**Web Mining (CSE3024)**

**Lab Assignment 8**

Name: **Kritika Mishra**

Registration Number: **16BCI0041**

Slot: L15+L16

Date: 3rd October 2018

Question:

**Implement a k-means algorithm with sklearn to partition observations in a dataset into a specific number of clusters in order to aid in analysis of the data.**

* **Use Sklearn Toolkit / Package to perform the process**
* **Import kmeans and PCA through the sklearn library**
* **Devise an elbow curve to select the optimal number of clusters (k)**
* **Generate and visualise a k-means clustering algorithms**

**Note : Dataset in CSV can be generated or downloaded from the internet. Please specify the source of the dataset in the documentation steps of this program.**

Dataset:

<http://www.michaeljgrogan.com/datasets/> ->sample\_stocks.csv

Code:

import pandas

import pylab as pl

from sklearn.cluster import KMeans

from sklearn.decomposition import PCA

variables = pandas.read\_csv('sample\_stocks.csv')

Y = variables[['returns']]

X = variables[['dividendyield']]

X\_norm = (X - X.mean()) / (X.max() - X.min())

Y\_norm = (Y - Y.mean()) / (Y.max() - Y.min())

Nc = range(1, 20)

kmeans = [KMeans(n\_clusters=i) for i in Nc]

kmeans

score = [kmeans[i].fit(Y).score(Y) for i in range(len(kmeans))]

score

pl.plot(Nc,score)

pl.xlabel('Number of Clusters')

pl.ylabel('Score')

pl.title('Elbow Curve')

pl.show()

pca = PCA(n\_components=1).fit(Y)

pca\_d = pca.transform(Y)

pca\_c = pca.transform(X)

kmeans=KMeans(n\_clusters=3)

kmeansoutput=kmeans.fit(Y)

kmeansoutput

pl.figure('3 Cluster K-Means')

pl.scatter(pca\_c[:, 0], pca\_d[:, 0], c=kmeansoutput.labels\_)

pl.xlabel('Dividend Yield')

pl.ylabel('Returns')

pl.title('3 Cluster K-Means')

pl.show()

Output:





