23/01/24

Bag of Words

Dictionary:

He likes machine learning. He did a course on machine learning. Now he does job in the area of machine learning and he likes it.

He - 4

likes - 2

machine - 3

learning - 3

⋮

→ Decide a static dictionary.

→ Every document will be represented using the dimension of the dictionary.

Read about tf.idf HW

→ Cosine Distance works well here.

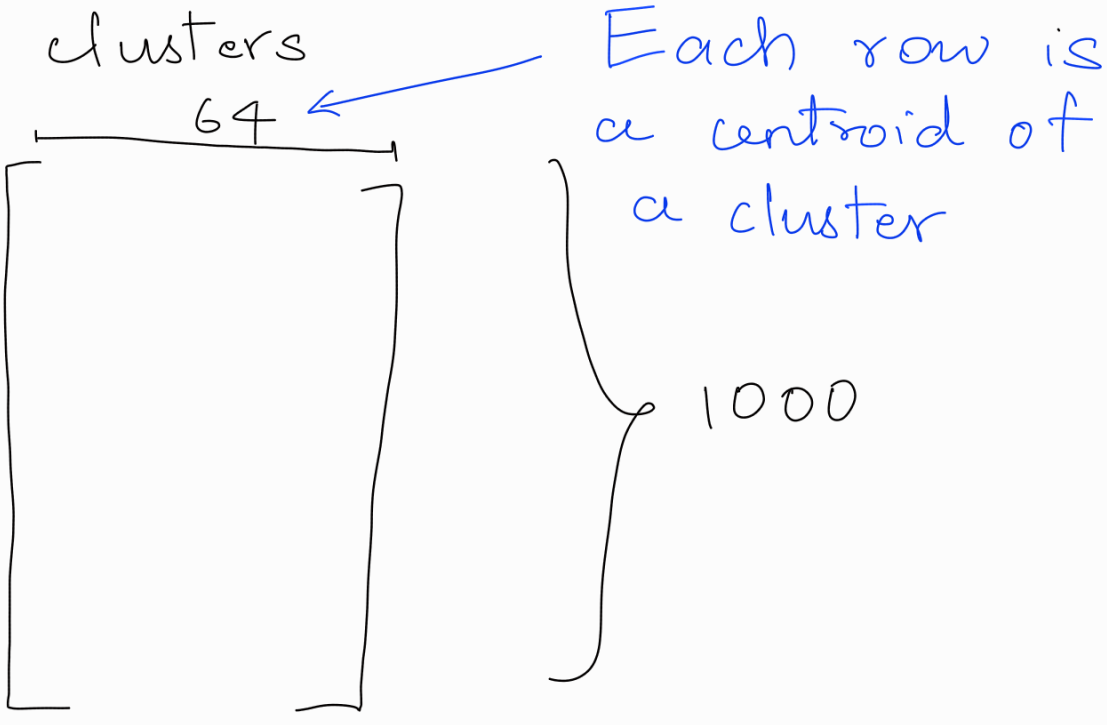
Visual Bag of Words

→ Similar kind of representation for images.

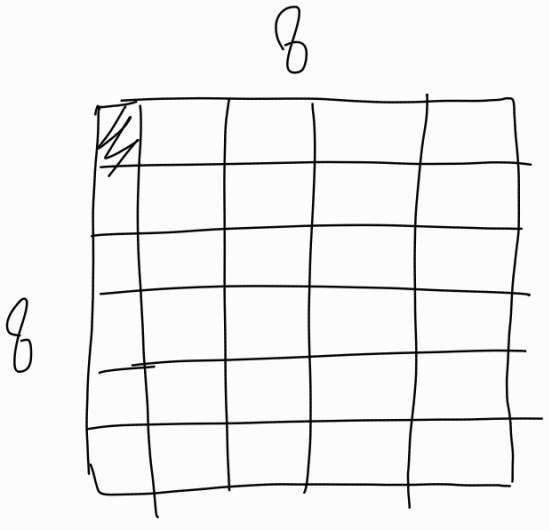
→ Consider 10,000 images

Break each image into 100 patches of 8x8 size each. Similar patches clustered together using K-means.

1000 clusters



The clusters which are very large very likely represent the equivalent of stop words. We throw away the top 50 clusters.



→ Find distance of each patch of an image from all the centroids and increment the closest centroid.

→ Now this is the representation of the image.