Clustering and LDA on Images and Texts

1. Pinterest data is not properly categorized. Hence direct application of any classification algorithm may not resulting in desirable results

3. Pinterest data is clustered in many sense — boards, related pins etc.

4. Clustering or other un-supervised algorithm may be better evaluated.

Caffe has a good object detection function and some well trained model.
Image => objects contained in: person, cat, ….
Use these words as input for later algorithm

Cluster directly with image words, description words
OR
LDA —> a set of topics of images: ImgSet
LDA —> a set of topics of descriptions: TxtSet
Cluster with ImgSet, TxtSet

Evaluation:

1. the cost of two images within a board is 0.
2. the cost of two images that are related is 1.
3. the cost of two images that are related by a third image is 2

compare the final costs.

AWS s3 access key