

```

#k_means_algo_mod.py
import math
import numpy
import random
import graphToolKit as gtk

def findDistance(obj1, obj2):
    distance = 0
    for i in range(len(obj1)):
        distance += (obj1[i] - obj2[i])**2
    return math.sqrt(distance)

def findSquaredDistance(obj1, obj2):
    distance = 0
    for i in range(len(obj1)):
        distance += (obj1[i] - obj2[i])**2
    return distance

def findCluster(obj1, cent1, cent2, cent3):
    distances = []
    distances.append(findDistance(obj1, cent1))
    distances.append(findDistance(obj1, cent2))
    distances.append(findDistance(obj1, cent3))
    return distances.index(min(distances)) + 1

def findMean(cluster):
    uval = wval = xval = yval = 0
    for obj in cluster:
        uval += obj[0]
        wval += obj[1]
        xval += obj[2]
        yval += obj[3]
    size = len(cluster)
    return [(uval/size), (wval/size), (xval/size), (yval/size)]

def findSSE(centroids, clusters):
    sse = 0
    for i in range(3):
        for obj in clusters["cluster" + str(i + 1)]:
            sse += findSquaredDistance(obj, centroids[i])
    return sse

# taking input from file
dataSet = []

```

```

dataFile = open("iris.data", "r")
for line in dataFile:
    obj = []
    x = line.strip().split(",")
    for i in range(4):
        obj.append((float)(x[i]))
    dataSet.append(obj)

random.shuffle(dataSet)

finalClusters = {}
minimalSSE = 0
sseValues = []

for i in range(150):
    # initialize centroid values
    cent = numpy.array(random.sample(dataSet, 3))
    # print(cent)
    flag = "all_good"
    cluster1 = []
    cluster2 = []
    cluster3 = []

    while True:
        cluster1.clear()
        cluster2.clear()
        cluster3.clear()
        for obj in dataSet:
            cluster = findCluster(obj, cent[0], cent[1], cent[2])
            if cluster == 1:
                cluster1.append(obj)
            elif cluster == 2:
                cluster2.append(obj)
            else:
                cluster3.append(obj)

        if len(cluster1) == 0 or len(cluster2) == 0 or len(cluster3) == 0:
            flag = "empty_cluster"
            break

        newCent = numpy.array([findMean(cluster1), findMean(cluster2),
findMean(cluster3)])

        compare = cent == newCent

```

```

        if compare.all() :
            break
        else:

            cent = numpy.delete(cent, [0,1,2],0)
            cent = newCent

    if flag == "empty_cluster":
        print("invalid clustering...iteration skipped!")
        continue

    # print("iteration: " + str(i))
    clusters = {}
    clusters["cluster1"] = cluster1
    clusters["cluster2"] = cluster2
    clusters["cluster3"] = cluster3

    newSSE = findSSE(cent, clusters)
    sseValues.append(newSSE)
    # print(newSSE)

    if newSSE < minimalSSE or i == 0:
        minimalSSE = newSSE
        finalClusters.clear()
        for cluster in clusters:
            finalClusters.update({cluster : clusters[cluster]})

print("minimal SSE = " + str(minimalSSE))
for cluster in finalClusters:
    print(cluster)
    print(finalClusters[cluster])

gtk.plot3DGraph(clusters)
gtk.plot4DGraph(clusters)
gtk.plotSSEGraph(sseValues)
gtk.plot2DGraph(clusters)

#graphToolKit

import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

```

```

import numpy

def plot4DGraph(clusters):

    fig = plt.figure()
    ax = fig.add_subplot(111, projection='3d')

    iter = 0
    for cluster in clusters:
        u_val = [obj[0] for obj in clusters[cluster]]
        v_val = [obj[1] for obj in clusters[cluster]]
        w_val = [obj[2] for obj in clusters[cluster]]
        x_val = [obj[3] for obj in clusters[cluster]]

        if iter == 0:
            img1 = ax.scatter(u_val, v_val, w_val, c = x_val, cmap = plt.winter(),
label = 'cluster1')
            cbar = fig.colorbar(img1, shrink = 0.5, aspect = 10)
        elif iter == 1:
            img2 = ax.scatter(u_val, v_val, w_val, c = x_val, cmap = plt.spring(),
label = 'cluster2')
            cbar = fig.colorbar(img2, shrink = 0.5, aspect = 10)
        else:
            img3 = ax.scatter(u_val, v_val, w_val, c = x_val, cmap = plt.gray(),
label = 'cluster3')
            cbar = fig.colorbar(img3, shrink = 0.5, aspect = 10)

        iter += 1
        cbar.ax.get_yaxis().labelpad = 15
        cbar.ax.set_ylabel('petal width in cm')
        cbar.ax.get_xaxis().labelpad = 15
        cbar.ax.set_xlabel('cluster' + str(iter))

    ax.set_xlabel('sepal length in cm', rotation=150)
    ax.set_ylabel('sepal width in cm')
    ax.set_zlabel(r'petal length in cm', rotation=60)

    plt.title("4D representation of clustering solution")
    plt.show()

def plot3DGraph(clusters):

```

```

fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
colorArray = ['red', 'green', 'blue']
iter = 0

for cluster in clusters:
    u_val = [obj[0] for obj in clusters[cluster]]
    v_val = [obj[1] for obj in clusters[cluster]]
    w_val = [obj[2] for obj in clusters[cluster]]

    ax.scatter(u_val, v_val, w_val, s = 75, c = colorArray[iter], label =
'cluster' + str(iter + 1))
    iter += 1

plt.legend()
ax.set_xlabel('sepal length in cm', fontsize=13, rotation=150)
ax.set_ylabel('sepal width in cm', fontsize=13)
ax.set_zlabel('petal length in cm', fontsize=13, rotation=60)
plt.title("3D representation of clustering solution")
plt.show()

def plotSSEGraph(sseValues):

    x_val = numpy.arange(1,151,1)
    y_val = sseValues
    plt.plot(x_val, y_val)
    plt.scatter(x_val, y_val, c = "red", marker= '+', label = "round 1")

    plt.xlabel("iteration")
    plt.ylabel("SSE values")
    plt.title("iteration vs SSE values")
    plt.grid()
    plt.legend()
    plt.show()

def plot2DGraph(clusters):

    colorArray = ['red', 'green', 'blue']
    attributes = ["sepal length", "sepal width", "petal length", "petal width"]
    for i in range(0,3,2):
        iter = 0
        for cluster in clusters:
            u_val = [obj[0 + i] for obj in clusters[cluster]]
            v_val = [obj[1 + i] for obj in clusters[cluster]]

```

```
plt.scatter(u_val, v_val, s = 50, c = colorArray[iter], label =
"cluster" + str(iter + 1))
    iter += 1

# plt.grid()
plt.xlabel(attributes[0 + i] + "(cm)", fontsize = 15)
plt.ylabel(attributes[1 + i] + "(cm)", fontsize = 15)
plt.title(attributes[0 + i] + " vs " + attributes[1 + i] + " of clusters",
fontsize = 20)
plt.show()
```

## ScreenShots









