**Lab Assignment 3**

**Team : 1**

Charan Kothapalli - 15

Sai Srinivas Vidiyala - 27

Greg brown - 4

Sai Sri - 18

Priyanka -7

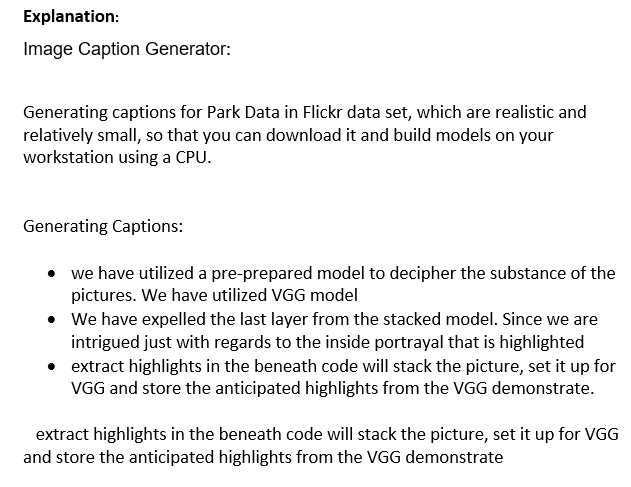
**Introduction** :

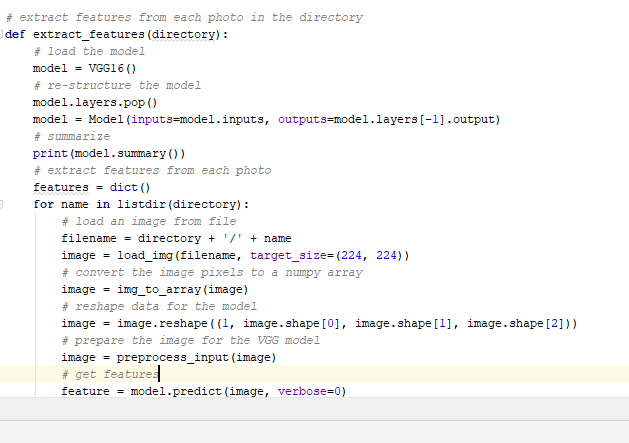
We are Generating Captions for Images on Dataset and evaluating performance, Data Analytics based on unsupervised learning.

GitHub Link:

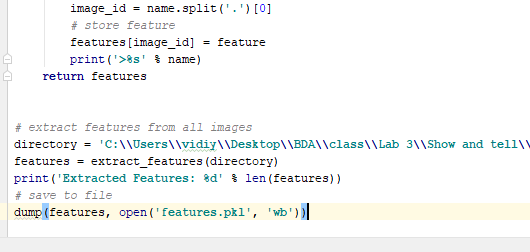
<https://github.com/vidiyalasaisrinivas/CS5542_BigdataApplications/tree/master/Lab%203>

Dataset :COCOQA

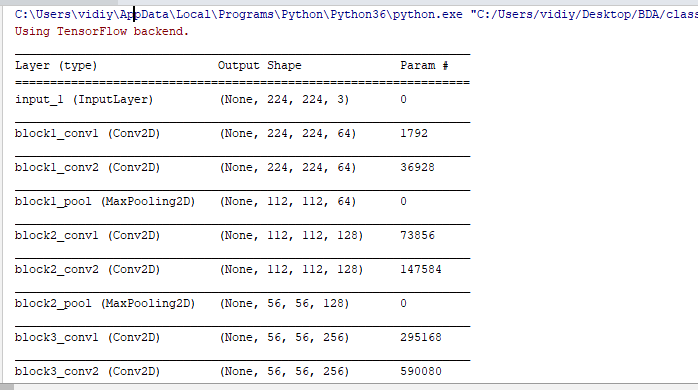




Save results in *features.pkl*‘ ,and results will be used later



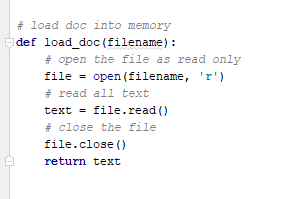
Output:



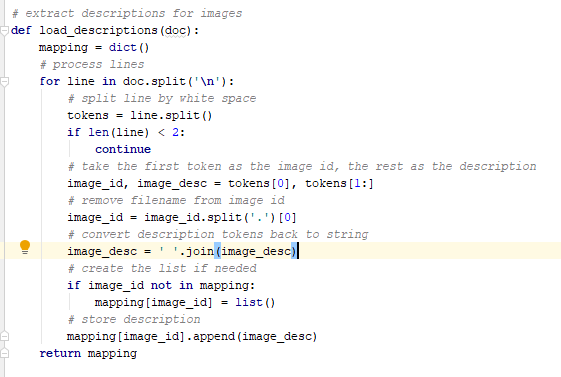
**Captions**:

We will input token.txt as input and we will be getting description.txt as outcome

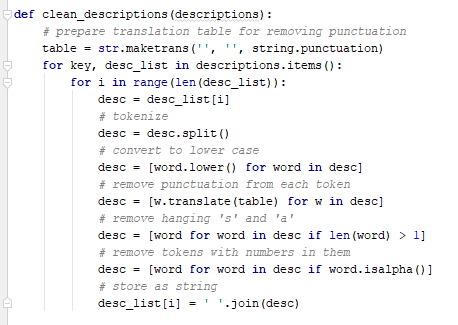
We will load token.txt using load\_doc()



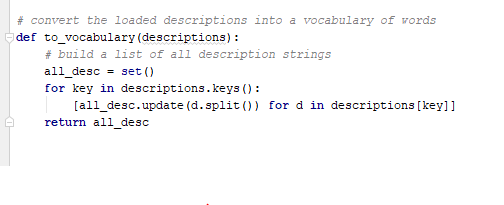
load\_descriptions() that, given the loaded document text, will return a dictionary of photo id to descriptions. Each photo id maps to a list of one or more textual descriptions.



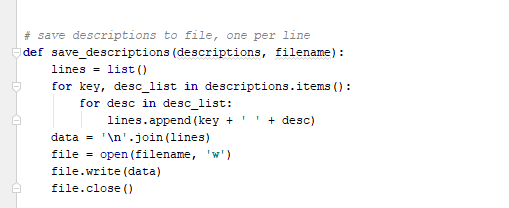
We cleaned the descriptions that are changing over all words to lowercase, expel all punctuations, expel all words less than one or break even with to one character and expel words with numbers in clean\_description().



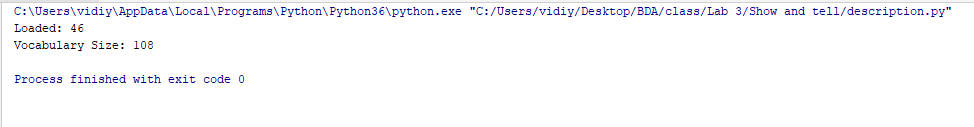
Converted loaded desc to vocabulary of words



At last, we put away these cleaned desc to descriptions.txt in save\_descriptions().



Outcome

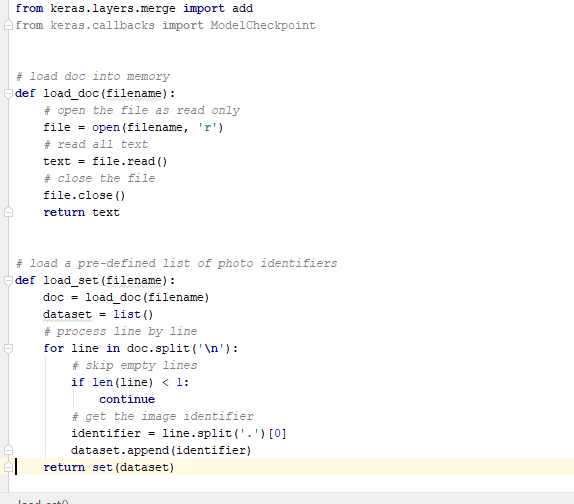


c)

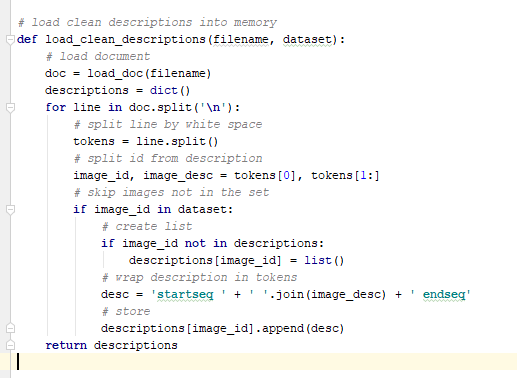
Train deep learning model

Images and description.txt,features.pk are input ,result is h5 model

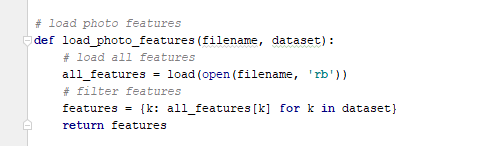
1.load data



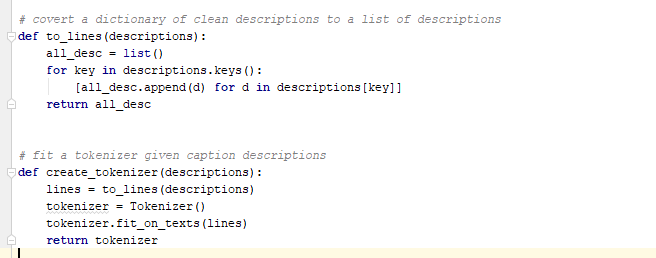
load\_clean\_descriptions() will presently stack the cleaned portrayals from "description.txt" and return the dictionary of identifiers to the list of content captions. we moreover include "startseq" and "endseq" as the conclusion demonstrate will require to begin with word to kick-off the content era prepare and the final word to flag the conclusion of the method.



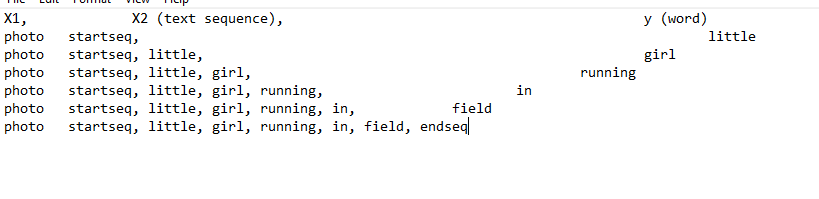
Load photo features



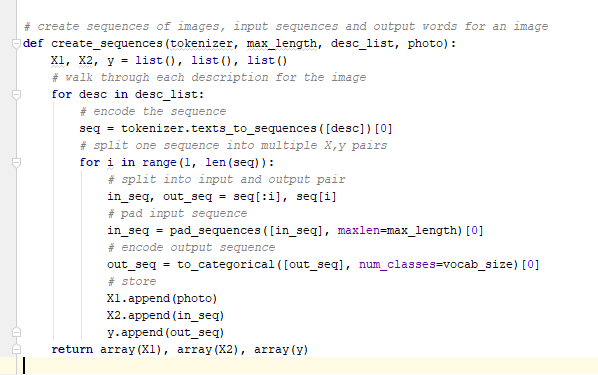
to\_lines() will change over the lexicon into a list of strings and create\_tokenizer to fit a tokenizer

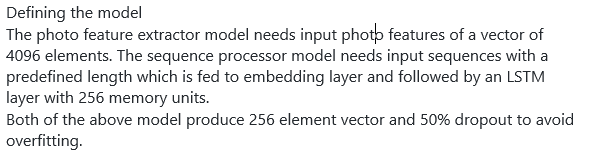


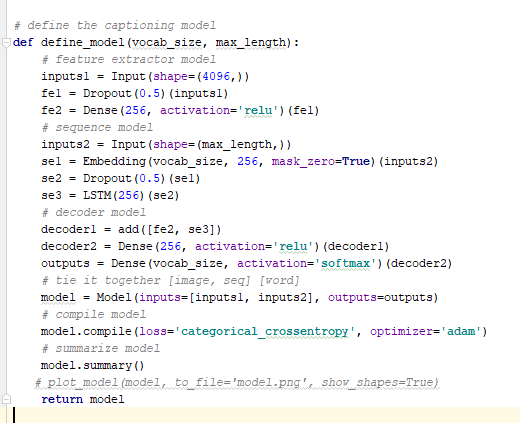
For example, the input sequence “little girl running in field” would be split into 6 input-output pairs to train the model:

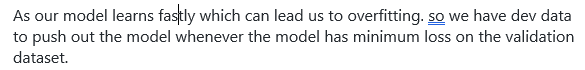


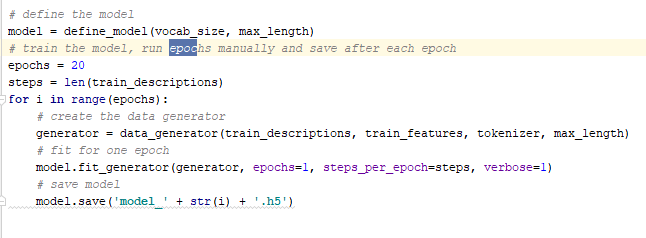




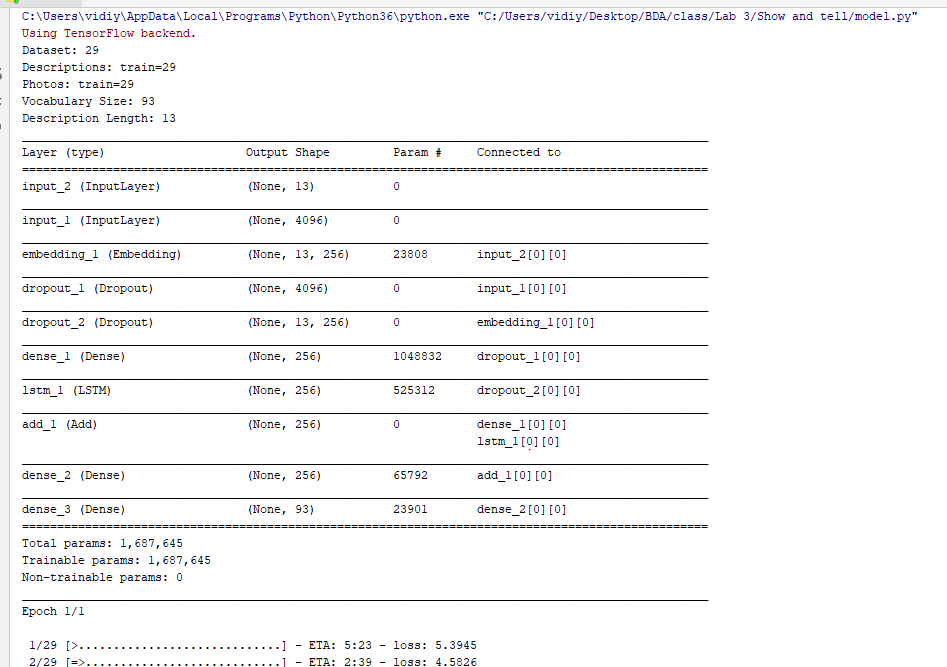




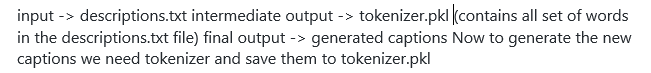


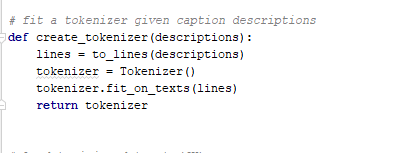


OutCome

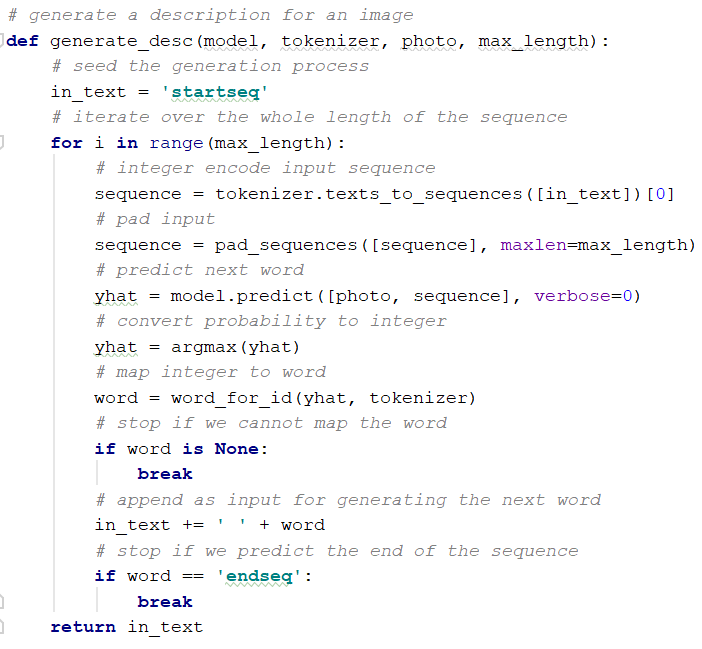


**Generate captions using trained model**

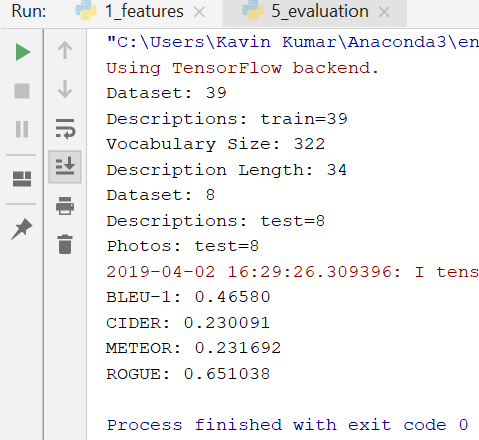




At final we make or produce the arrangement with startseq for our test picture



**Accuracy**

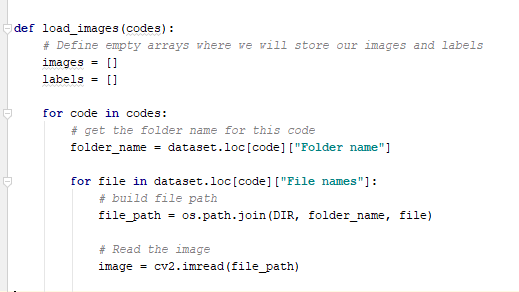


2.Analytics using unsupervised learning

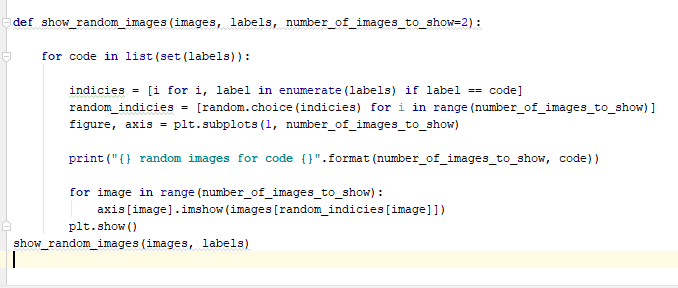
We given input images in different folders ,it categorised based on unsupervised learning

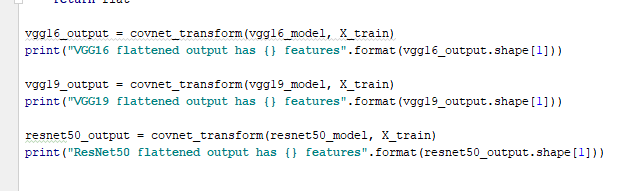
We performed various clustering techniques

Load images

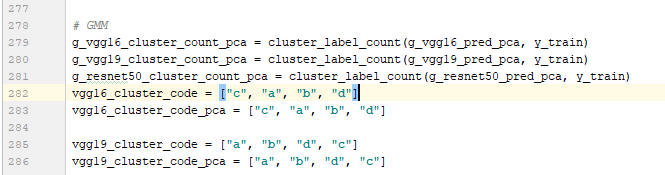


Random images

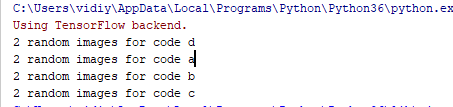




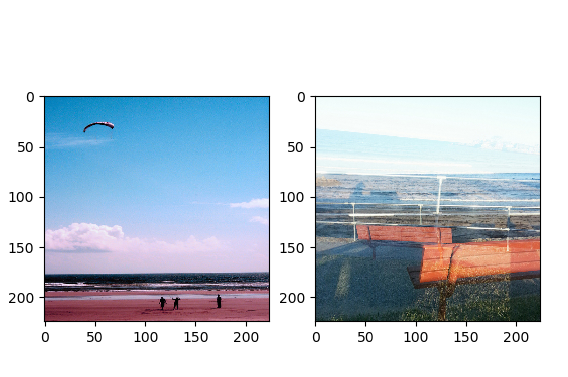




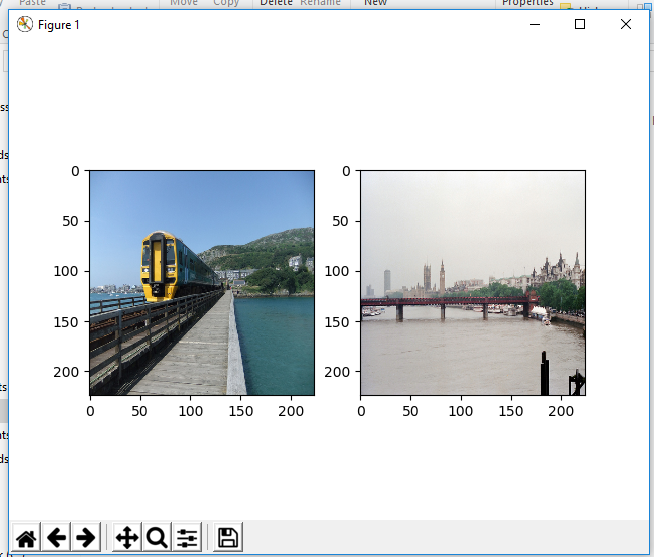
Below are some of the outputs



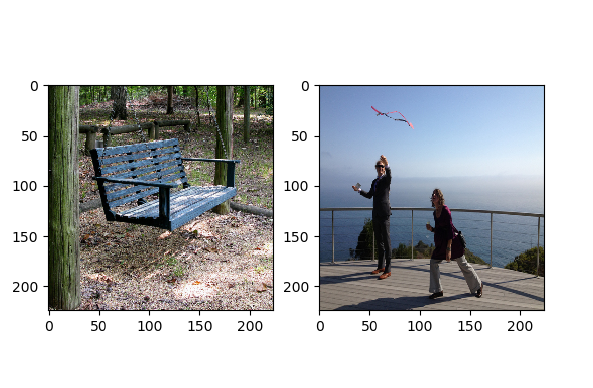
Code d



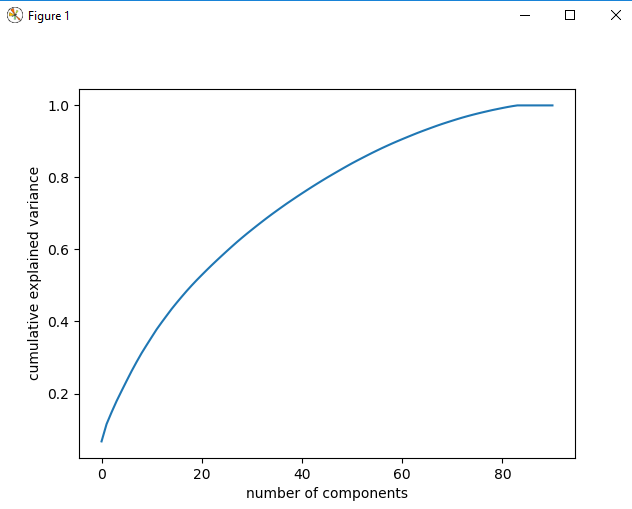
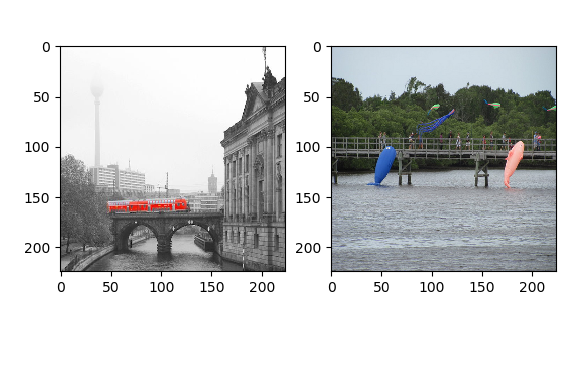
Code a



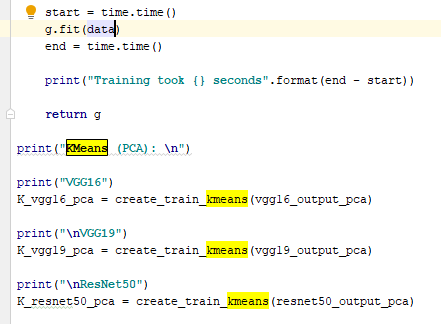
Code b

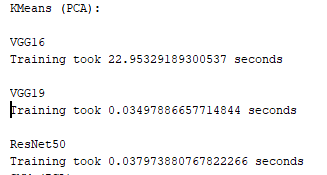


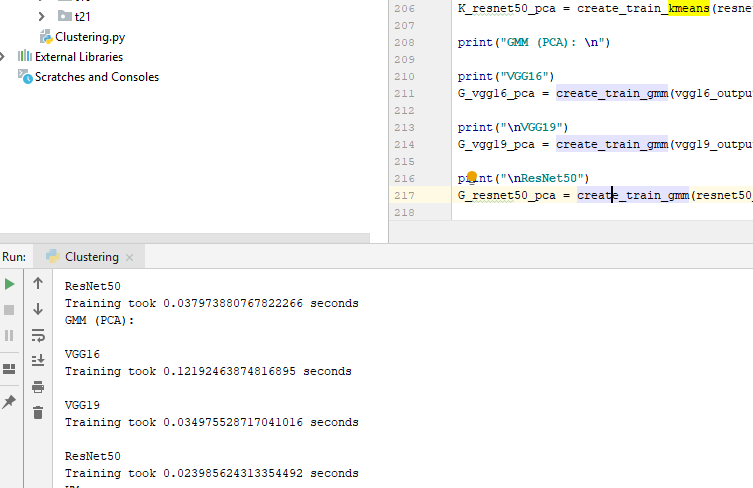
Code c

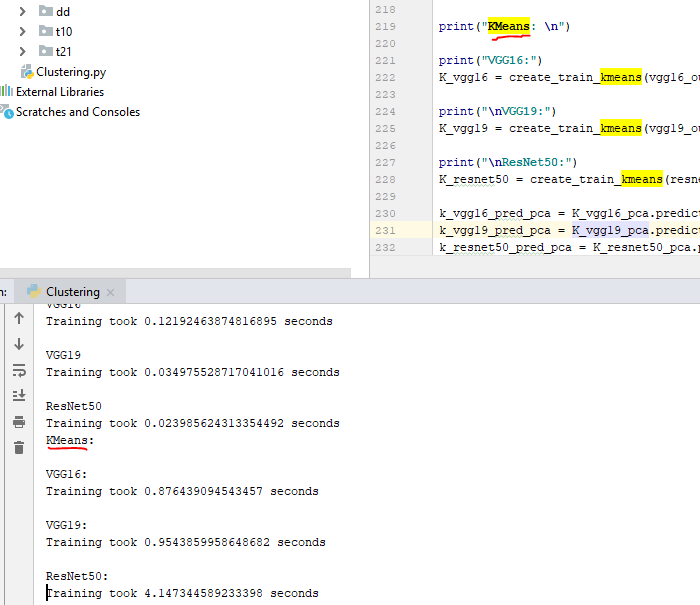


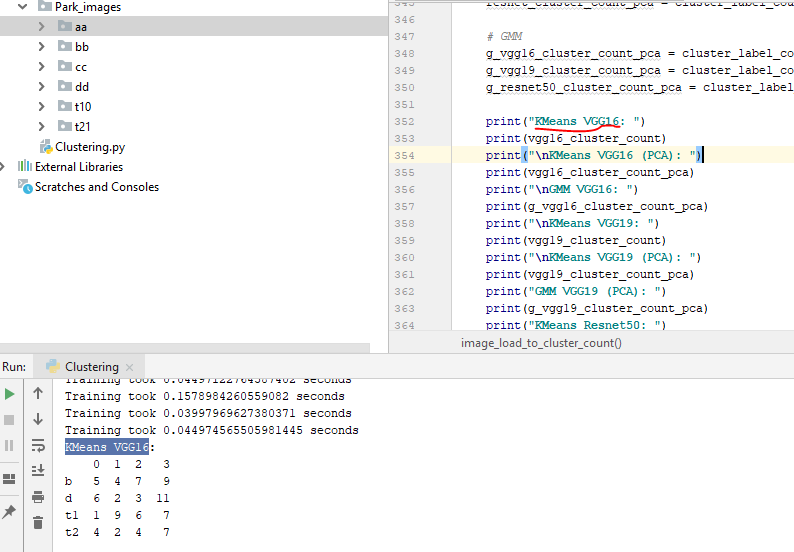
Kmeans

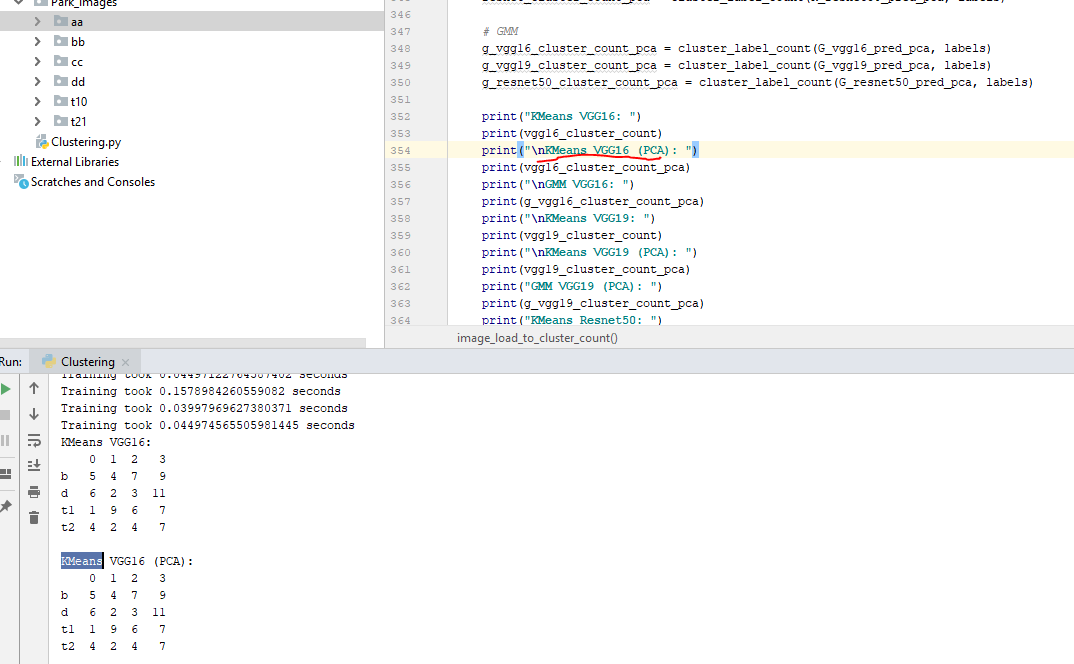












Different types of clustering technique results for my data

