

Clustering Demos

Pinak, Srashti

March 2020

Document classification with Hierarchical clustering

Dataset

20 newsgroups dataset available in [scikit](#)

Notebook

1. Load and look at sample data
2. Filter text (section 5.6.2.3) and look at the filtered text
3. Vectorize text
4. HAC with [sklearn.cluster](#) or [scipy.cluster](#)
5. Plot truncated dendrogram
6. Figure out the number of clusters based on [silhouettes](#)
7. Figure out the number of misclassifications

To Learn

1. Hierarchical Clustering
2. Bag of words model and TF-IDF
3. Silhouettes
4. Manipulation of 20 newsgroups dataset
5. Using Python to implement HAC

Edge detection with k-means

Dataset

Images collected off the web

Notebook

1. Load and create greyscale image
2. Compute features
3. Use k-means to find edges