

INFO7374 Algorithmic Digital Marketing

Summary	In this assignment, we implemented all the three implementations. And create two respective apps to fulfill the two modes.
Author	Yifu Liu and Pengfei He

Data Preprocessing

Sampling Data:

We used the **bson_and_image preprocessing.ipynb** to sample dataset, to generate sample images, as well as to have an insight of the category-id

```
pool.join()

print('Images saved at %s' % images_dir)
print('Products: \t%d\nCategories: \t%d\nPictures: \t%d' % (pro

file = open(os.path.join(base_dir, 'retrained_labels.txt'), 'w'

rootdir_glob = images_dir + '/*/*/*'
folder_list = [f for f in iglob(rootdir_glob, recursive=True) i
for folder in folder_list:
    category = folder.split('/')[-1]
    file.write(category + '\n')

file.close()

print('"retrained_labels.txt" saved at %s' % base_dir)
```

```
82it [00:00, 101.84it/s]
Images saved at /Users/check4068/images
Products:      2
Categories:    2
Pictures:      2
"retrained_labels.txt" saved at /Users/check4068
```

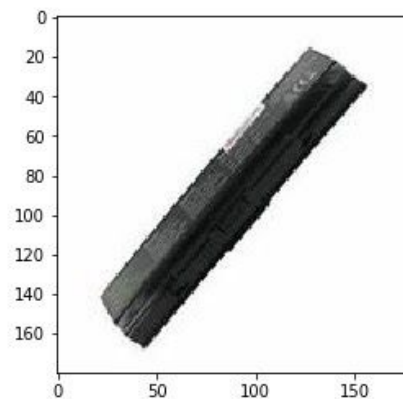
```
[5]: prod_to_category
```

```
[5]: category_id
```

_id	
0	1000010653
1	1000010653
2	1000004079
3	1000004141
4	1000015539
...	...
95	1000010653
97	1000010683
98	1000010667
99	1000014053
101	1000004085

82 rows x 1 columns

```
[6]: plt.imshow(picture);
```



Sample Images to Use:

Found [110] images

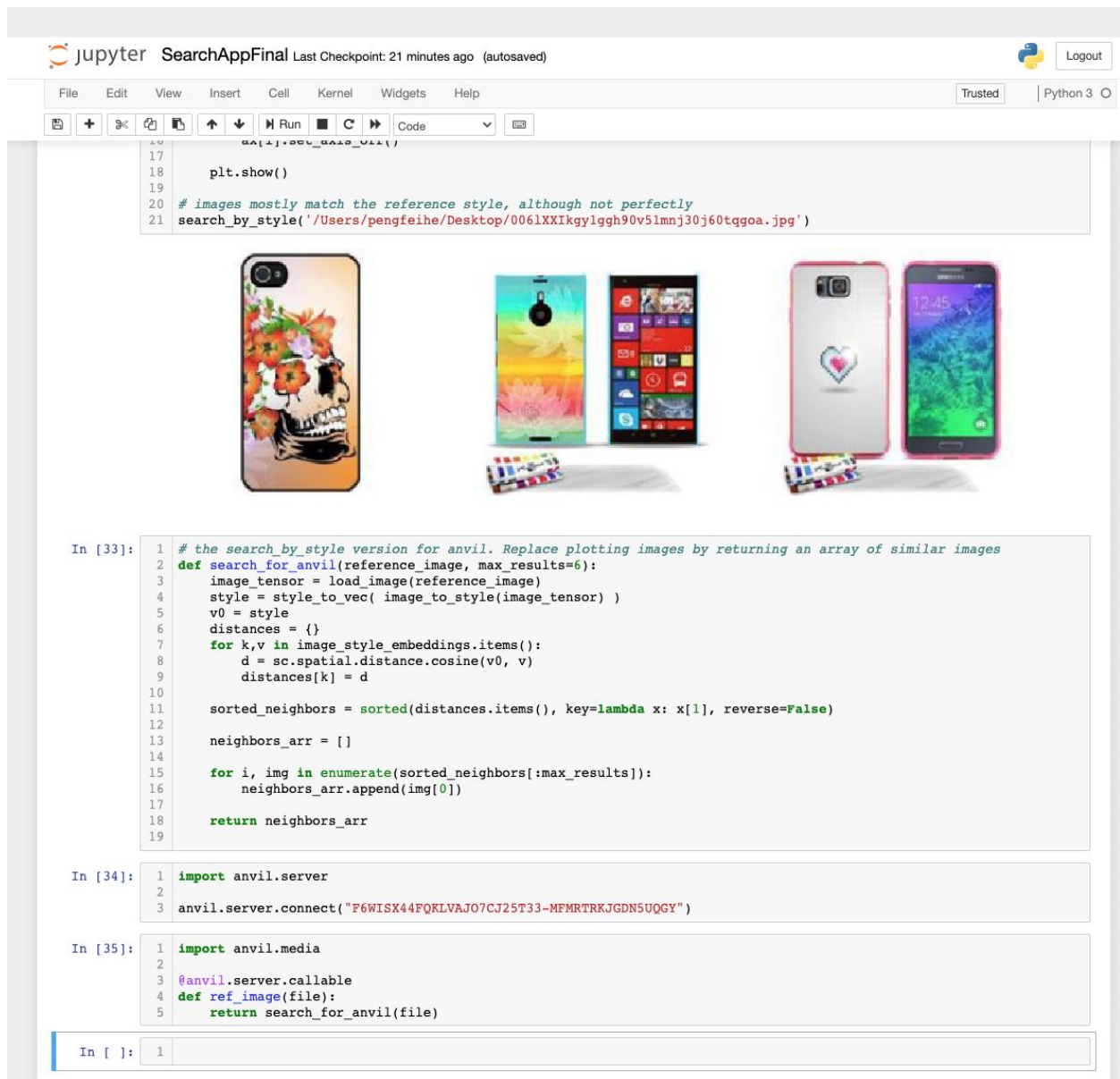


Implement of Version 1 (App mode 2)

We used the **first implement version** to create an app **fulfill the mode two**, which allows user to upload a new image and return k images similar to it. This app is published on **Anvil cloud**. The public url is **bitter-general-top.anvil.app**

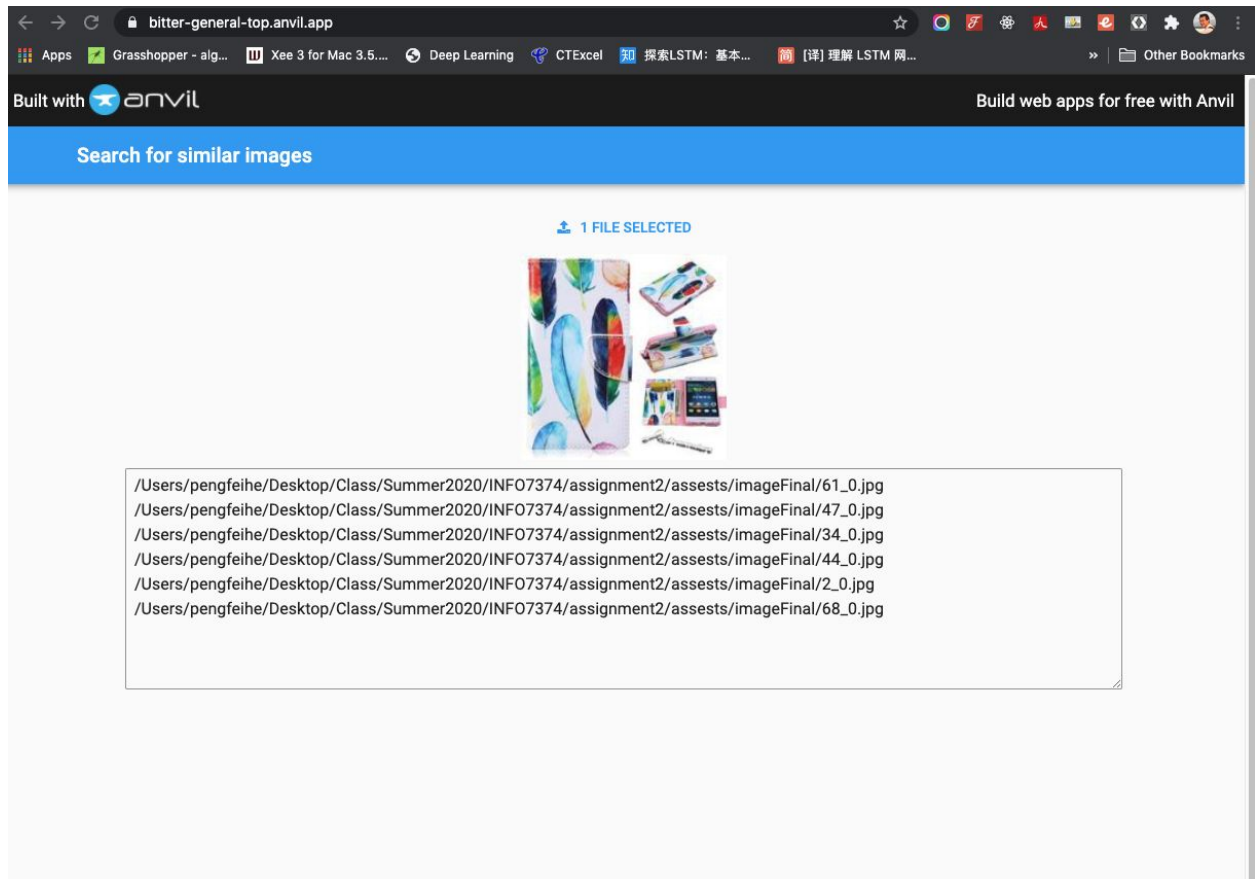
Implementing version_1 locally:

We modified the `search_by_style` function, by generating a new style for each new uploaded image, and then match this new style to the existing style array. In this way, we can allow users to **randomly upload any images**, and still can return the best similar images.



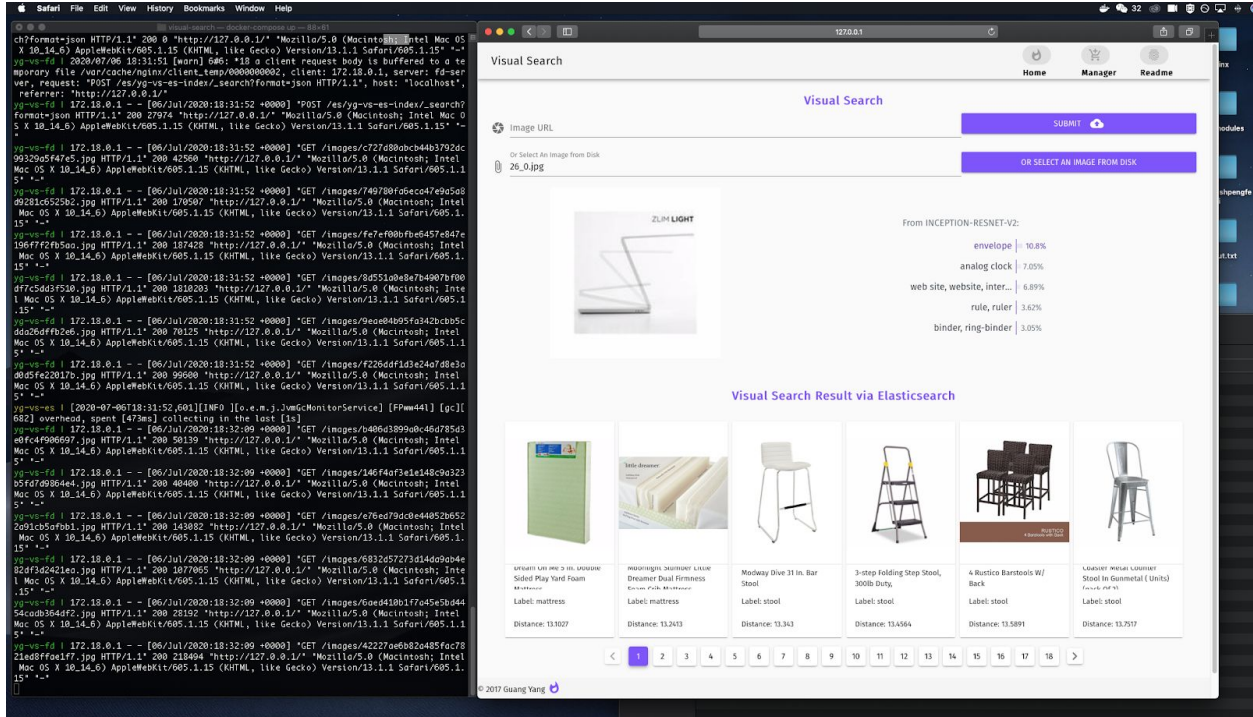
Deployed to Anvil Cloud (url:bitter-general-top.anvil.app)

We used Anvil to turn the jupyter notebook into a web app. And deployed it into Anvil Cloud server.



Implement of Version 2 (Run on docker)

Since the reference app for the **second implement version** is already a fully developed search app. We just **run it on docker** to use it.

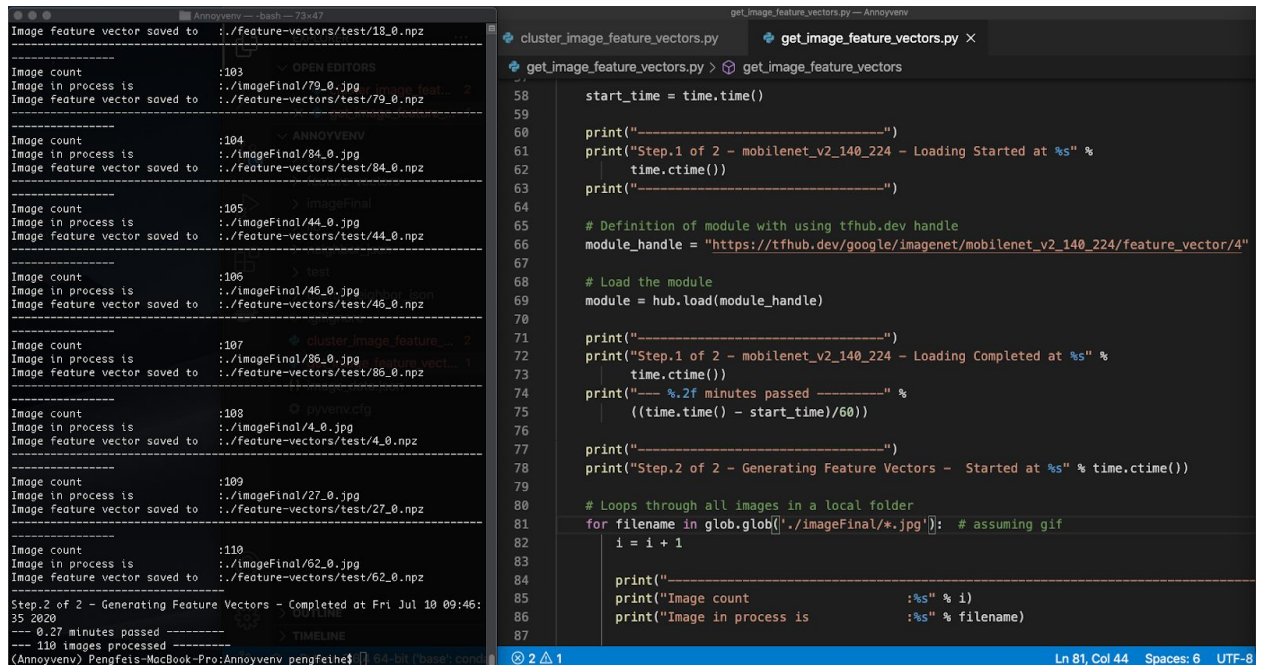


Implement of Version 3 (App mode 1)

We used the **third implement version** to create an app fulfill the **mode one**, which allows the user to select one image and get k similar images based on **elasticsearch**. Our web app is developed using django framework.

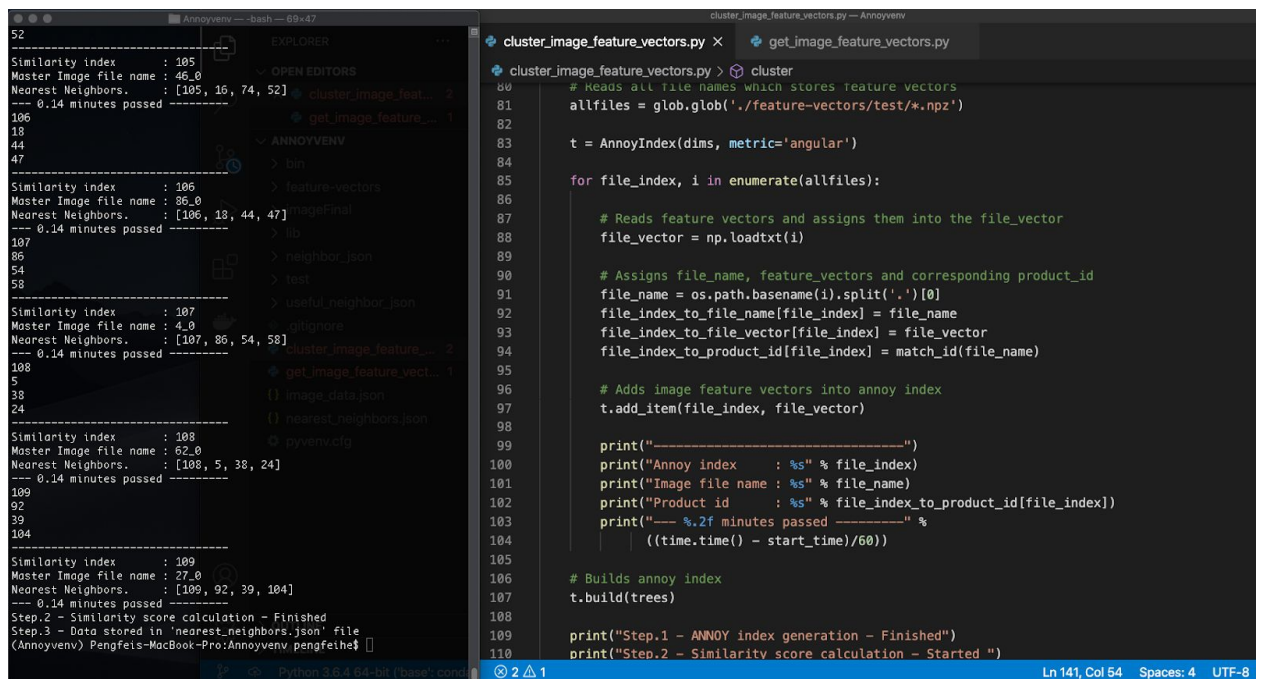
Generate N-Nearest Neighbor json file

We firstly create the vector file for each image, and then use those vectors to generate the nearest_neighbors.json file.



```
Image Feature vector saved to : ./feature-vectors/test/18_0.npz
-----
Image count : 103
Image in process is : ./imageFinal/79_0.jpg
Image Feature vector saved to : ./feature-vectors/test/79_0.npz
-----
Image count : 104
Image in process is : ./imageFinal/84_0.jpg
Image Feature vector saved to : ./feature-vectors/test/84_0.npz
-----
Image count : 105
Image in process is : ./imageFinal/44_0.jpg
Image Feature vector saved to : ./feature-vectors/test/44_0.npz
-----
Image count : 106
Image in process is : ./imageFinal/46_0.jpg
Image Feature vector saved to : ./feature-vectors/test/46_0.npz
-----
Image count : 107
Image in process is : ./imageFinal/86_0.jpg
Image Feature vector saved to : ./feature-vectors/test/86_0.npz
-----
Image count : 108
Image in process is : ./imageFinal/4_0.jpg
Image Feature vector saved to : ./feature-vectors/test/4_0.npz
-----
Image count : 109
Image in process is : ./imageFinal/27_0.jpg
Image Feature vector saved to : ./feature-vectors/test/27_0.npz
-----
Image count : 110
Image in process is : ./imageFinal/62_0.jpg
Image Feature vector saved to : ./feature-vectors/test/62_0.npz
-----
Step.2 of 2 - Generating Feature Vectors - Completed at Fri Jul 10 09:46:35 2020
--- 0.27 minutes passed ---
--- 110 images processed ---
(Annoyenv) PengFeis-MacBook-Pro:Annoyenv pengfeihe$
```

```
cluster_image_feature_vectors.py
get_image_feature_vectors.py
get_image_feature_vectors
58 start_time = time.time()
59
60 print("-----")
61 print("Step.1 of 2 - mobilenet_v2_140_224 - Loading Started at %s" %
62       time.ctime())
63 print("-----")
64
65 # Definition of module with using tfhub.dev handle
66 module_handle = "https://tfhub.dev/google/imagenet/mobilenet_v2_140_224/feature_vector/4"
67
68 # Load the module
69 module = hub.load(module_handle)
70
71 print("-----")
72 print("Step.1 of 2 - mobilenet_v2_140_224 - Loading Completed at %s" %
73       time.ctime())
74 print("---- %.2f minutes passed ----" %
75       ((time.time() - start_time)/60))
76
77 print("-----")
78 print("Step.2 of 2 - Generating Feature Vectors - Started at %s" % time.ctime())
79
80 # Loops through all images in a local folder
81 for filename in glob.glob('./imageFinal/*.jpg'): # assuming gif
82     i = i + 1
83
84     print("-----")
85     print("Image count : %s" % i)
86     print("Image in process is : %s" % filename)
87
```



```
Similarity index : 105
Master Image file name : 46_0
Nearest Neighbors. : [105, 16, 74, 52]
--- 0.14 minutes passed ---
106
18
44
47
Similarity index : 106
Master Image file name : 86_0
Nearest Neighbors. : [106, 13, 44, 47]
--- 0.14 minutes passed ---
107
54
58
Similarity index : 107
Master Image file name : 4_0
Nearest Neighbors. : [107, 86, 54, 58]
--- 0.14 minutes passed ---
108
5
38
24
Similarity index : 108
Master Image file name : 62_0
Nearest Neighbors. : [108, 5, 38, 24]
--- 0.14 minutes passed ---
109
92
39
104
Similarity index : 109
Master Image file name : 27_0
Nearest Neighbors. : [109, 92, 39, 104]
--- 0.14 minutes passed ---
Step.2 - Similarity score calculation - Finished
Step.3 - Data stored in 'nearest_neighbors.json' file
(Annoyenv) PengFeis-MacBook-Pro:Annoyenv pengfeihe$
```

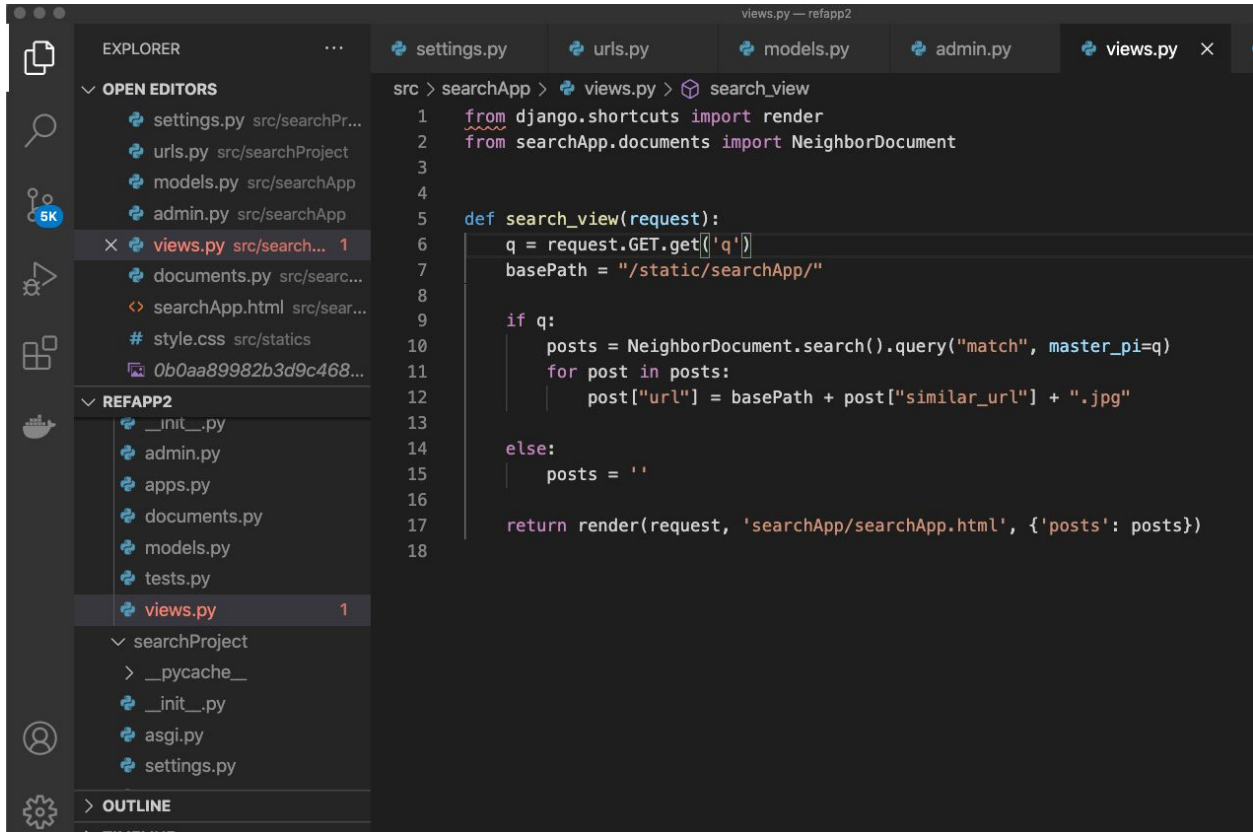
```
cluster_image_feature_vectors.py
get_image_feature_vectors.py
cluster
80 # Reads all file names which stores feature vectors
81 allfiles = glob.glob('./feature-vectors/test/*.npz')
82
83 t = AnnoyIndex(dims, metric='angular')
84
85 for file_index, i in enumerate(allfiles):
86
87     # Reads feature vectors and assigns them into the file_vector
88     file_vector = np.loadtxt(i)
89
90     # Assigns file_name, feature_vectors and corresponding product_id
91     file_name = os.path.basename(i).split('.')[0]
92     file_index_to_file_name[file_index] = file_name
93     file_index_to_file_vector[file_index] = file_vector
94     file_index_to_product_id[file_index] = match_id(file_name)
95
96     # Adds image feature vectors into annoy index
97     t.add_item(file_index, file_vector)
98
99     print("-----")
100     print("Annoy index : %s" % file_index)
101     print("Image file name : %s" % file_name)
102     print("Product id : %s" % file_index_to_product_id[file_index])
103     print("---- %.2f minutes passed ----" %
104           ((time.time() - start_time)/60))
105
106 # Builds annoy index
107 t.build(trees)
108
109 print("Step.1 - ANNOY index generation - Finished")
110 print("Step.2 - Similarity score calculation - Started ")
```

Integrate elasticsearch with django framework

We used **django-elasticsearch-dsl** for the integration. Run **python manage.py search_index --rebuild** to indexing the data model.

```
28 ALLOWED_HOSTS = []
29
30
31 # Application definition
32
33 INSTALLED_APPS = [
34     'django.contrib.admin',
35     'django.contrib.auth',
36     'django.contrib.contenttypes',
37     'django.contrib.sessions',
38     'django.contrib.messages',
39     'django.contrib.staticfiles',
40     'searchApp',
41     'django_elasticsearch_dsl',
42 ]
43
44 ELASTICSEARCH_DSL = {
45     'default': {
46         'hosts': 'localhost:9200'
47     },
48 }
49
50 MIDDLEWARE = [
```

Search based on the N-Nearest Neighbor json file



Generate UI for the app

