Cluster size from 3.6 Clustering features to form a codebook

Oct 10

1 / 7
Oct 11

7d ago

marchampson **8d**

Hi,

Is there a formula for calculating the cluster size based on the amount of images in a dataset? I want datasets to be built up by users over time, so initially one may only have 1 or 2 images in it.

My sample dataset has in fact only got 2 images now but fails to generate a codebook using the following arguments: --clusters 1536 --percentage 0.25

I get:
ValueError: Number of samples smaller than number of clusters.

which calls from k_means.py around line 1300:
if n_samples < self.n_clusters:
raise ValueError("Number of samples smaller than number "
"of clusters.")

I can only get the codebook to be generated if I change clusters to be '1' - I'm sure I've just misread something or overlooking something simple but I've been looking for a few hours and need some pointers if possible as to how those values should be calculated.

Thanks

Adrian Chief PyImageSearcher **8d**

Let's say you have X images and you want to generate K clusters. In order to generate K clusters the number of images X needs to be greater than K, which implies X + 1 >= K

This same logic extends to keypoints and local invariant descriptors. Using the command above you are setting K=1536. But if you inspect the number of extracted keypoints and local invariant descriptors you'll find this value is less than 1536.

In short, check the number of features extracted from your images and then adjust K to be smaller than this value.

marchampson **8d**

Great, thanks Adrian. Will take another look.

marchampson **7d**

OK, I might get this by the time you come online in which case I'll post, but just in case. My 2 image dataset has found 1,654 features in total. I then run the cluster script with the following values:

vocab = voc.fit(300, 0.25)

If I go for 25% of my total features, I get a sample size of 414 features and I'm assuming your last answer was to make my n_clusters value less than this figure? At 300 it is, but for some reason I still get the same error. If my logic is right then it's a bug and I'll find and fix, but it it's wrong then I'll never find it. Command line out put below.

Thanks

(cv) marcs-mbp-2:cbir marchampson$ py index_features.py
dj3HPl2Pfk.jpg
Num keypoints: 589
qmK7j3x45u.jpg
Num keypoints: 1065

[WARN] minimum init buffer not reached - 2016-10-11 10:02:25.717524
[INFO] creating datasets... - 2016-10-11 10:02:25.717605
[INFO] writing un-empty buffers... - 2016-10-11 10:02:25.732455
[INFO] writing `image_ids` buffer - 2016-10-11 10:02:25.732538
[INFO] writing `index` buffer - 2016-10-11 10:02:25.739462
[INFO] writing `features` buffer - 2016-10-11 10:02:25.740191
[INFO] compacting datasets... - 2016-10-11 10:02:25.742405
[INFO] old size of `image_ids`: 2; new size: 2 - 2016-10-11 10:02:25.742510
[INFO] old size of `index`: 2; new size: 2 - 2016-10-11 10:02:25.742561
[INFO] old size of `features`: 1,654; new size: 1,654 - 2016-10-11 10:02:25.742602
(cv) marcs-mbp-2:cbir marchampson$ py cluster_features.py
Total features 1654
sampleSize 414
numClusters: 300
[INFO] starting sampling... - 2016-10-11 10:02:31.091681
[INFO] sampled 414 features from a population of 1,654 - 2016-10-11 10:02:31.092623
[INFO] clustering with k=300 - 2016-10-11 10:02:31.092652
n_clusters 300 random_state None
Traceback (most recent call last):
File "cluster_features.py", line 11, in
vocab = voc.fit(300, 0.25)
File "/Users/marchampson/Desktop/cbir/mainwaring/ir/vocabulary.py", line 44, in fit
clt.fit(data)
File "/Users/marchampson/.virtualenvs/cv/lib/python2.7/site-packages/sklearn/cluster/k_means_.py", line 1238, in fit
raise ValueError("Number of samples smaller than number "
ValueError: Number of samples smaller than number of clusters.
(cv) marcs-mbp-2:cbir marchampson$

---

marchampson **7d**

It was me, I had a tab error in vocabulary.py so the clustering section was under the for i in idxs: loop

All sorted now.

Cheers

Adrian Chief PyImageSearcher **7d**

> marchampson:
>
> It was me, I had a tab error in vocabulary.py so the clustering section was under the for i in idxs: loop

Nice job resolving the bug 🙂 Also, in the future if you want to post code or terminal output make sure you use the `pre` HTML tag to format the code/output blocks to make it more readable.

---

marchampson **7d**

`(cv) marcs-mbp-2:cbir marchampson$ Thanks Adrian, will do.`