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

To Learn

1. k-means
2. Greyscaling an image in Python using Pillow
3. What are the features?
4. How to detect an edge pixel?
5. Creating an image from edge pixels

Notebook

1. Load and create greyscale image
2. Show that greyscaling preserves the edges
3. Compute features
4. Use k-means to find edge pixels
5. Create an image from edge pixels using Pillow

Spatial Clustering with four algorithms

Dataset

Mall Customers data with DBSCAN algorithm applied on github.

Notebook

1. Replicate the python script for DBSCAN, explaination in youtube
2. Apply the other three algorithms: HAC, Spectral and K-means
3. Compare: pros and cons of each algorithm

ref: scikit implementation of all clustering algorithms.

To Learn

1. DBSCAN Clustering
2. Spectral Clustering
3. Using python for implementing clustering algorithms

Image Segmentation with DBSCAN

Dataset

Images collected off the web