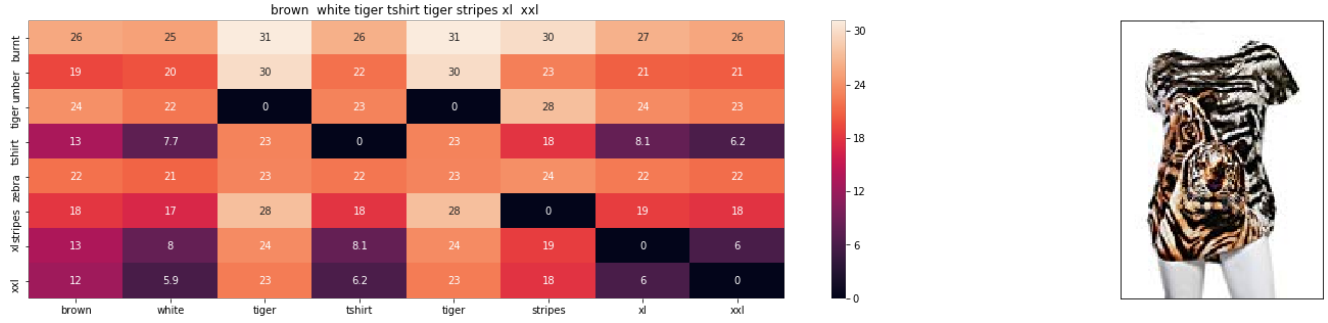
ASIN : B00JXQB5FQ
Brand : 1 Mad Fit
euclidean distance from input : 4.645917892817431
==========

In [44]:

```
# brand and color weight =50
# title vector weight = 5

idf_w2v_brand(1416, 5, 50, 20)
```

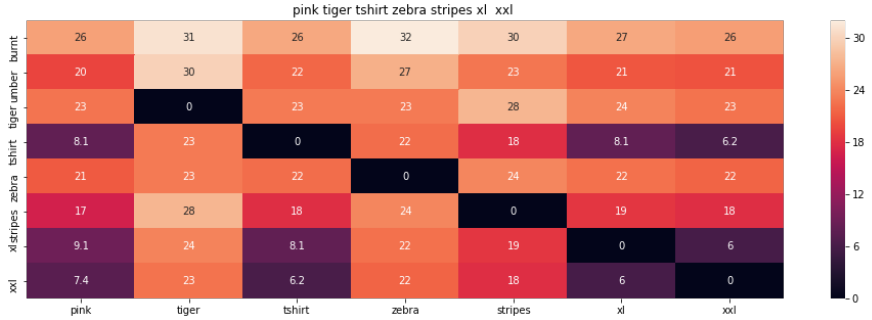


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 0.0
==========

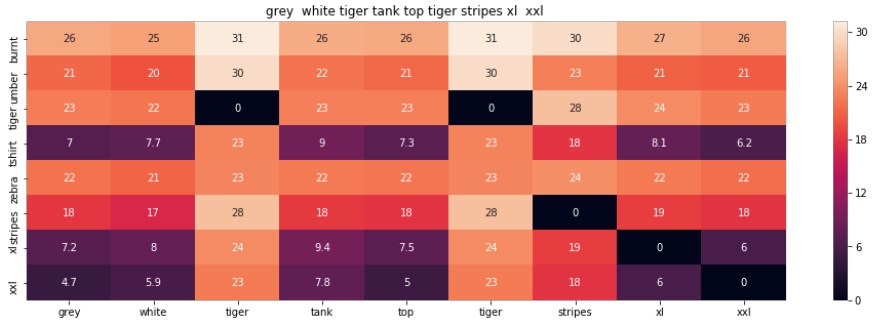


ASIN : B00JXQB5FQ

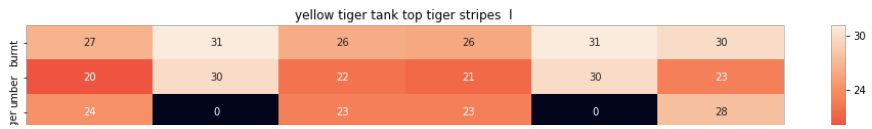
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 0.433721923828125
==========

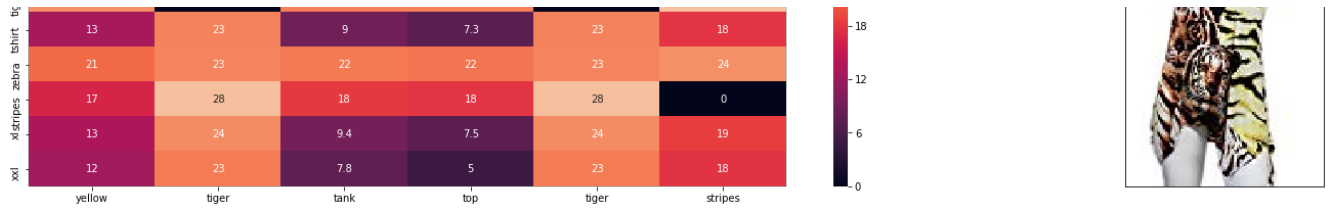


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.6550929332897277
==========



ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.7729360063543584
==========

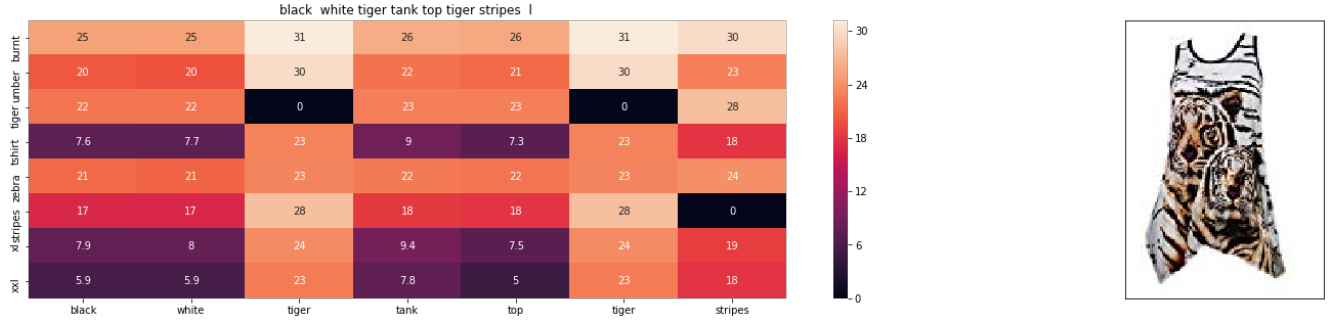




ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.8028780160201077

==========

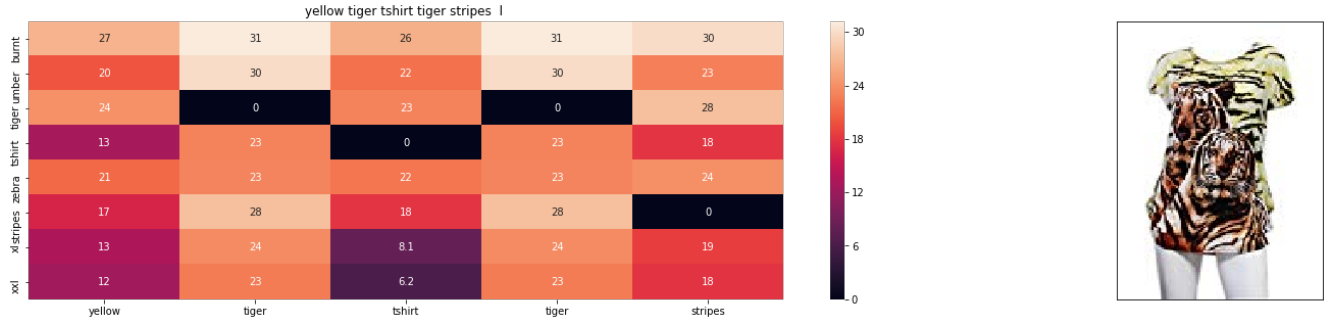
==========



ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.8031960230557966

==========

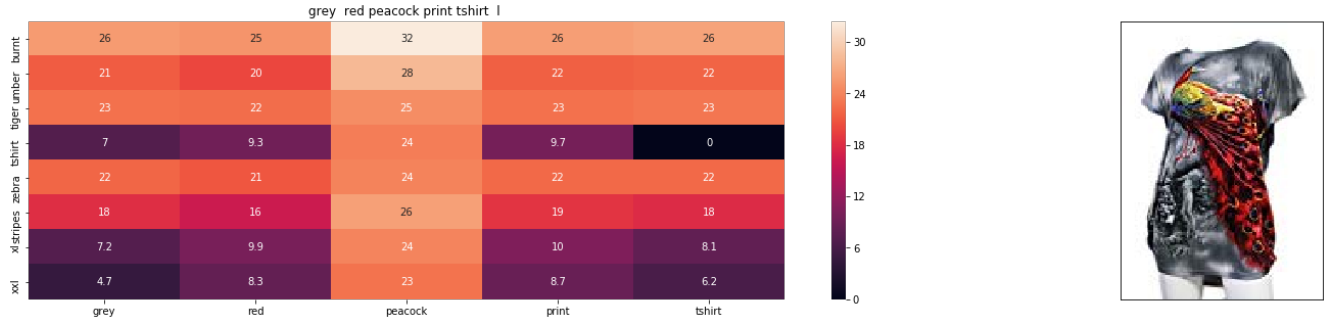
==========



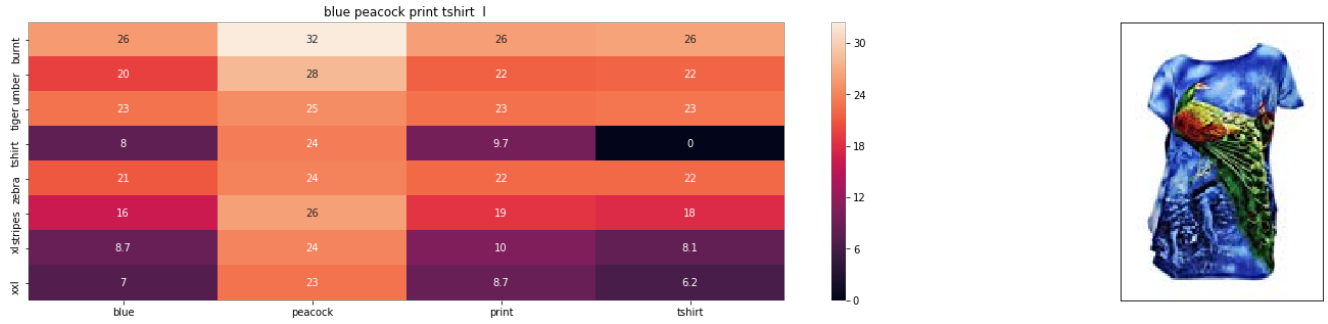
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.8214161616056013

==========

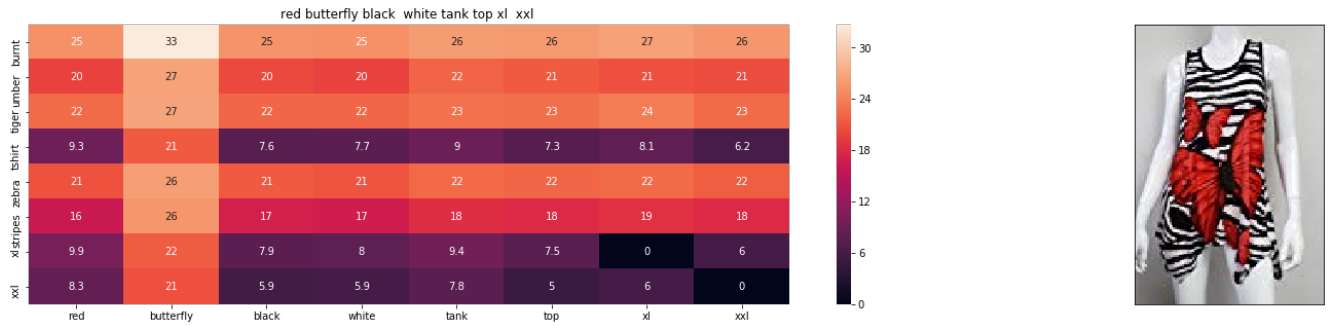
==========



ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.9077767808904913
==========

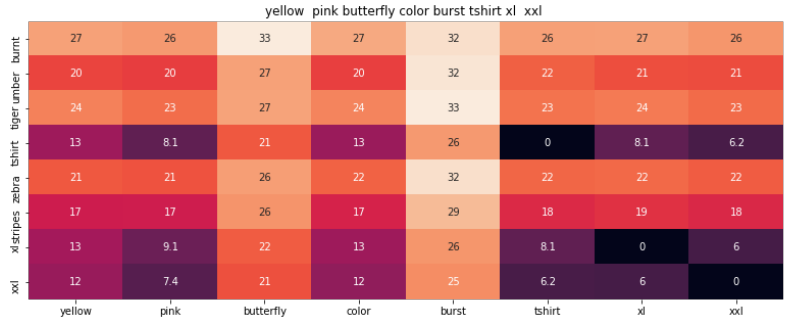


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.9214293049716347
==========

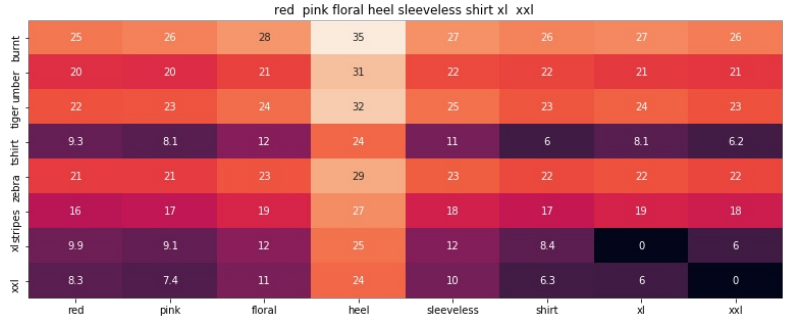


ASIN : B00JXQB5FQ

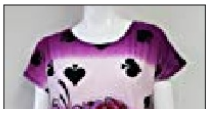
Brand : Si Row
euclidean distance from input : 1.936463165611727

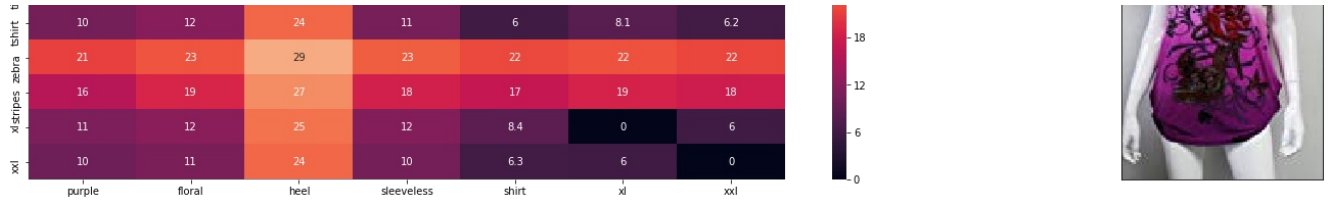


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.956703810586869

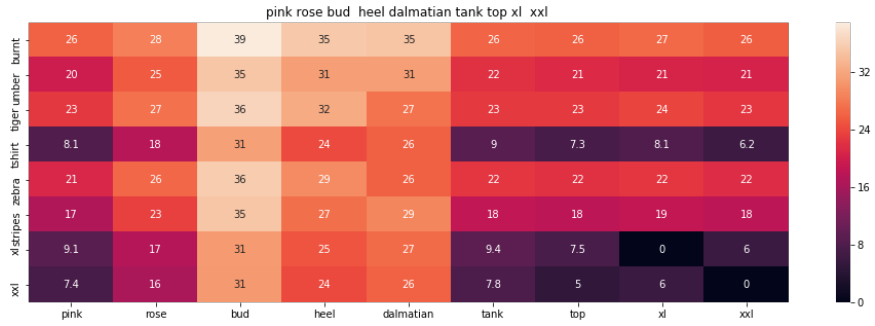


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.9806064955788791

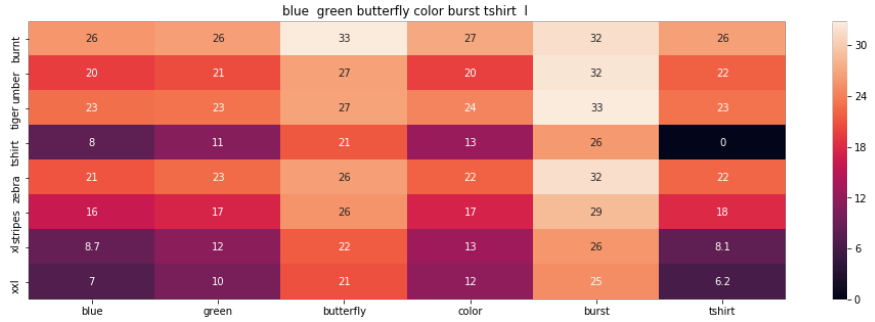




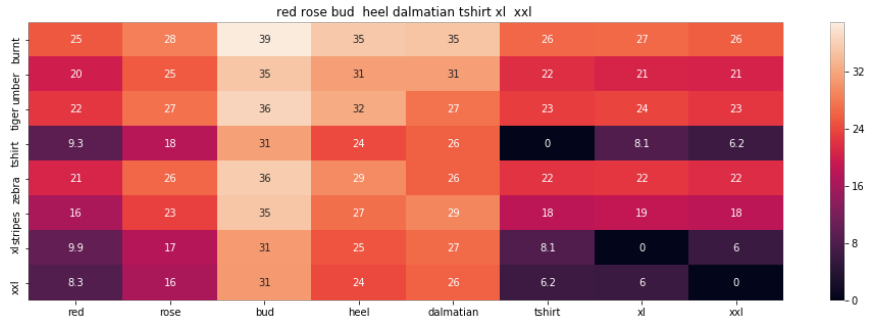
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.012185530557572
==========



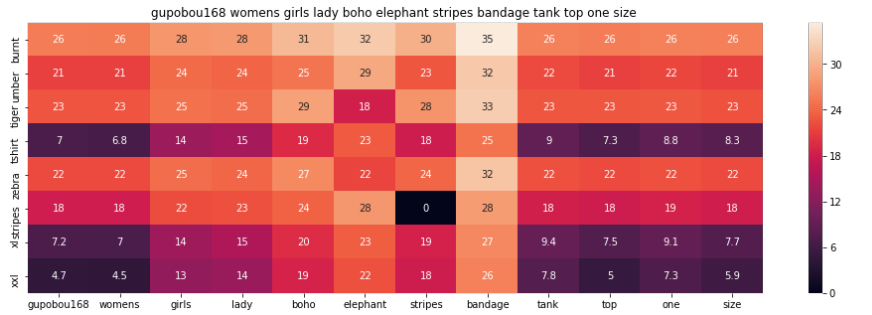
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.01335171819074
==========



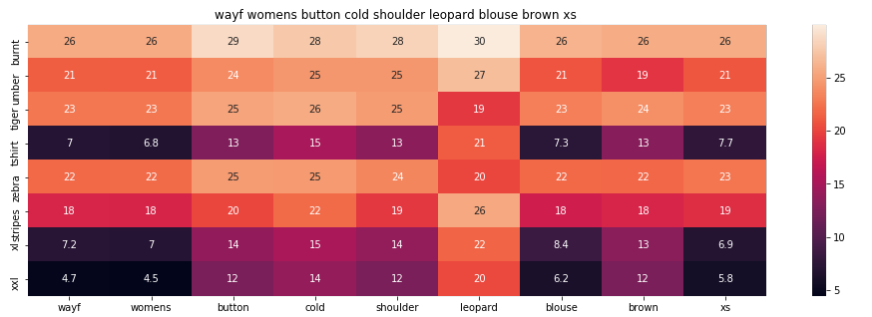
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.0138832789151717
==========



ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.0367257554998663
==========
==========

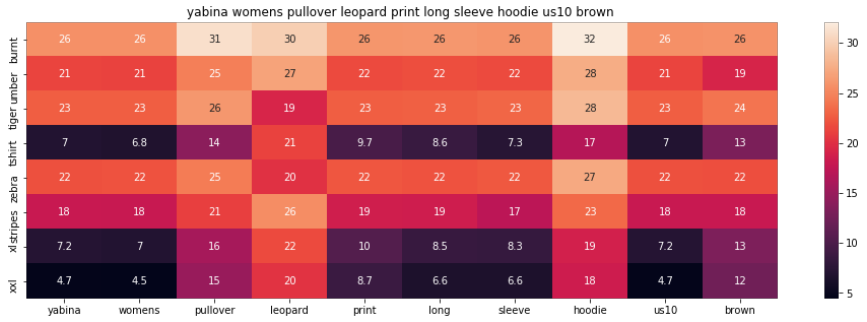


ASIN : B00JXQB5FQ
Brand : GuPoBoU168
euclidean distance from input : 2.656204098419553
==========
==========

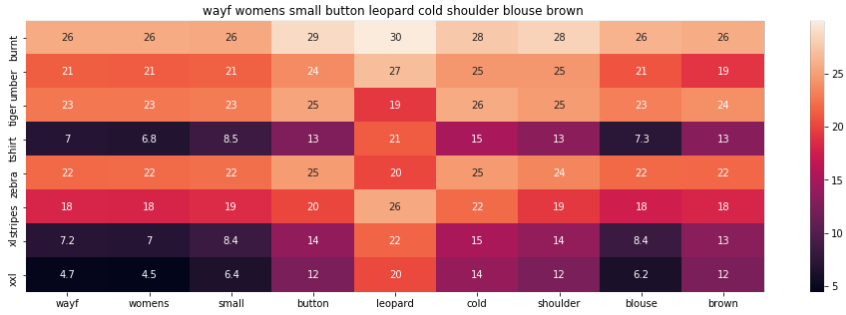


ASIN : B00JXQB5FQ
Brand : WAYF

Brand : WAYF
euclidean distance from input : 2.6849067129437008
==========



ASIN : B00JXQB5FQ
Brand : YABINA
euclidean distance from input : 2.6858381232997024
==========



ASIN : B00JXQB5FQ
Brand : WAYF
euclidean distance from input : 2.694761948650377
==========

[10.2] Keras and Tensorflow to extract features

In [3]:

```
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
```

```
from PIL import Image
import pandas as pd
import pickle
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-3-854a3fb72b56> in <module>()
      1 import numpy as np
----> 2 from keras.preprocessing.image import ImageDataGenerator
      3 from keras.models import Sequential
      4 from keras.layers import Dropout, Flatten, Dense
      5 from keras import applications

ModuleNotFoundError: No module named 'keras'
```

In [35]:

```
# https://gist.github.com/fchollet/f35fbc80e066a49d65f1688a7e99f069
# Code reference: https://blog.keras.io/building-powerful-image-classification-models-using-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()

'''
```

Out[35]:

```
"\n# dimensions of our images.\nimg_width, img_height = 224, 224\n\ntop_model_weights_path = 'bott
leneck_fc_model.h5'\ntrain_data_dir = 'images2/'\nnb_train_samples = 16042\nepochs =
50\nbatch_size = 1\n\n\ndef save_bottlebeck_features():\n    \n    #Function to compute VGG-16 CNN
for image feature extraction.\n    \n    asins = []\n    datagen = ImageDataGenerator(rescale=1. /
255)\n    \n    # build the VGG16 network\n    model = applications.VGG16(include_top=False, weight
s='imagenet')\n    generator = datagen.flow_from_directory(\n        train_data_dir,\n        targe
t_size=(img_width, img_height),\n        batch_size=batch_size,\n        class_mode=None,\n
shuffle=False)\n\n    for i in generator.filenames:\n        asins.append(i[2:-5])\n\n    bottlenec
k_features_train = model.predict_generator(generator, nb_train_samples // batch_size)\n
bottleneck_features_train = bottleneck_features_train.reshape((16042,25088))\n    \n
np.save(open('16k_data_cnn_features.npy', 'wb'), bottleneck_features_train)\n
np.save(open('16k_data_cnn_feature_asins.npy', 'wb'), np.array(asins))\n
\n\nsave_bottlebeck_features()\n\n"
```

In [23]:

```
data[data.title=='burnt umber tiger tshirt zebra stripes xl  xxl ']
```

Out[23]:

	asin	brand	color	medium_image_url	product_type_name	title	formatted_price
1416	B00JXQB5FQ	Si Row	Brown	https://images-na.ssl-images-amazon.com/images...	TOYS_AND_GAMES	burnt umber tiger tshirt zebra stripes xl xxl	$19.61

[10.3] Visual features based product similarity.

In [56]:

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

# load the original 16K dataset
data = pd.read_pickle('pickels/16k_apperal_data_preprocessed')

#ordering dataponts of data according to asin
df_asin=pd.DataFrame(asins,columns=['asin'])
df1 = data.set_index('asin')
df1 = df1.reindex(index=df_asin['asin'])
data = df1.reset_index()
df_asins = list(data['asin'])
from IPython.display import display, Image, SVG, Math, YouTubeVideo


#get similar products using CNN features (VGG-16)
def get_similar_products_cnn(doc_id, num_results):
    doc_id = asins.index(df_asins[doc_id])
    pairwise_dist = pairwise_distances(bottleneck_features_train, bottleneck_features_train[doc_id]
.reshape(1,-1))
    print(pairwise_dist)
    indices = np.argsort(pairwise_dist.flatten())[0:num_results]
    pdists  = np.sort(pairwise_dist.flatten())[0:num_results]
    print("In",indices)
    df_indices = list(data.index[indices])
    print(df_indices)

    for i in range(len(indices)):
        rows = data[['medium_image_url','title','asin']].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]])

    return(pairwise_dist)
```

```
_=get_similar_products_cnn(1416, 20)
```

```
[[56.22038 ]
 [52.389206]
 [54.218445]
 ...
 [58.046803]
 [53.92738 ]
 [60.428642]]
In [ 1416  1413  1421  1422 11752  6868 12191  4758  5406  6370  3732  1398
  7439  4739 12318  5896  2690  8386 13481  7779]
[1416, 1413, 1421, 1422, 11752, 6868, 12191, 4758, 5406, 6370, 3732, 1398, 7439, 4739, 12318, 5896,
2690, 8386, 13481, 7779]
```

Product Title: burnt umber tiger tshirt zebra stripes xl xxl
Euclidean Distance from input image: 6.32596e-06
Amazon Url: www.amzon.com/dp/B00JXQB5FQ

Product Title: pink tiger tshirt zebra stripes xl xxl
Euclidean Distance from input image: 30.05017
Amazon Url: www.amzon.com/dp/B00JXQASS6

Product Title: yellow tiger tshirt tiger stripes l
Euclidean Distance from input image: 41.261116
Amazon Url: www.amzon.com/dp/B00JXQCUIC

Product Title: brown white tiger tshirt tiger stripes xl xxl
Euclidean Distance from input image: 44.000156
Amazon Url: www.amzon.com/dp/B00JXQCWTO

Product Title: kawaii pastel tops tees pink flower design
Euclidean Distance from input image: 47.38248
Amazon Url: www.amzon.com/dp/B071FCWD97

Product Title: womens thin style tops tees pastel watermelon print
Euclidean Distance from input image: 47.71842
Amazon Url: www.amzon.com/dp/B01JUNHBRM

Product Title: kawaii pastel tops tees baby blue flower design
Euclidean Distance from input image: 47.90206
Amazon Url: www.amzon.com/dp/B071SBCY9W

Product Title: edv cheetah run purple multi xl
Euclidean Distance from input image: 48.046482
Amazon Url: www.amzon.com/dp/B01CUPYBM0

Product Title: danskin womens vneck loose performance tee xsmall pink ombre
Euclidean Distance from input image: 48.101837
Amazon Url: www.amzon.com/dp/B01F7PHXY8

Product Title: summer alpaca 3d pastel casual loose tops tee design
Euclidean Distance from input image: 48.118866
Amazon Url: www.amzon.com/dp/B01I80A93G

Product Title: miss chievous juniors striped peplum tank top medium shadowpeach
Euclidean Distance from input image: 48.13122
Amazon Url: www.amzon.com/dp/B0177DM70S

Product Title: red pink floral heel sleeveless shirt xl xxl
Euclidean Distance from input image: 48.16945
Amazon Url: www.amzon.com/dp/B00JV63QQE

Product Title: moana logo adults hot v neck shirt black xxl
Euclidean Distance from input image: 48.256786
Amazon Url: www.amzon.com/dp/B01LX6H43D

Product Title: abaday multicolor cartoon cat print short sleeve longline shirt large
Euclidean Distance from input image: 48.265686
Amazon Url: www.amzon.com/dp/B01CR57YY0

Product Title: kawaii cotton pastel tops tees peach pink cactus design
Euclidean Distance from input image: 48.362602
Amazon Url: www.amzon.com/dp/B071WYLBZS

Product Title: chicago chicago 18 shirt women pink
Euclidean Distance from input image: 48.383606
Amazon Url: www.amzon.com/dp/B01GXAZTRY

Product Title: yichun womens tiger printed summer tshirts tops
Euclidean Distance from input image: 48.449356
Amazon Url: www.amzon.com/dp/B010NN9RXO

Product Title: nancy lopez whimsy short sleeve whiteblacklemon drop xs
Euclidean Distance from input image: 48.47889
Amazon Url: www.amzon.com/dp/B01MPX6IDX

Product Title: womens tops tees pastel peach ice cream cone print
Euclidean Distance from input image: 48.557957
Amazon Url: www.amzon.com/dp/B0734GRKZL

```
Product Title:  uswomens mary j blige without tshirts shirt
Euclidean Distance from input image: 48.614372
Amazon Url: www.amzon.com/dp/B01M0XXFKK
```

In [65]:

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

# load the original 16K dataset
data = pd.read_pickle('pickels/16k_apperal_data_preprocessed')

#ordering dataponts of data according to asin
df_asin=pd.DataFrame(asins,columns=['asin'])
df1 = data.set_index('asin')
df1 = df1.reindex(index=df_asin['asin'])
data = df1.reset_index()
df_asins = list(data['asin'])
from IPython.display import display, Image, SVG, Math, YouTubeVideo


#get similar products using CNN features (VGG-16)
def get_similar_products_cnn_2(doc_id, num_results):
    doc_id = asins.index(df_asins[doc_id])
    pairwise_dist = pairwise_distances(bottleneck_features_train, bottleneck_features_train[doc_id]
.reshape(1,-1))

    return(pairwise_dist)
```

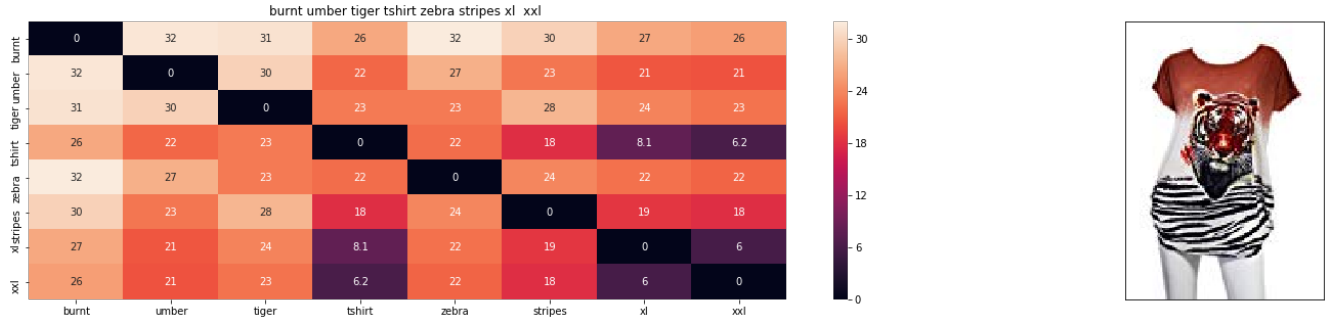
In [69]:

```
def idf_w2v_brand_img(doc_id, w1, w2,w3, num_results):
    # doc_id: apparel's id in given corpus
    # w1: weight for  w2v features
    # w2: weight for brand and color features

    # pairwise_dist will store the distance from given input apparel to all remaining apparels
    # 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
    idf_w2v_dist  = pairwise_distances(w2v_title_weight, w2v_title_weight[doc_id].reshape(1,-1))
    ex_feat_dist = pairwise_distances(extra_features, extra_features[doc_id])
    dist_img = get_similar_products_cnn_2(doc_id, num_results)

    pairwise_dist   = (w1 * idf_w2v_dist +  w2 * ex_feat_dist + w3 *dist_img)/float(w1 + w2+w3)

    # 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)):
        heat_map_w2v_brand(data['title'].loc[df_indices[0]],data['title'].loc[df_indices[i]], data[
'medium_image_url'].loc[df_indices[i]], indices[0], indices[i],df_indices[0], df_indices[i], 'weigh
ted')
        print('ASIN :',data['asin'].loc[df_indices[i]])
```

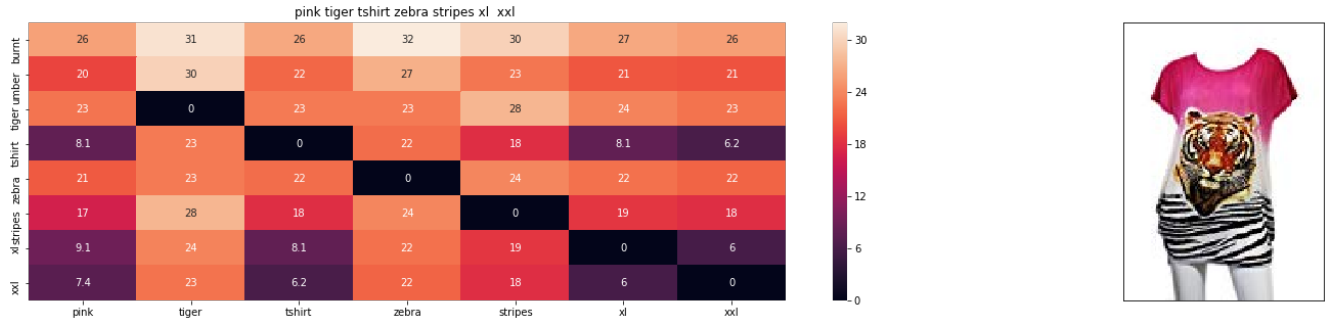
```
        print('Brand :',data['brand'].loc[df_indices[i]])
        print('euclidean distance from input :', pdists[i])
        print('='*125)
```

In [75]:

```
idf_w2v_brand_img(1416,5,5,5, 20)#providing same weight for text,color and imgage
```

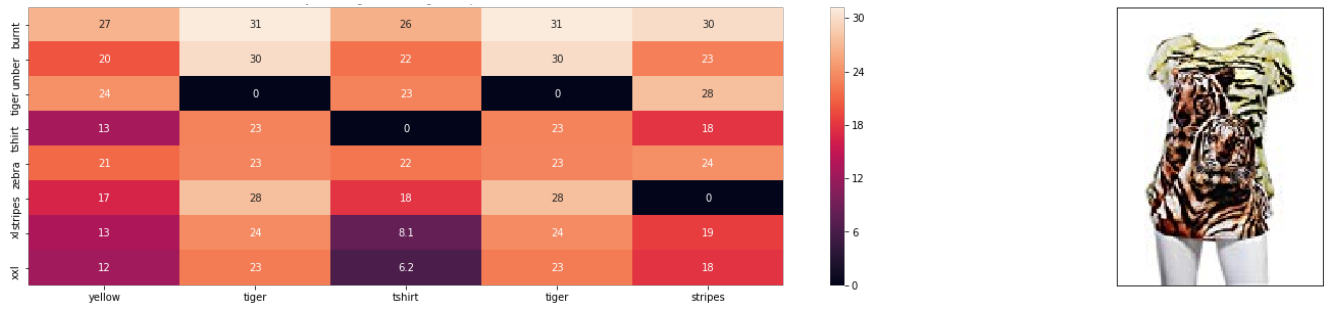


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.108653279719874e-06
==========



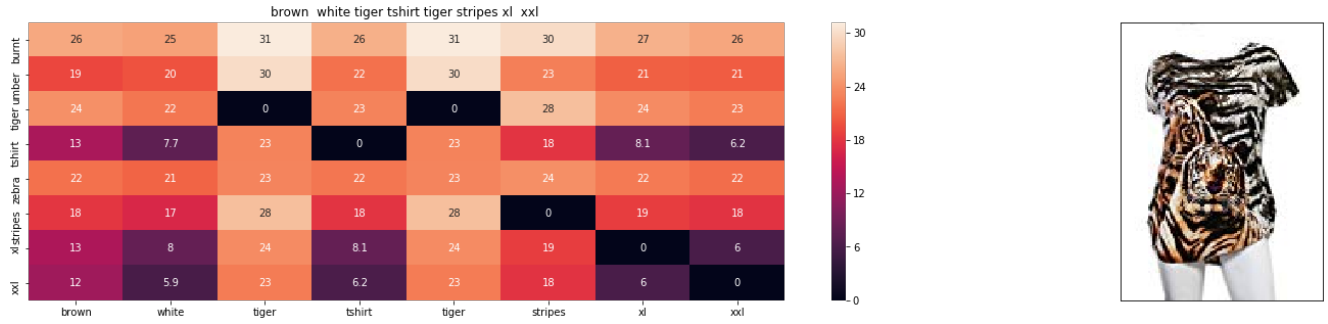
ASIN : B00JXQASS6
Brand : Si Row
euclidean distance from input : 11.842757034422217
==========

yellow tiger tshirt tiger stripes l



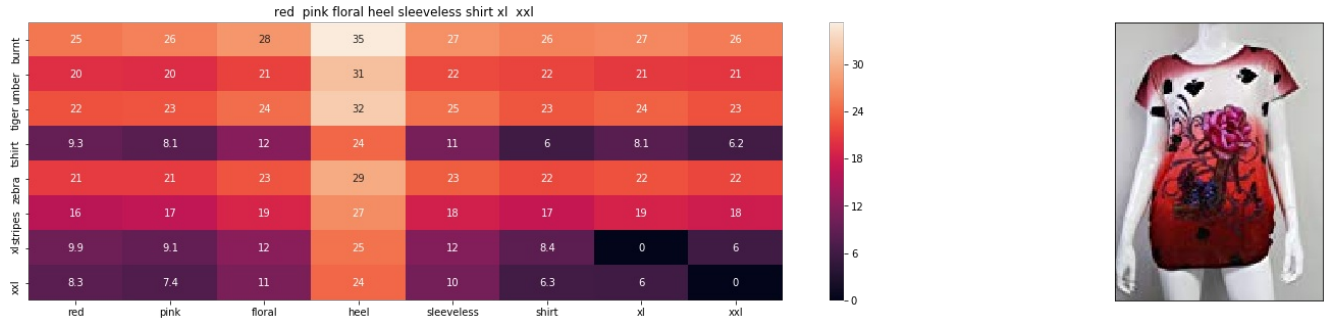
ASIN : B00JXQCUIC
Brand : Si Row
euclidean distance from input : 16.189590072752296

==========



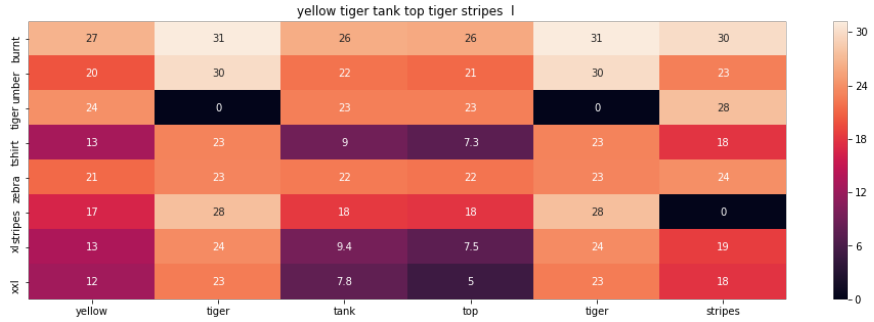
ASIN : B00JXQCWTO
Brand : Si Row
euclidean distance from input : 16.257032267252605

==========

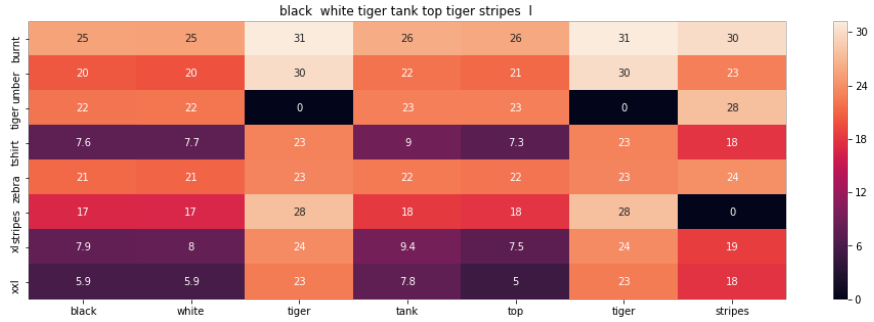


ASIN : B00JV63QQE
Brand : Si Row
euclidean distance from input : 19.076066080849625

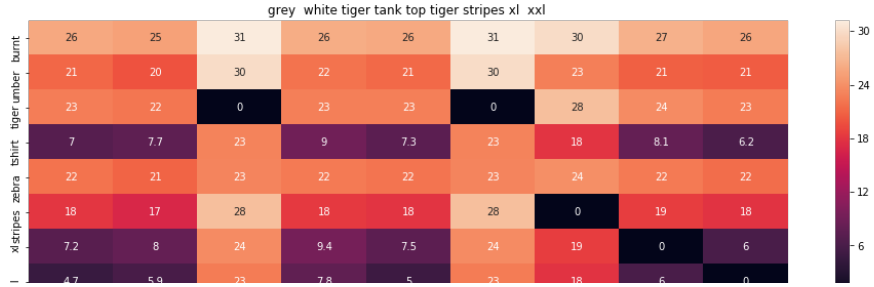
==========



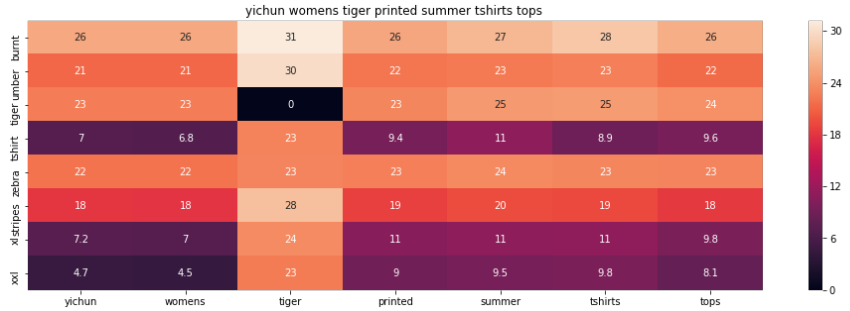
ASIN : B00JXQAUWA
Brand : Si Row
euclidean distance from input : 19.097982025266944



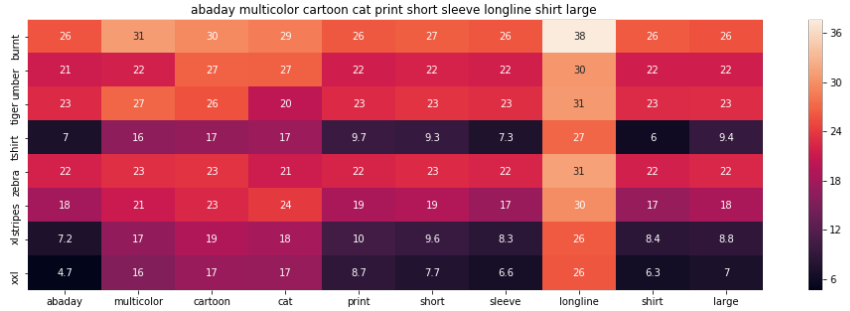
ASIN : B00JXQAO94
Brand : Si Row
euclidean distance from input : 19.13750559500978



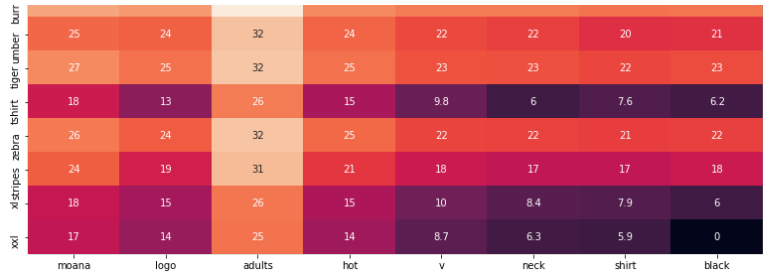
ASIN : B00JXQAFZ2
Brand : Si Row
euclidean distance from input : 19.28427416495607
==========



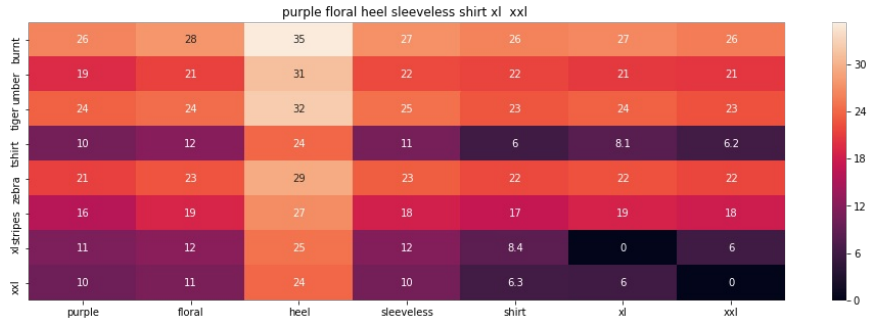
ASIN : B010NN9RXO
Brand : YICHUN
euclidean distance from input : 19.32008696452804
==========



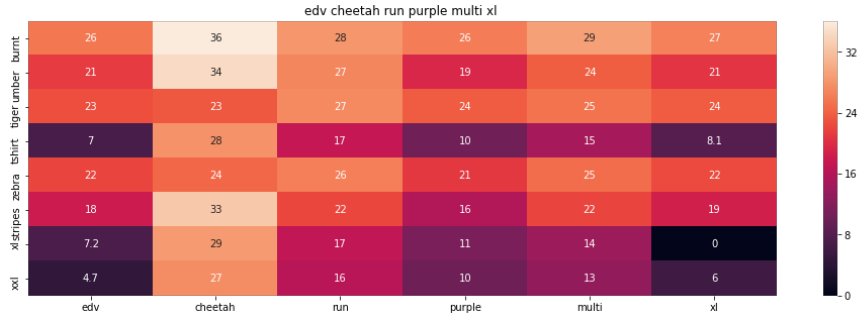
ASIN : B01CR57YY0
Brand : ABADAY
euclidean distance from input : 19.48145171062179
==========



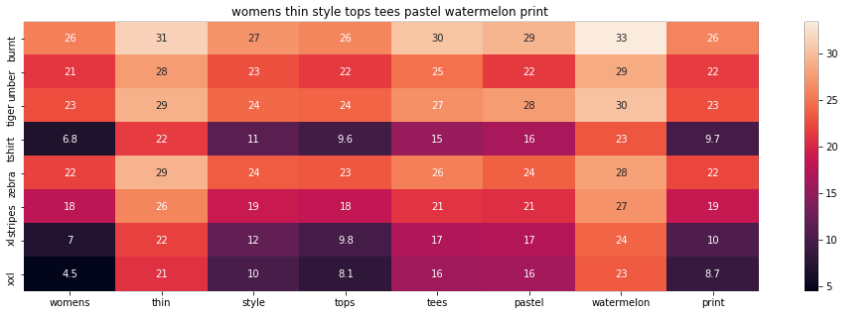
ASIN : B01LX6H43D
Brand : BOBOB
euclidean distance from input : 19.48374459799158
==========



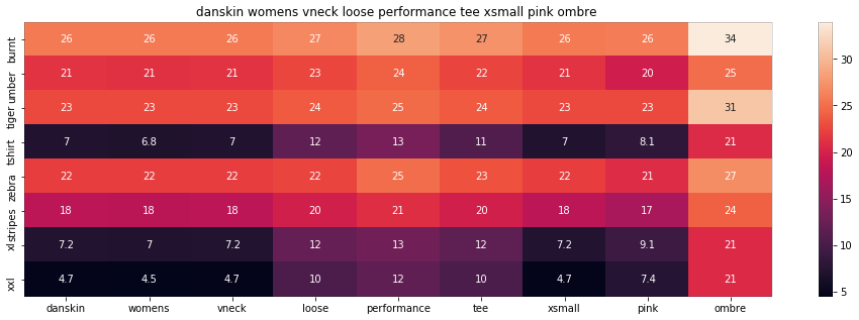
ASIN : B00JV63VC8
Brand : Si Row
euclidean distance from input : 19.487446085732437
==========



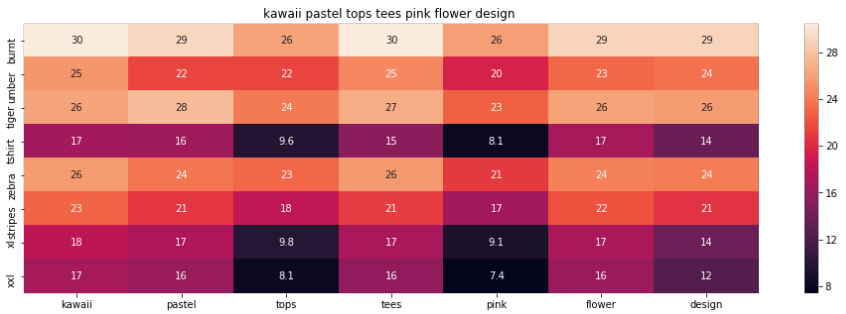
ASIN : B01CUPYBM0
Brand : Styleco
euclidean distance from input : 19.525803669850305
==========



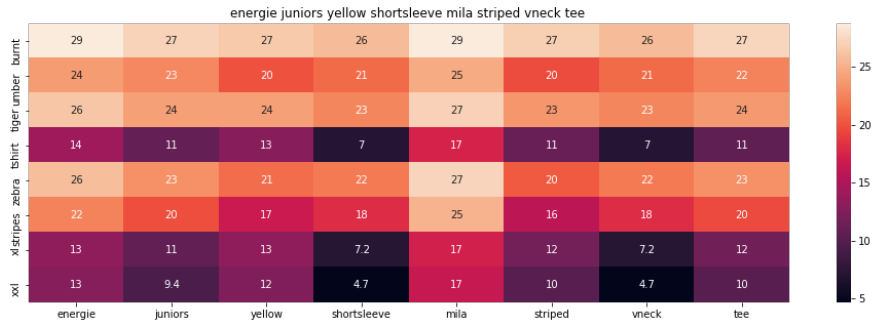
ASIN : B01JUNHBRM
Brand : Namnoi Clothing Store
euclidean distance from input : 19.641300093265762
==========



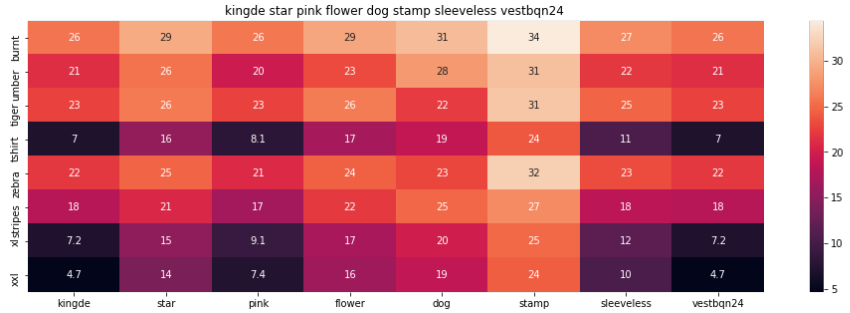
ASIN : B01F7PHXY8
Brand : Danskin Now
euclidean distance from input : 19.690636444091798
==========



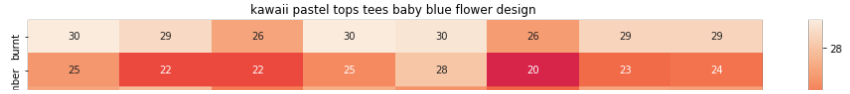
ASIN : B071FCWD97
Brand : Namnoi Clothing Store
euclidean distance from input : 19.74858769092201

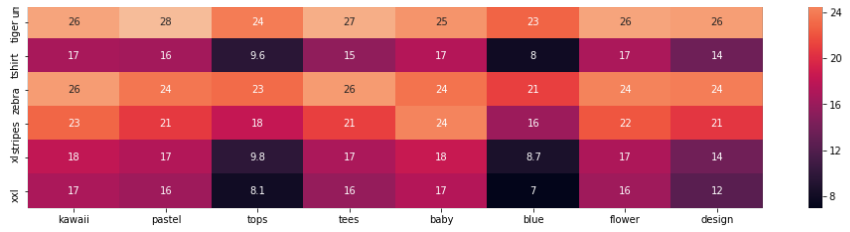


ASIN : B01MAV3S42
Brand : Energie
euclidean distance from input : 19.813749989430384

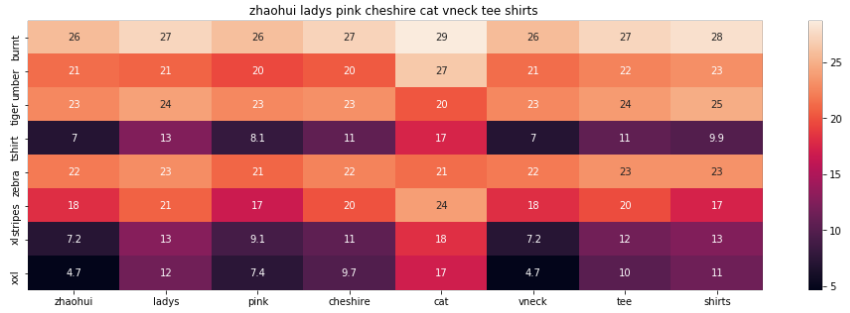


ASIN : B015H3W9BM
Brand : KINGDE
euclidean distance from input : 19.81972151652999





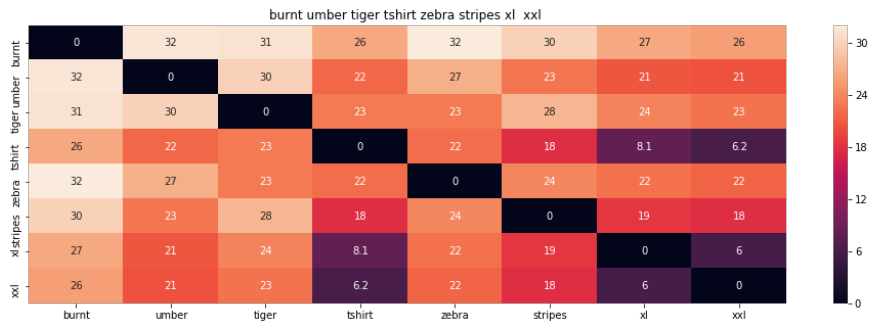
ASIN : B071SBCY9W
Brand : Namnoi Clothing Store
euclidean distance from input : 19.85036254240313
==========



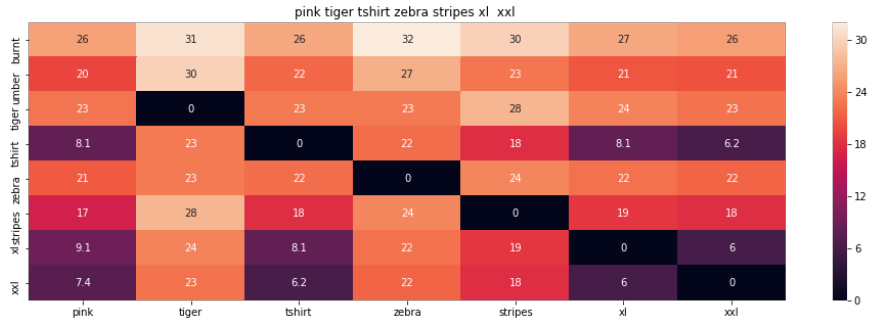
ASIN : B018DDFT4M
Brand : ZhaoHui
euclidean distance from input : 19.85594181275395
==========

In [76]:

```
idf_w2v_brand_img(1416,5,5,2, 20)#providing same weight for text and color
```

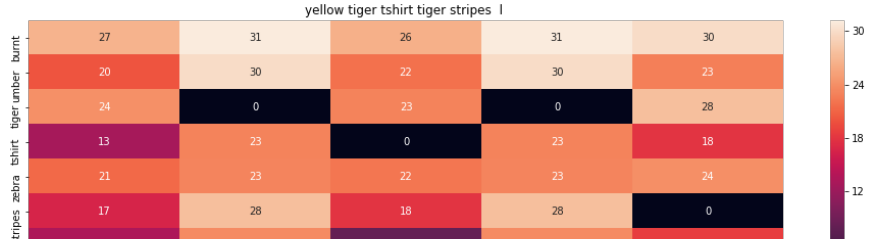


ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 1.0543266550181822e-06
==========

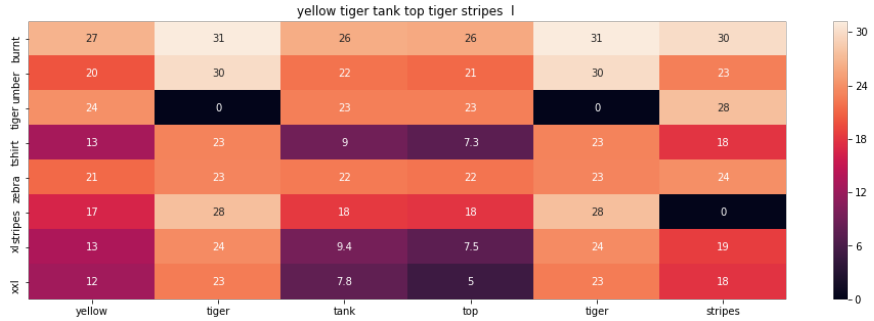


ASIN : B00JXQASS6
Brand : Si Row
euclidean distance from input : 7.290903568418396

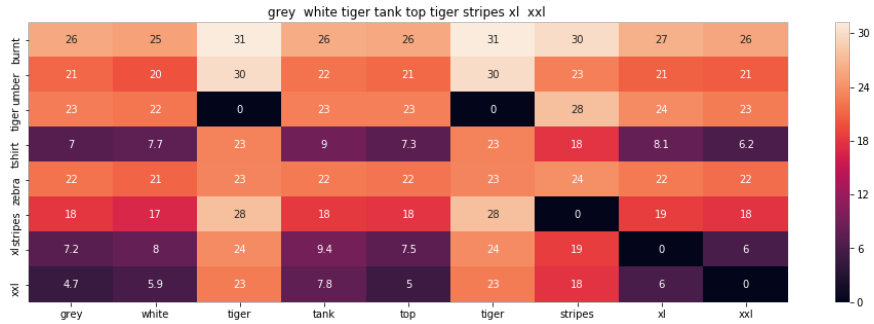
ASIN : B00JXQCWTO
Brand : Si Row
euclidean distance from input : 9.32125155131022



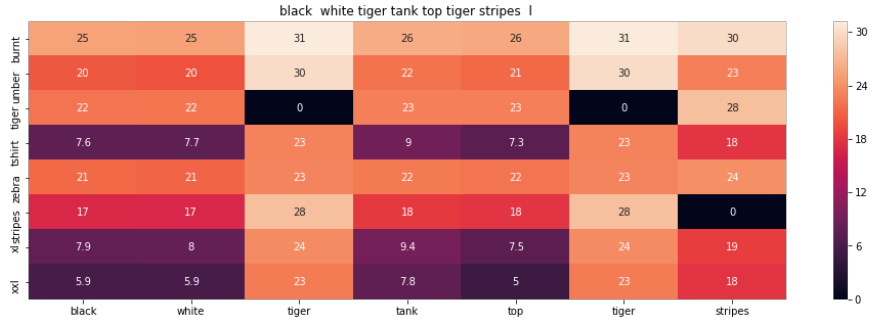
ASIN : B00JXQCUIC
Brand : Si Row
euclidean distance from input : 9.921709219765239



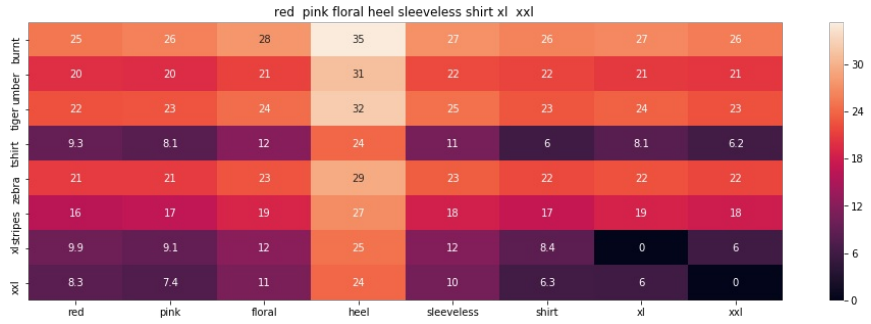
ASIN : B00JXQAUWA
Brand : Si Row
euclidean distance from input : 11.324924787036153



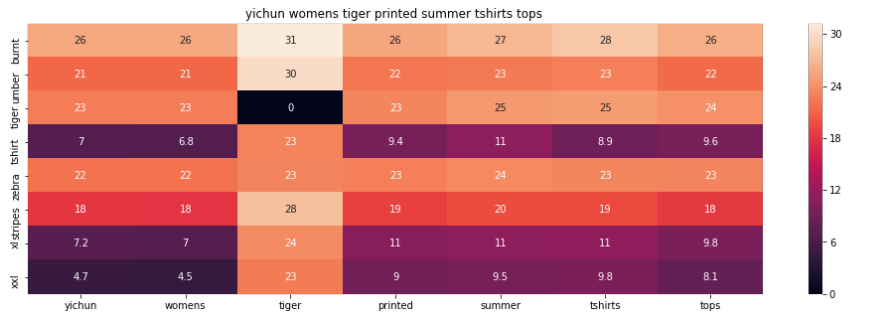
ASIN : B00JXQAFZ2
Brand : Si Row
euclidean distance from input : 11.335730711769633



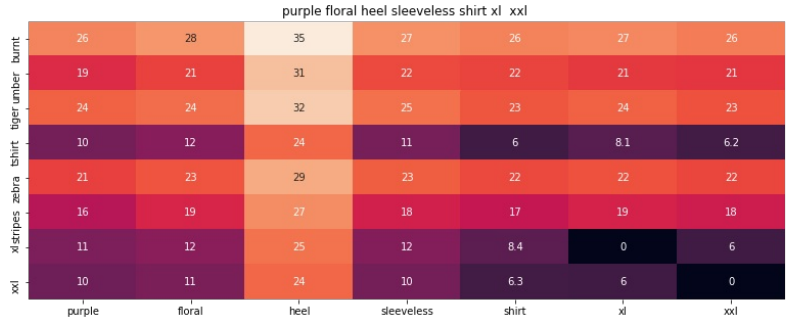
ASIN : B00JXQAO94
Brand : Si Row
euclidean distance from input : 11.345561345568868
==========



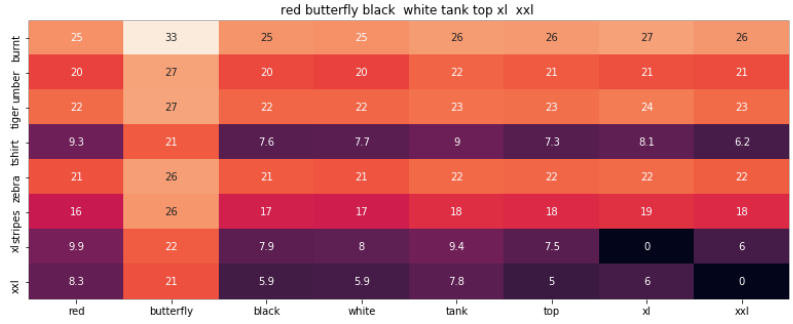
ASIN : B00JV63QQE
Brand : Si Row
euclidean distance from input : 11.802720387927266
==========



ASIN : B010NN9RXO
Brand : YICHUN
euclidean distance from input : 12.03776968588466
==========

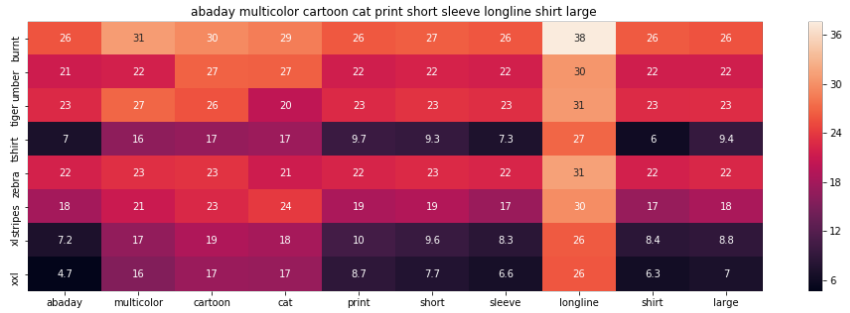


ASIN : B00JV63VC8
Brand : Si Row
euclidean distance from input : 12.09525299087323

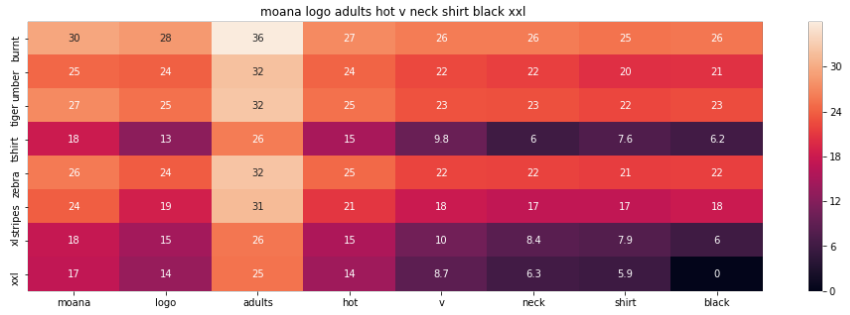


ASIN : B00JV63CW2
Brand : Si Row
euclidean distance from input : 12.099754651538106

ASIN : B00JXQCFRS
Brand : Si Row
euclidean distance from input : 12.26988506332196

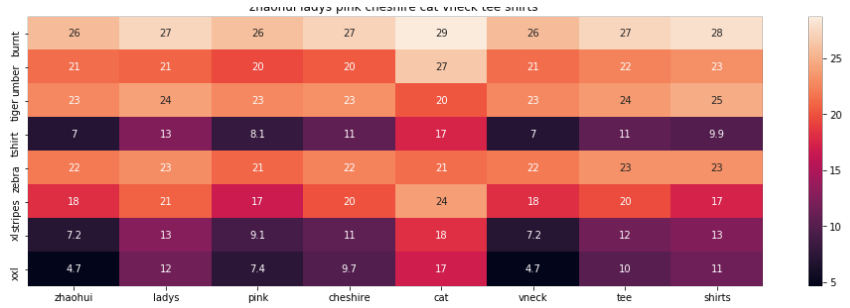


ASIN : B01CR57YY0
Brand : ABADAY
euclidean distance from input : 12.285393129488176

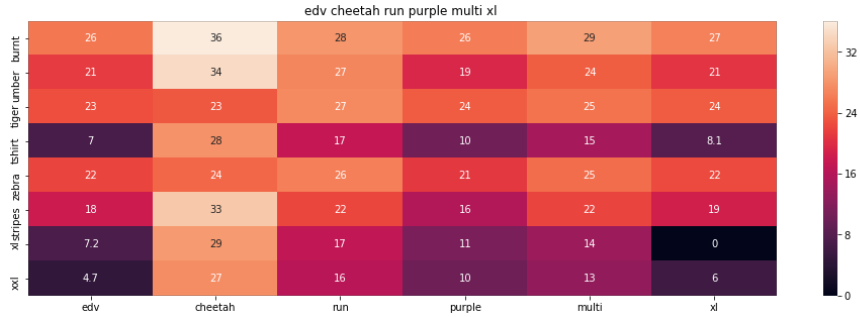


ASIN : B01LX6H43D
Brand : BOBOB
euclidean distance from input : 12.290483842989152

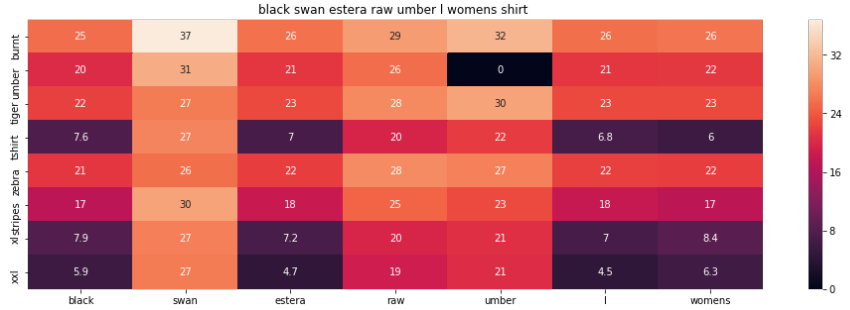
zhaohui ladys pink cheshire cat vneck tee shirts



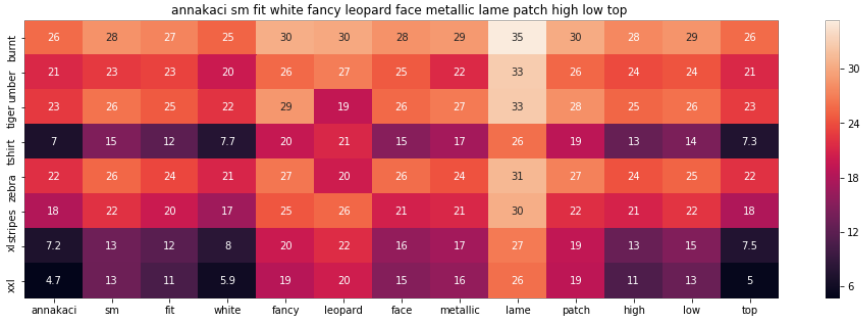
ASIN : B018DDFT4M
Brand : ZhaoHui
euclidean distance from input : 12.331318587760473



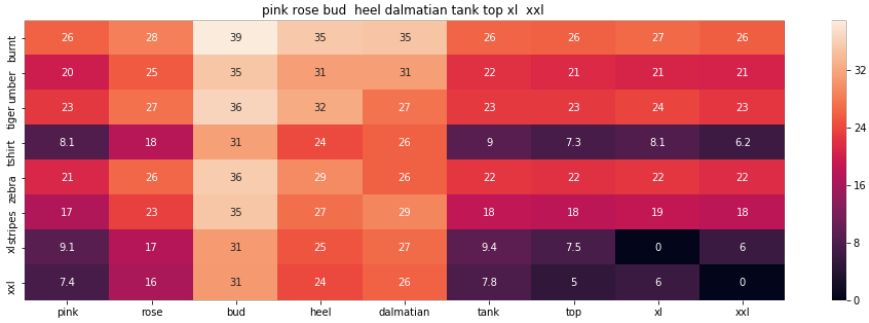
ASIN : B01CUPYBM0
Brand : Styleco
euclidean distance from input : 12.39563438365891



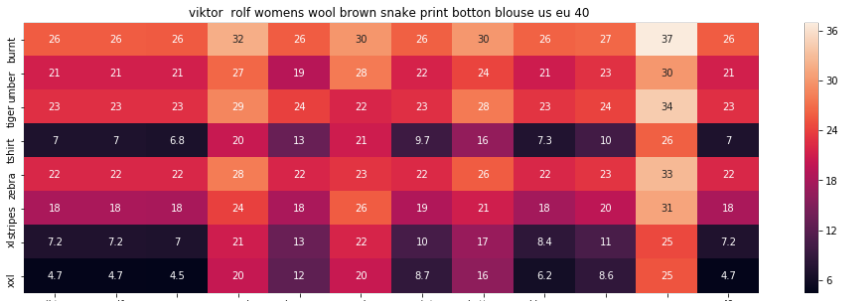
ASIN : B06Y1VN8WQ
Brand : Black Swan
euclidean distance from input : 12.442610422770182



ASIN : B00F4IJRES
Brand : Anna-Kaci
euclidean distance from input : 12.447298686010782



ASIN : B00JXQAX2C
Brand : Si Row
euclidean distance from input : 12.463132222644063

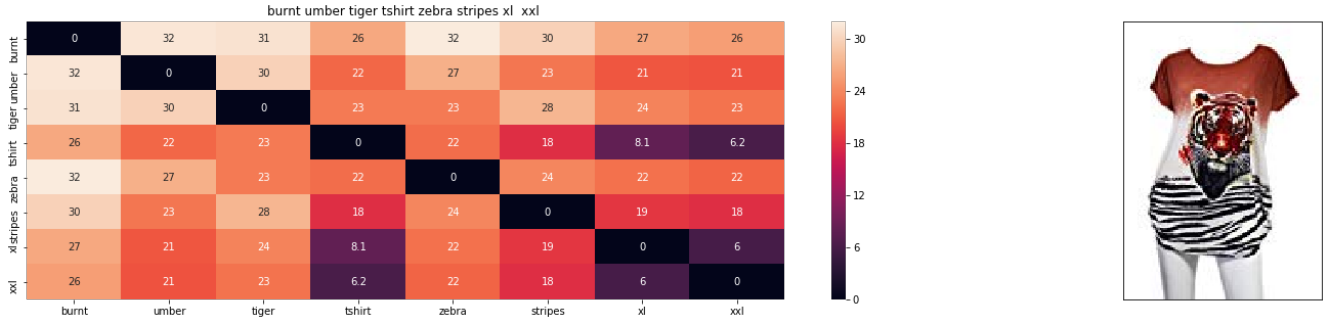


viktor rolf womens wool brown snake print botton blouse us eu 40

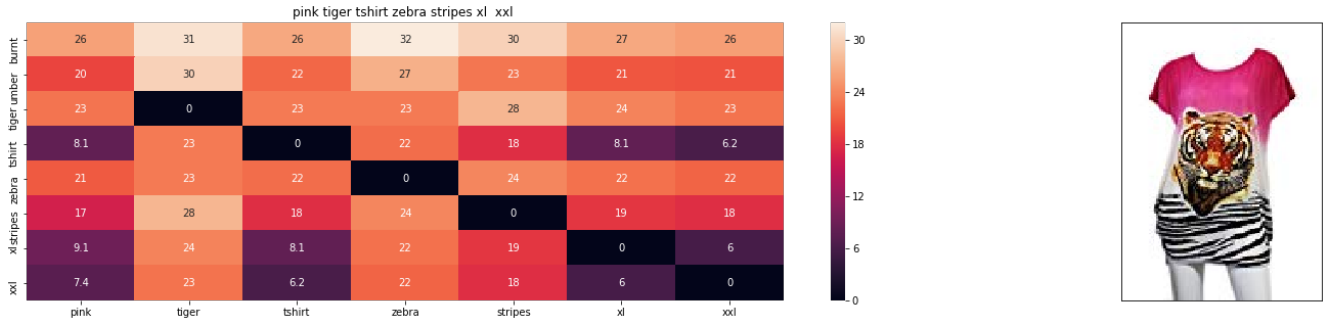
ASIN : B00LEHNVZ4
Brand : Viktor & Rolf
euclidean distance from input : 12.466990850976552
==========

In [78]:

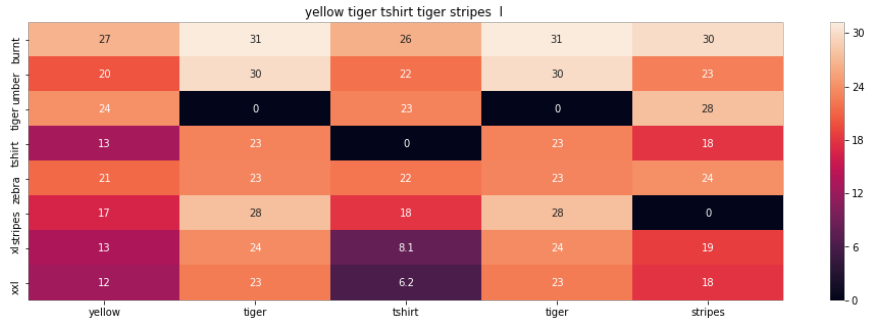
```
idf_w2v_brand_img(1416,4,1,4, 20)#providing same weight for text and image
```



ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 2.8115377467151525e-06
==========

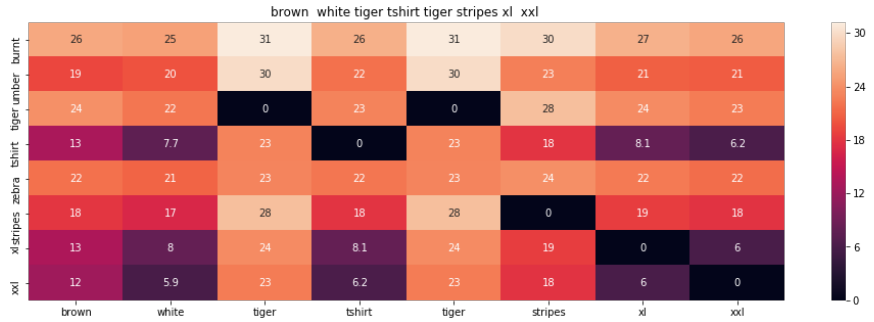


ASIN : B00JXQASS6
Brand : Si Row
euclidean distance from input : 15.318938191771926
==========



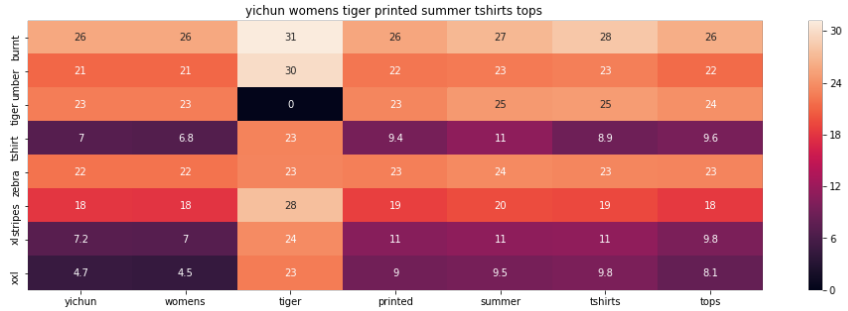
ASIN : B00JXQCUIC
Brand : Si Row
euclidean distance from input : 21.114716254380433

==========



ASIN : B00JXQCWTO
Brand : Si Row
euclidean distance from input : 21.676043404473198

==========

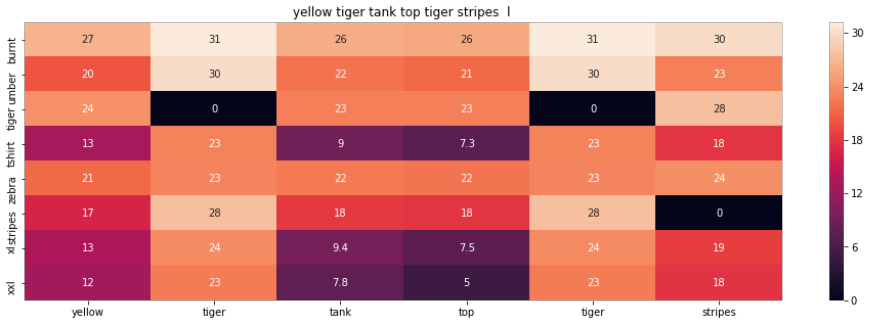


ASIN : B010NN9RXO
Brand : YICHUN
euclidean distance from input : 24.878198764244807

==========

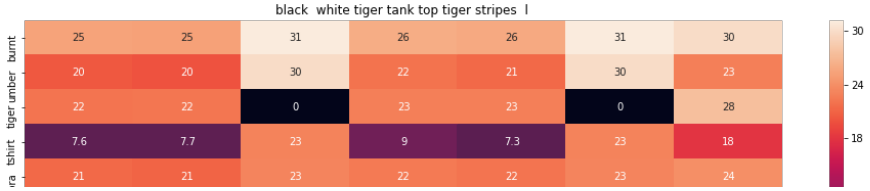
ASIN : B00JV63QQE
Brand : Si Row
euclidean distance from input : 24.963350338446187

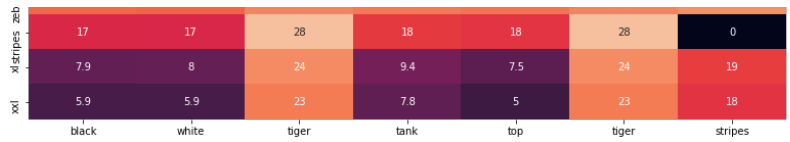
==========



ASIN : B00JXQAUWA
Brand : Si Row
euclidean distance from input : 24.992570919500874

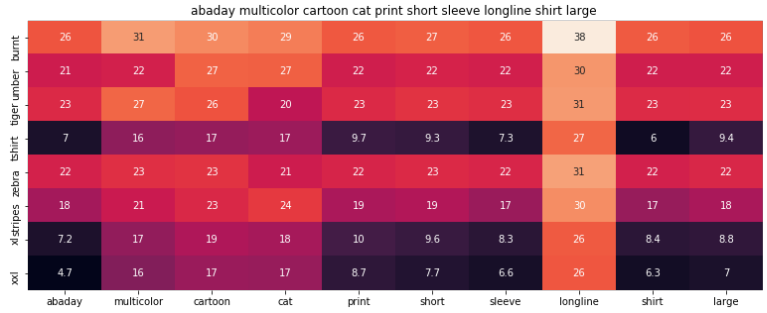
==========





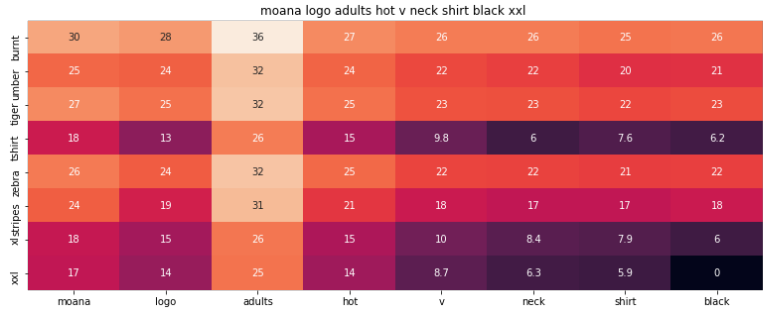
ASIN : B00JXQAO94
Brand : Si Row
euclidean distance from input : 25.04526969065973

==========



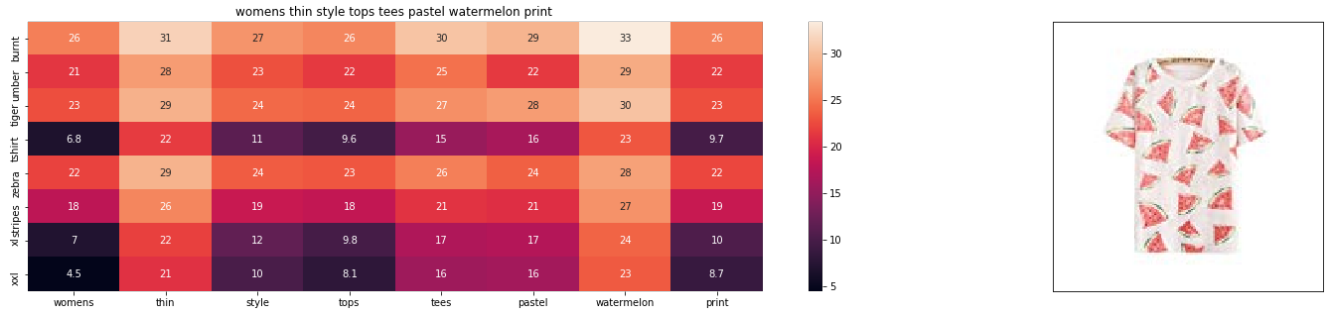
ASIN : B01CR57YY0
Brand : ABADAY
euclidean distance from input : 25.09335171665095

==========

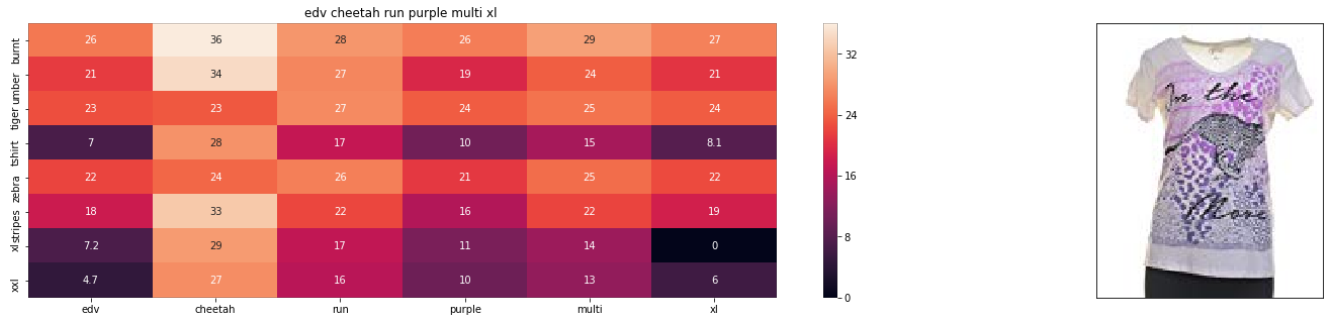


ASIN : B01LX6H43D
Brand : BOBOB
euclidean distance from input : 25.096408772654097

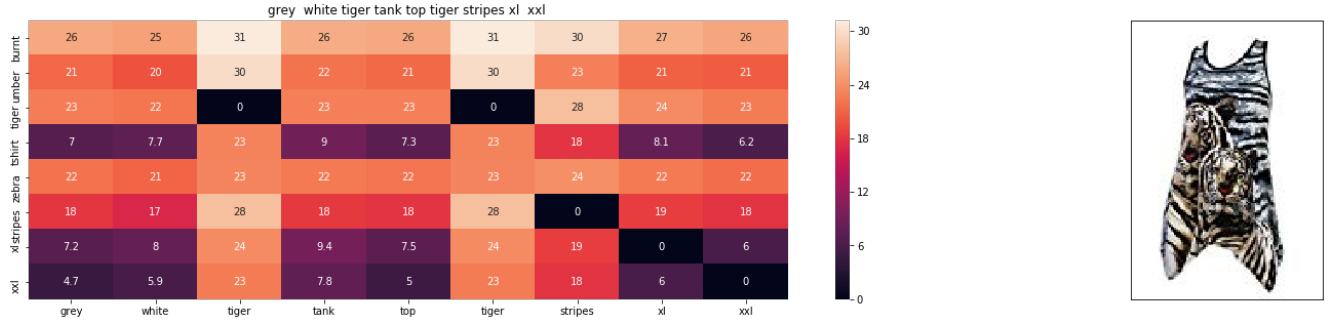
==========



ASIN : B01JUNHBRM
Brand : Namnoi Clothing Store
euclidean distance from input : 25.13430791004909

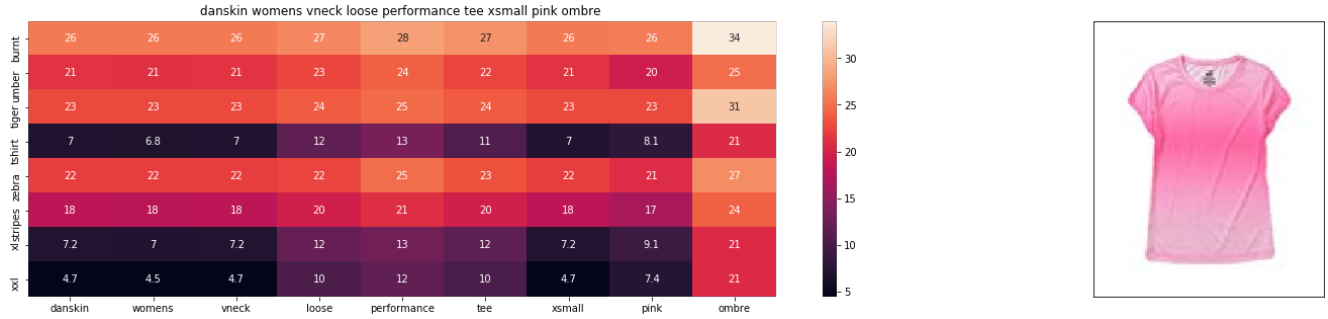


ASIN : B01CUPYBM0
Brand : Styleco
euclidean distance from input : 25.15248800137317



ASIN : B00JXQAFZ2
Brand : Si Row
euclidean distance from input : 25.240961329182408

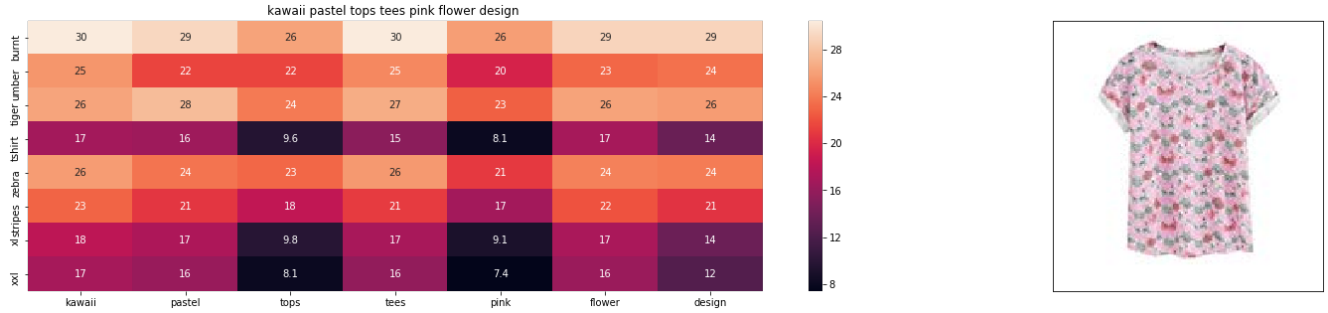
==========



ASIN : B01F7PHXY8
Brand : Danskin Now
euclidean distance from input : 25.254181967841255

==========

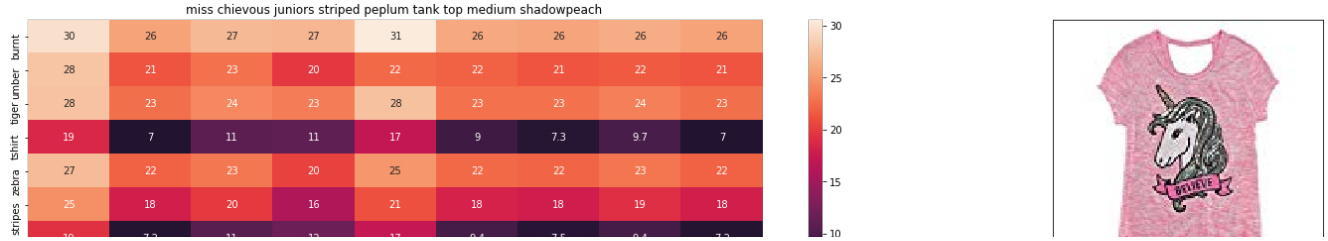
==========

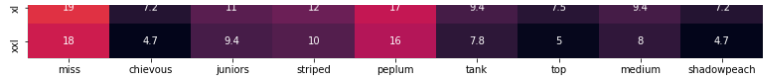


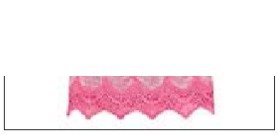
ASIN : B071FCWD97
Brand : Namnoi Clothing Store
euclidean distance from input : 25.277358209799523

==========

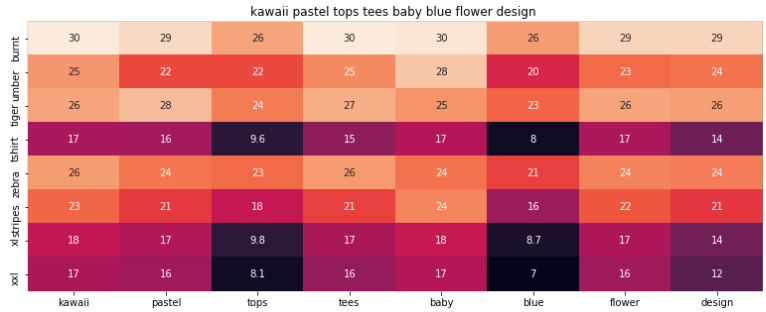
==========



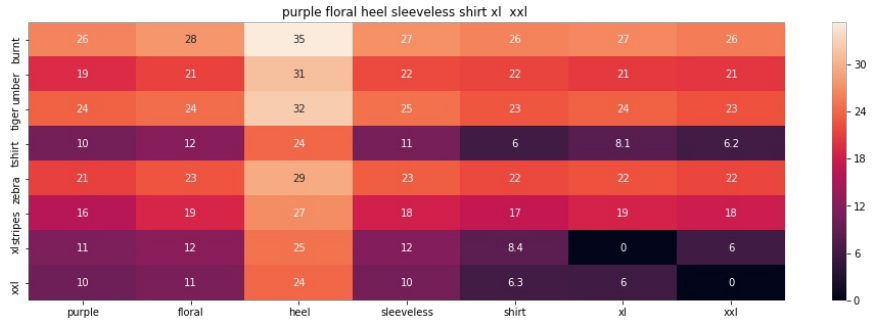




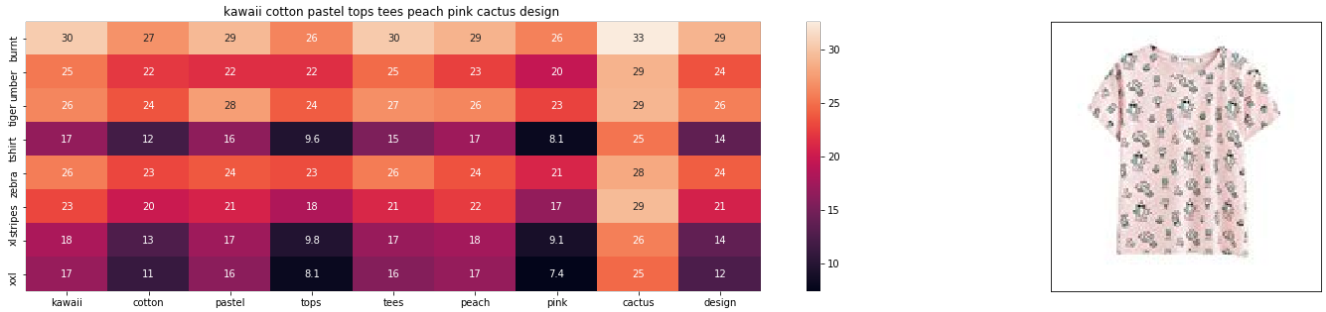
ASIN : B0177DM70S
Brand : MISS CHIEVOUS/AROUND THE WORLD APP
euclidean distance from input : 25.390335275352765
==========



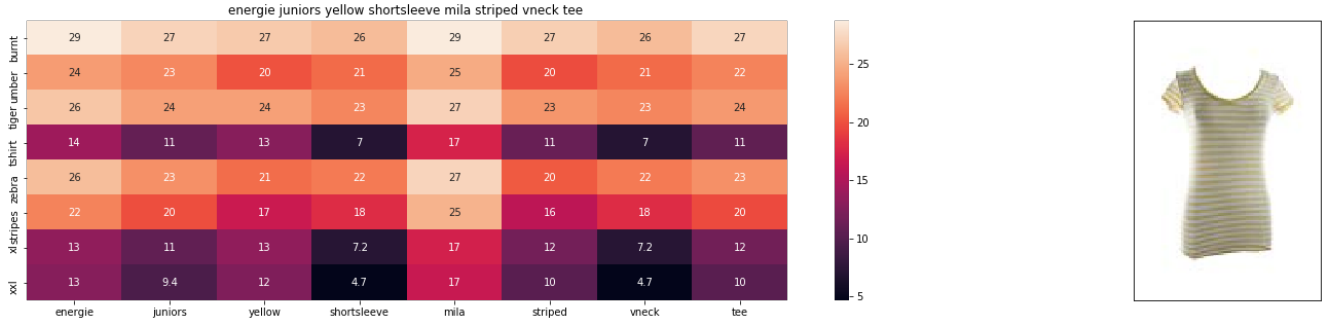
ASIN : B071SBCY9W
Brand : Namnoi Clothing Store
euclidean distance from input : 25.413056740208596
==========



ASIN : B00JV63VC8
Brand : Si Row
euclidean distance from input : 25.511857605020623
==========



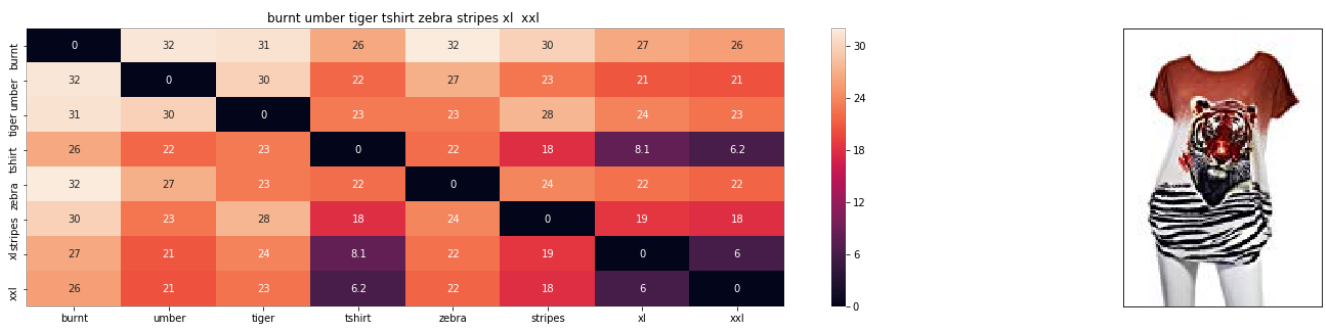
ASIN : B071WYLBZS
Brand : Namnoi Clothing Store
euclidean distance from input : 25.520664263808854
==========



ASIN : B01MAV3S42
Brand : Energie
euclidean distance from input : 25.536416936106246
==========

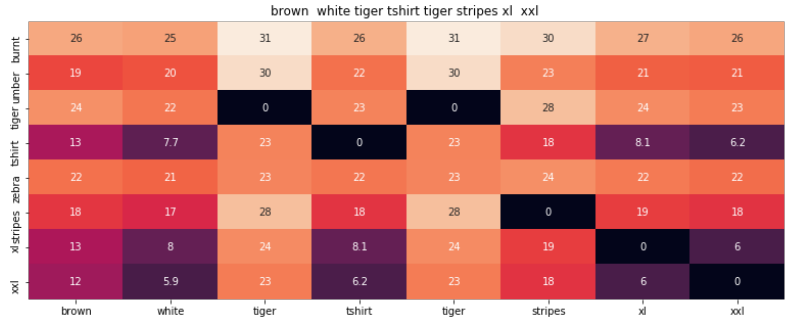
In [79]:

```
idf_w2v_brand_img(1416,5,4,0,20)#providing less weight for imgage
```

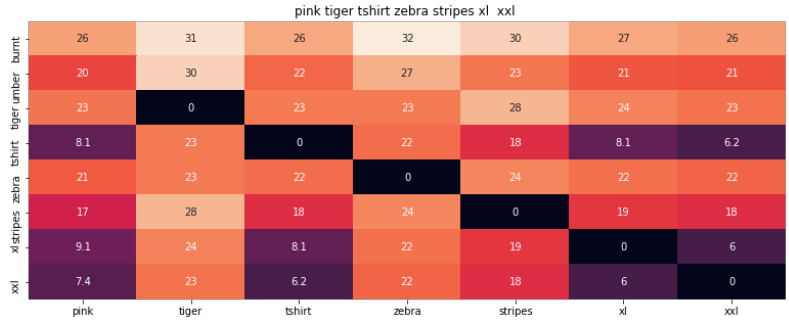


ASIN : B00JXQB5FQ

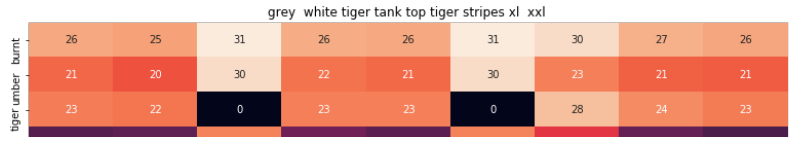
Brand : Si Row
euclidean distance from input : 0.0
==========

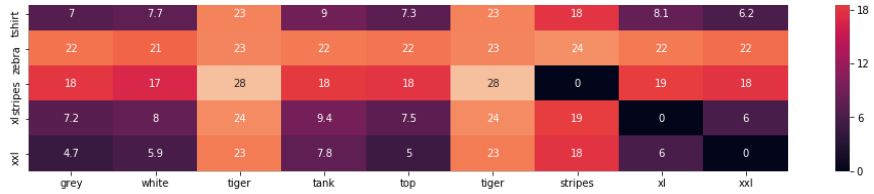


ASIN : B00JXQCWTO
Brand : Si Row
euclidean distance from input : 2.6505228678385415
==========

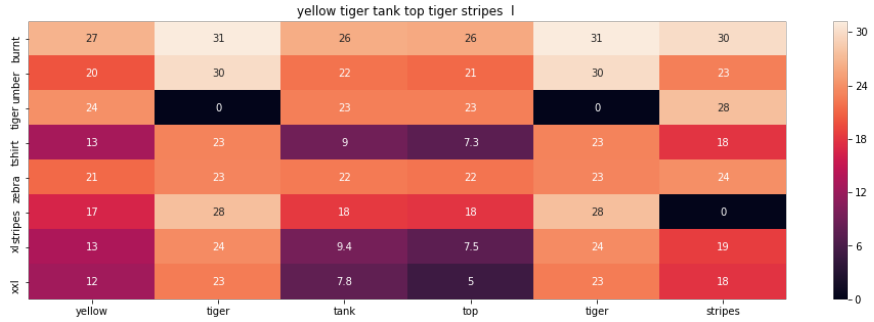


ASIN : B00JXQASS6
Brand : Si Row
euclidean distance from input : 2.8862541624191835
==========

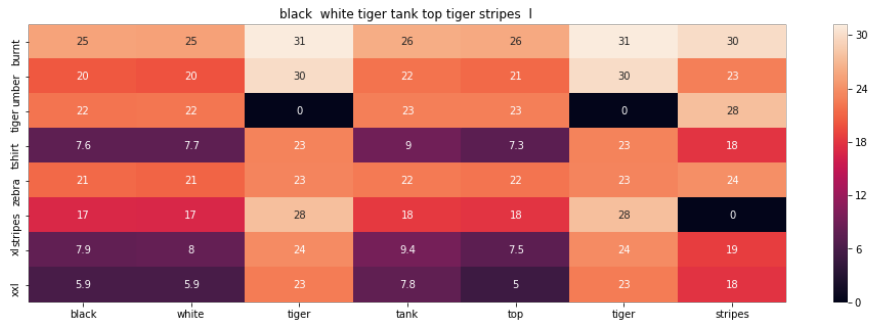




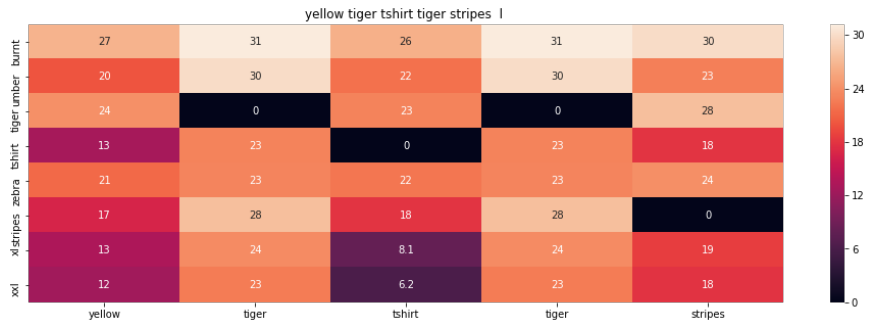
ASIN : B00JXQAFZ2
Brand : Si Row
euclidean distance from input : 3.6064062755919273



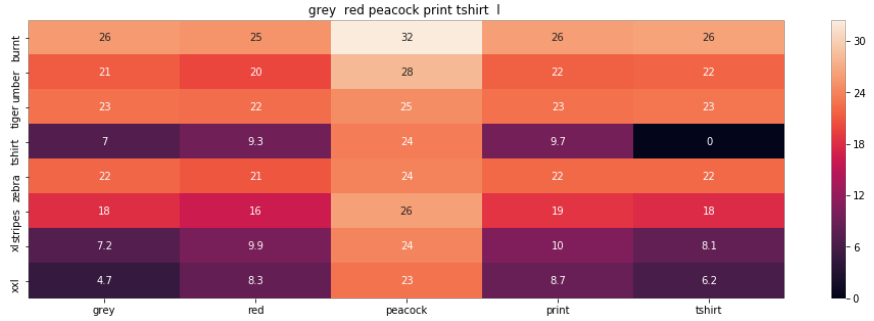
ASIN : B00JXQAUWA
Brand : Si Row
euclidean distance from input : 3.7893852235492838



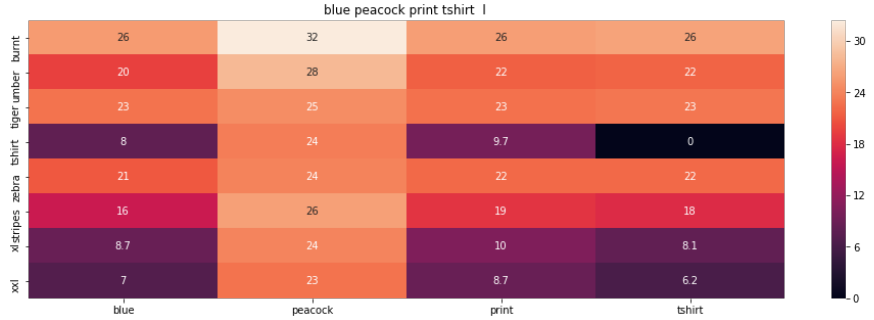
ASIN : B00JXQAO94
Brand : Si Row
euclidean distance from input : 3.791328599878494



ASIN : B00JXQCUIC
Brand : Si Row
euclidean distance from input : 3.9026738910161893

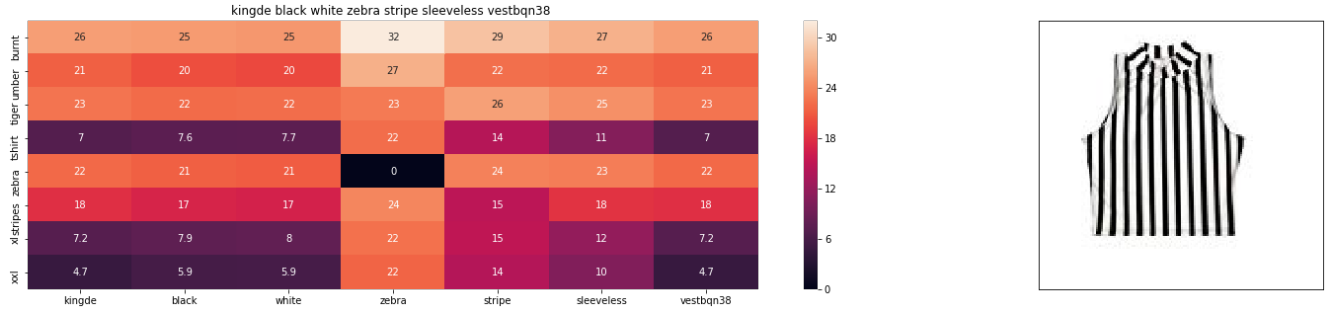


ASIN : B00JXQCFRS
Brand : Si Row
euclidean distance from input : 4.430433231090516

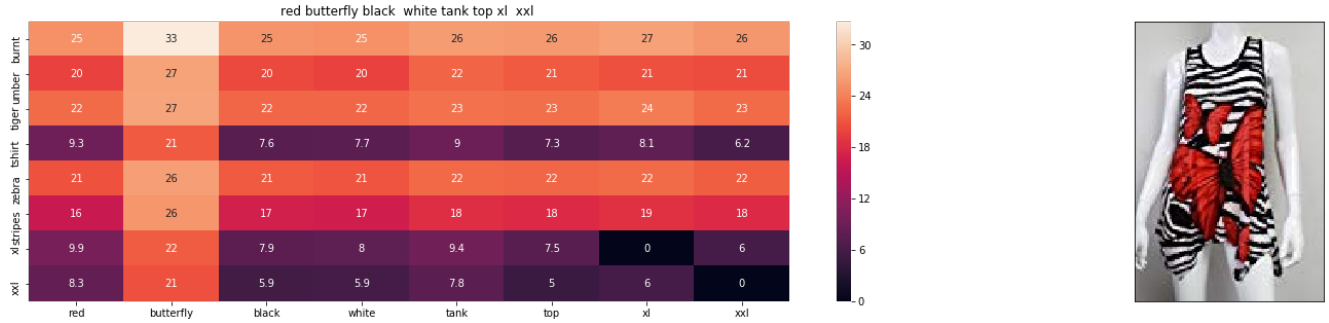


ASIN : B00JXQC8L6

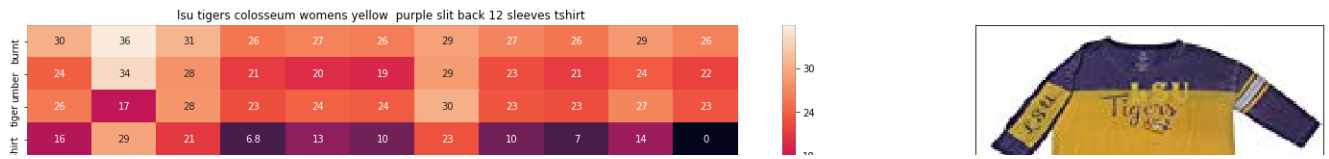
Brand : Si Row
euclidean distance from input : 4.513865322697504
==========

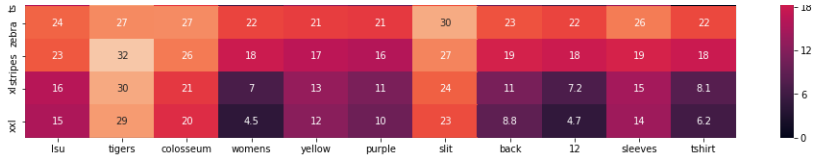


ASIN : B015H41F6G
Brand : KINGDE
euclidean distance from input : 4.58310586155822
==========

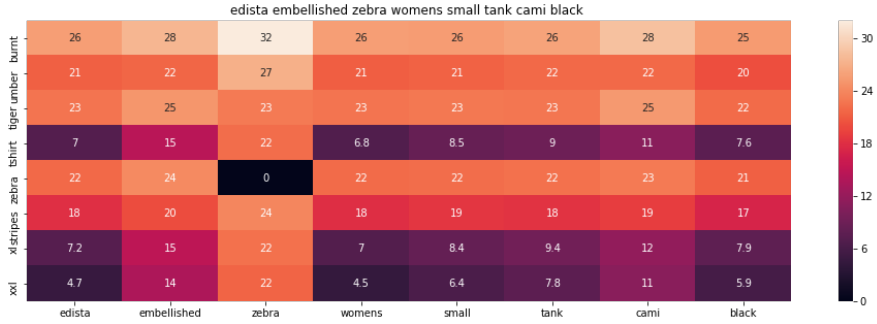


ASIN : B00JV63CW2
Brand : Si Row
euclidean distance from input : 4.605738915498069
==========

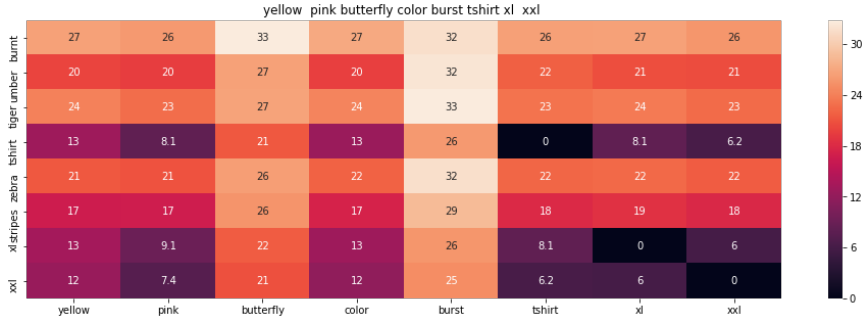




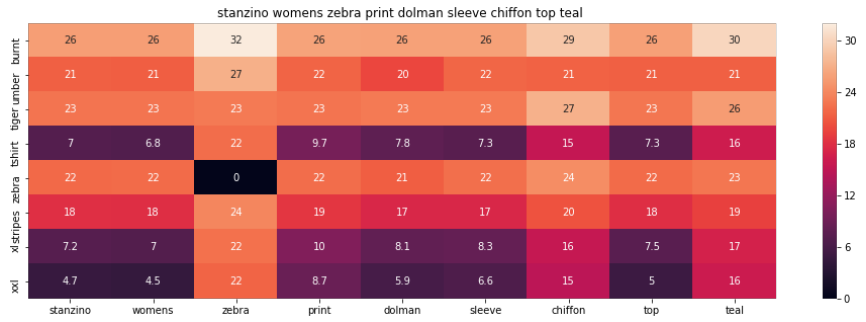
ASIN : B073R5Q8HD
Brand : Colosseum
euclidean distance from input : 4.651836957613993



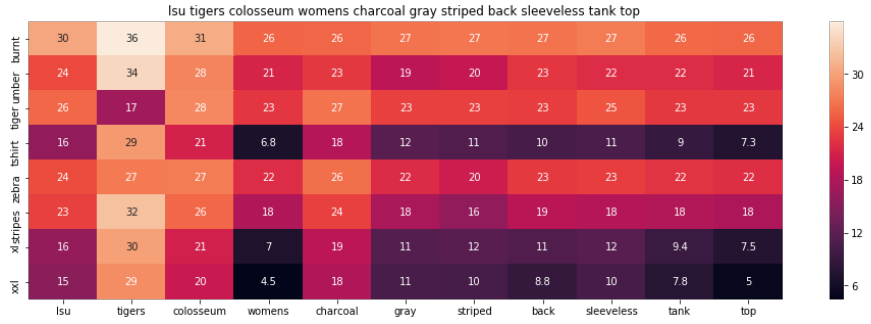
ASIN : B074P8MD22
Brand : Edista
euclidean distance from input : 4.727113650322327



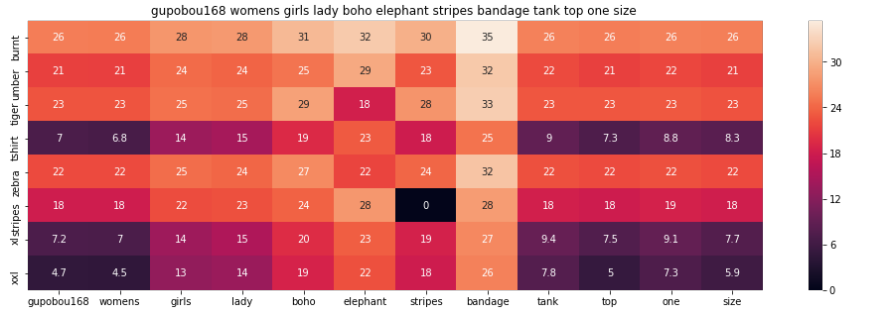
ASIN : B00JXQBBMI
Brand : Si Row
euclidean distance from input : 4.729431745901715



ASIN : B00C0I3U3E
Brand : Stanzino
euclidean distance from input : 4.739722920206224

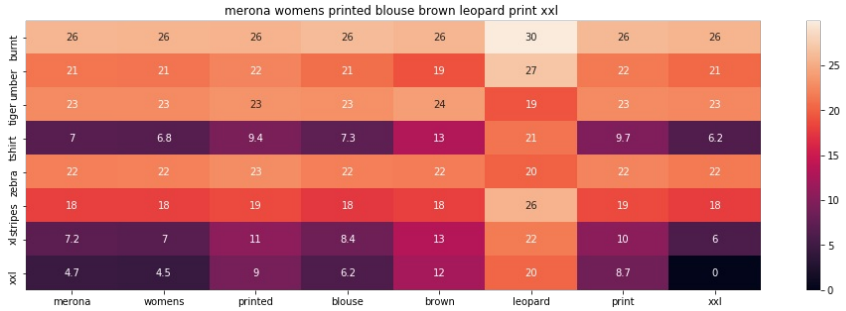


ASIN : B073R4ZM7Y
Brand : Colosseum
euclidean distance from input : 4.759755590969028



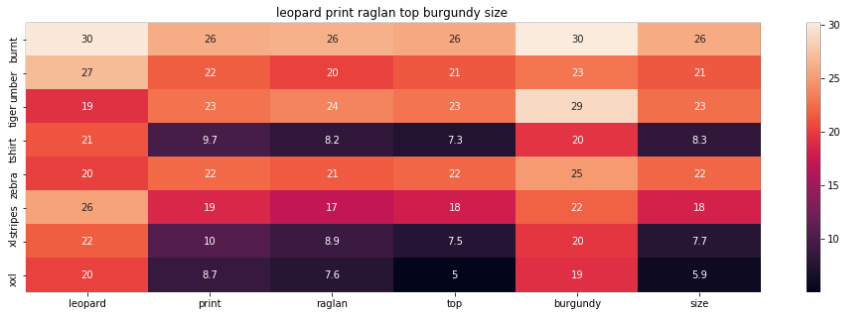
ASIN : B01ER184O6
Brand : GuPoBoU168
euclidean distance from input : 4.803566494231678

==========



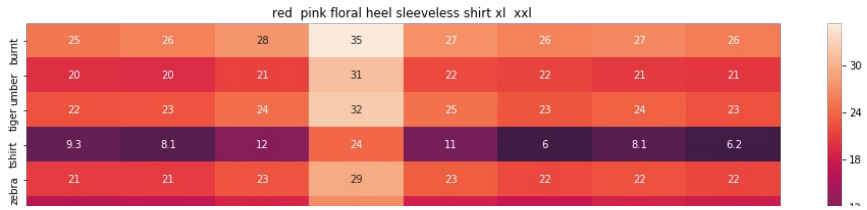
ASIN : B071YF3WDD
Brand : Merona
euclidean distance from input : 4.828946146117683

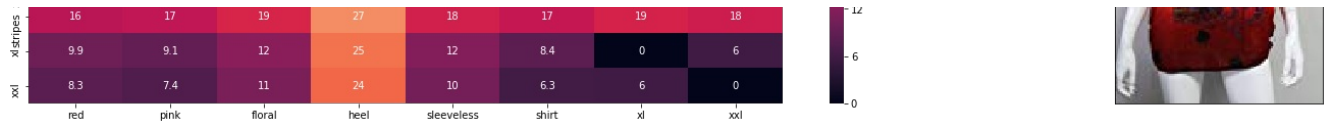
==========



ASIN : B01C6ORLDQ
Brand : 1 Mad Fit
euclidean distance from input : 4.8478613114920135

==========

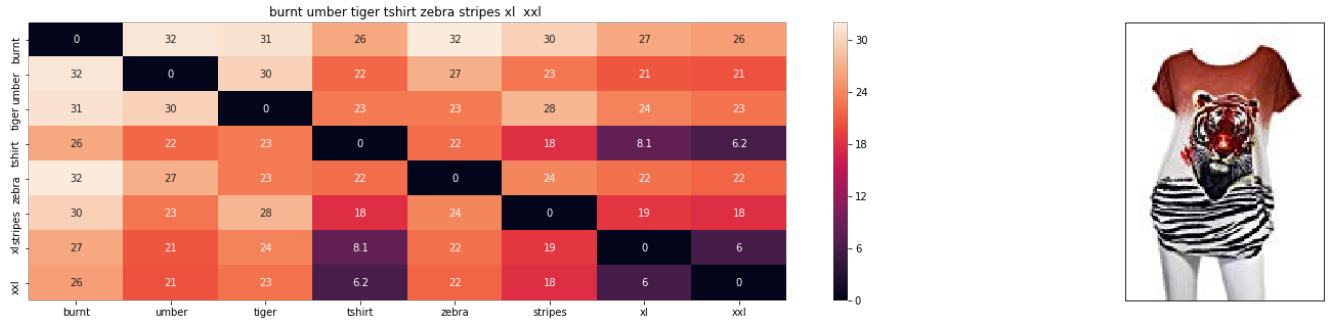




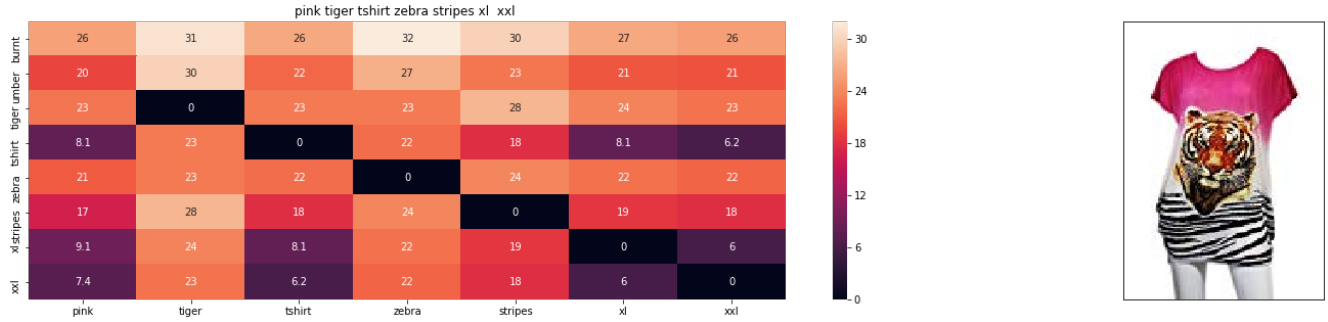
ASIN : B00JV63QQE
Brand : Si Row
euclidean distance from input : 4.875503709741775
==========

In [80]:

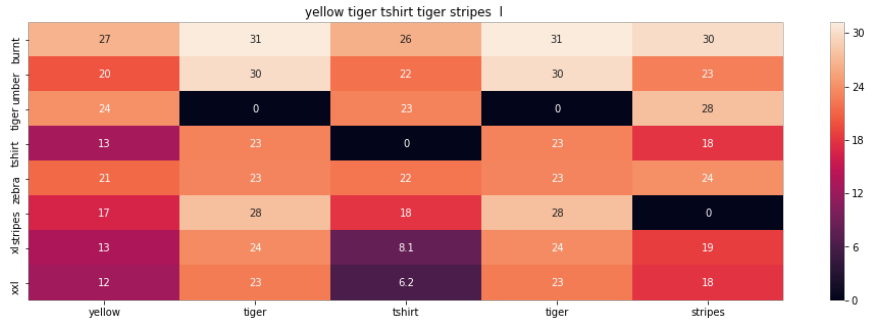
```
idf_w2v_brand_img(1416,0,0,10,20)#providing more weight for imgage
```



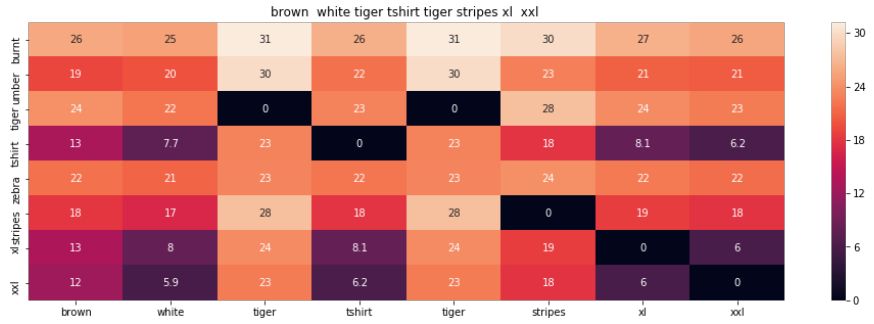
ASIN : B00JXQB5FQ
Brand : Si Row
euclidean distance from input : 6.325959839159623e-06
==========



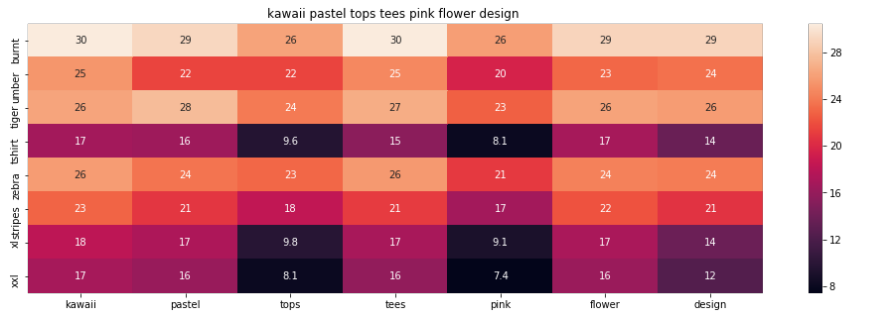
ASIN : B00JXQASS6
Brand : Si Row
euclidean distance from input : 30.0501708984375
==========



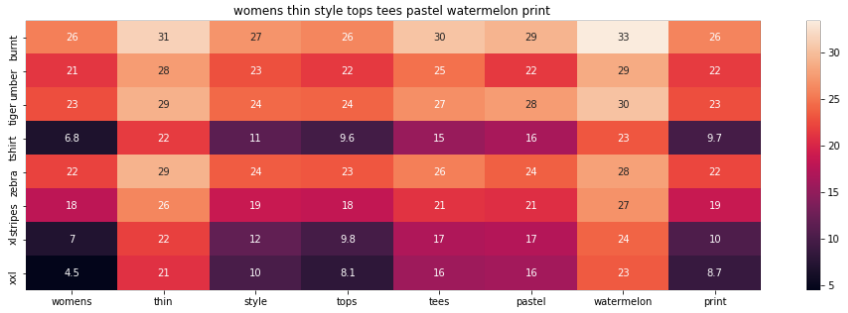
ASIN : B00JXQCUIC
Brand : Si Row
euclidean distance from input : 41.261114501953124
==========
==========



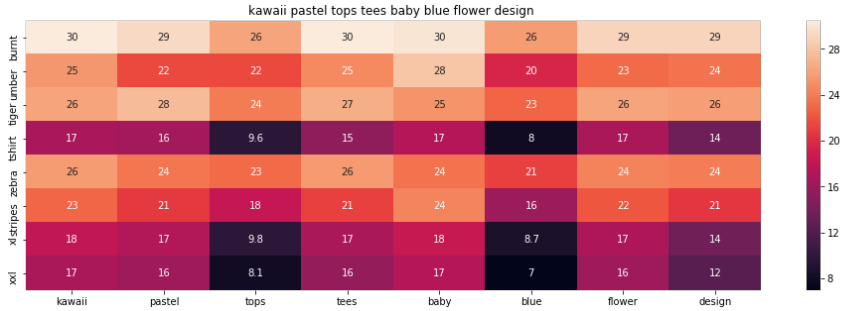
ASIN : B00JXQCWTO
Brand : Si Row
euclidean distance from input : 44.00015563964844
==========
==========



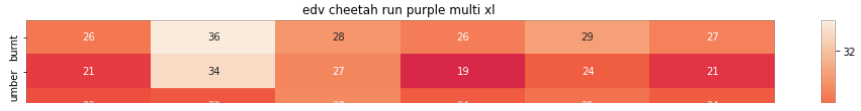
ASIN : B071FCWD97
Brand : Namnoi Clothing Store
euclidean distance from input : 47.38247985839844
==========

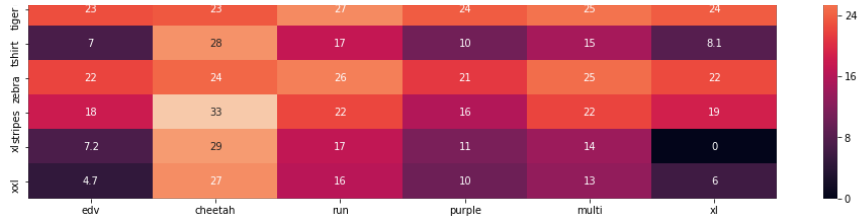


ASIN : B01JUNHBRM
Brand : Namnoi Clothing Store
euclidean distance from input : 47.71841735839844
==========

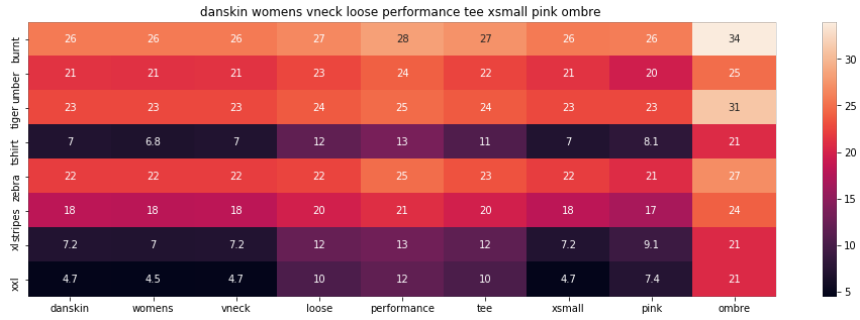


ASIN : B071SBCY9W
Brand : Namnoi Clothing Store
euclidean distance from input : 47.90206298828125
==========

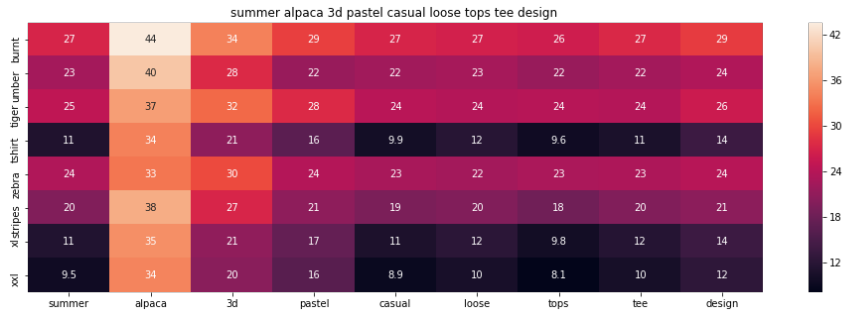




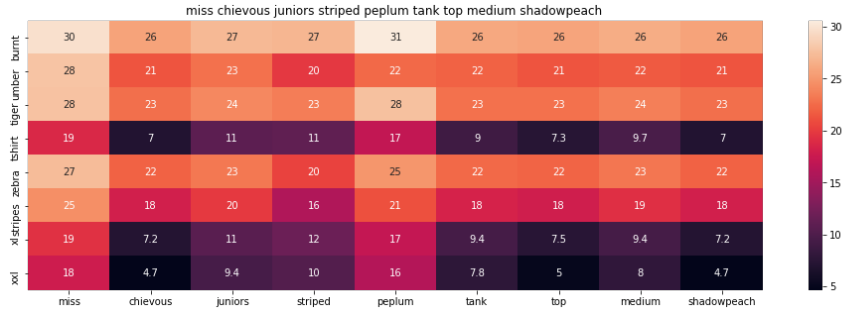
ASIN : B01CUPYBM0
Brand : Styleco
euclidean distance from input : 48.04648132324219
==========



ASIN : B01F7PHXY8
Brand : Danskin Now
euclidean distance from input : 48.101837158203125
==========

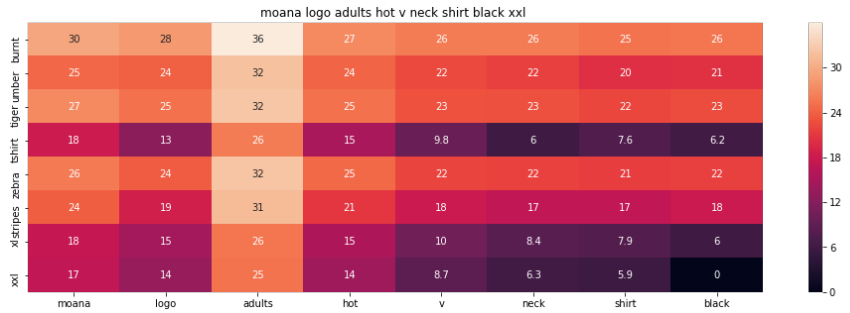


ASIN : B01I80A93G
Brand : Namnoi Clothing Store
euclidean distance from input : 48.118865966796875
==========



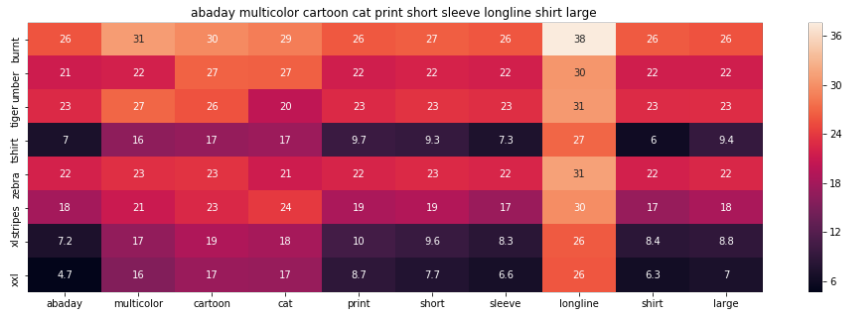
ASIN : B0177DM70S
Brand : MISS CHIEVOUS/AROUND THE WORLD APP
euclidean distance from input : 48.131222534179685
==========

ASIN : B00JV63QQE
Brand : Si Row
euclidean distance from input : 48.16944885253906
==========

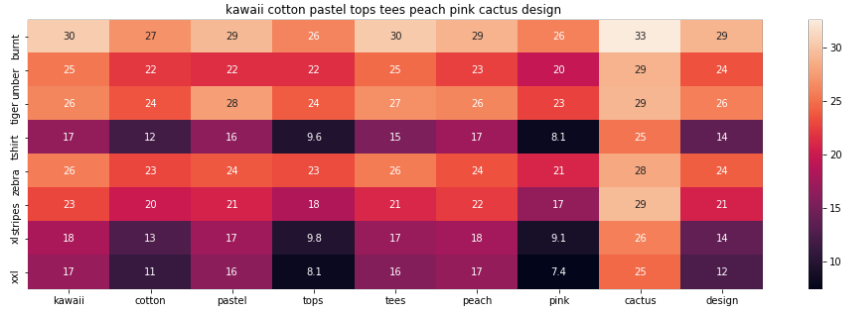


ASIN : B01LX6H43D
Brand : BOBOR

Brand : BOBOB
euclidean distance from input : 48.256787109375
==========

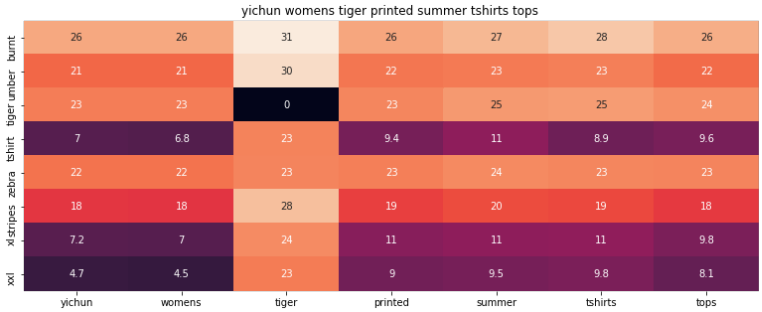


ASIN : B01CR57YY0
Brand : ABADAY
euclidean distance from input : 48.26568603515625
==========

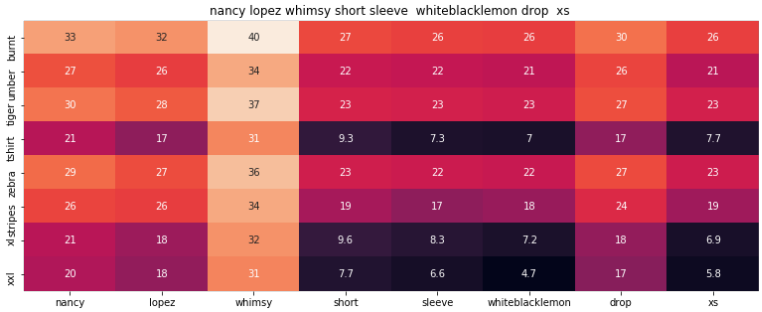


ASIN : B071WYLBZS
Brand : Namnoi Clothing Store
euclidean distance from input : 48.362603759765626
==========

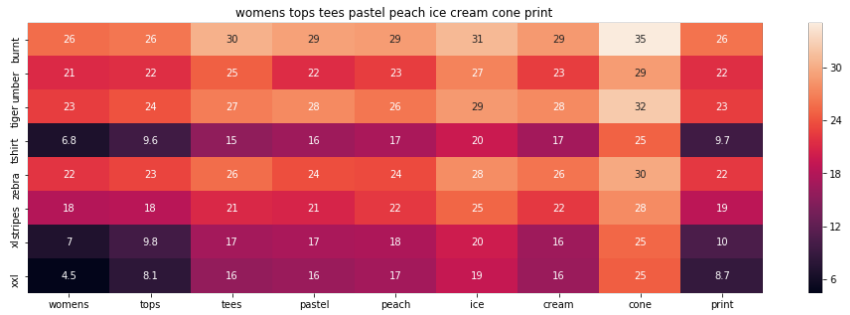
ASIN : B01GXAZTRY
Brand : Tony Arden
euclidean distance from input : 48.38360595703125



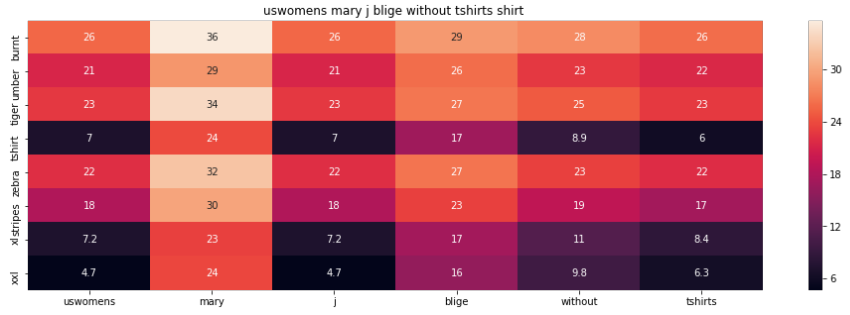
ASIN : B010NN9RXO
Brand : YICHUN
euclidean distance from input : 48.44935607910156



ASIN : B01MPX6IDX
Brand : Nancy Lopez
euclidean distance from input : 48.478887939453124



ASIN : B0734GRKZL
Brand : Namnoi Clothing Store
euclidean distance from input : 48.55795593261719



ASIN : B01M0XXFKK
Brand : South Dakota A.
euclidean distance from input : 48.614373779296876
