作业七

林恺越 181098163

实验步骤

因为给的源代码直接可以跑起来，所以kmeans主体部分代码没有需要改动的。

在使用bdkit的时候一开始是因为无法访问gitlab导致集群无法创建成功，使用github后成功创建，另外在vscode的终端上会无法push，网页显示被forbidden，需要到控制台上，安装git再输入命令才能成功push。

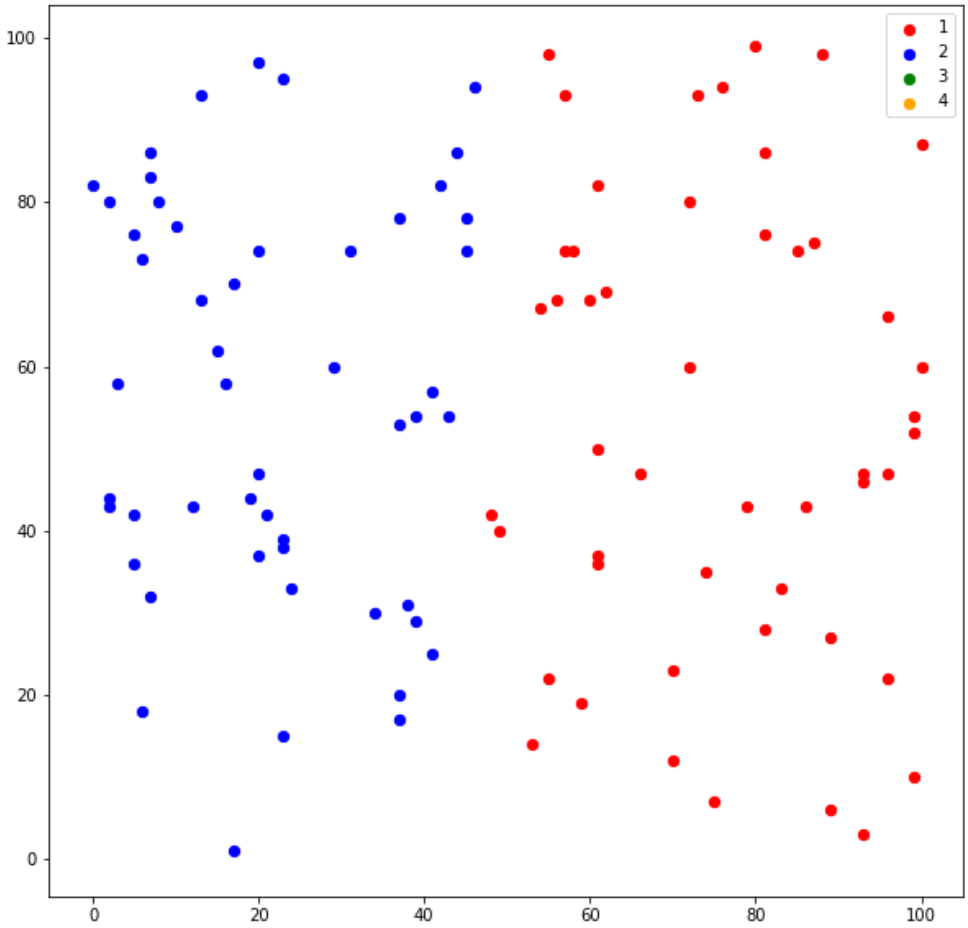
另外第一次创建之后，slave节点没有启动，导致mapreduce的任务卡住，等了很久都显示任务未完成，后来重新建了一次集群，只有slave2启动了，可以运行kmeans。但是隔了一天之后节点又挂掉，重新加载窗口后发现之前上传到hdfs的文件都没了，重新上传报错could only be replicated to 0 nodes instead of minReplication (=1). There are 0 datanode(s)。于是又建了一次集群……正常了。

可视化部分，使用python实现

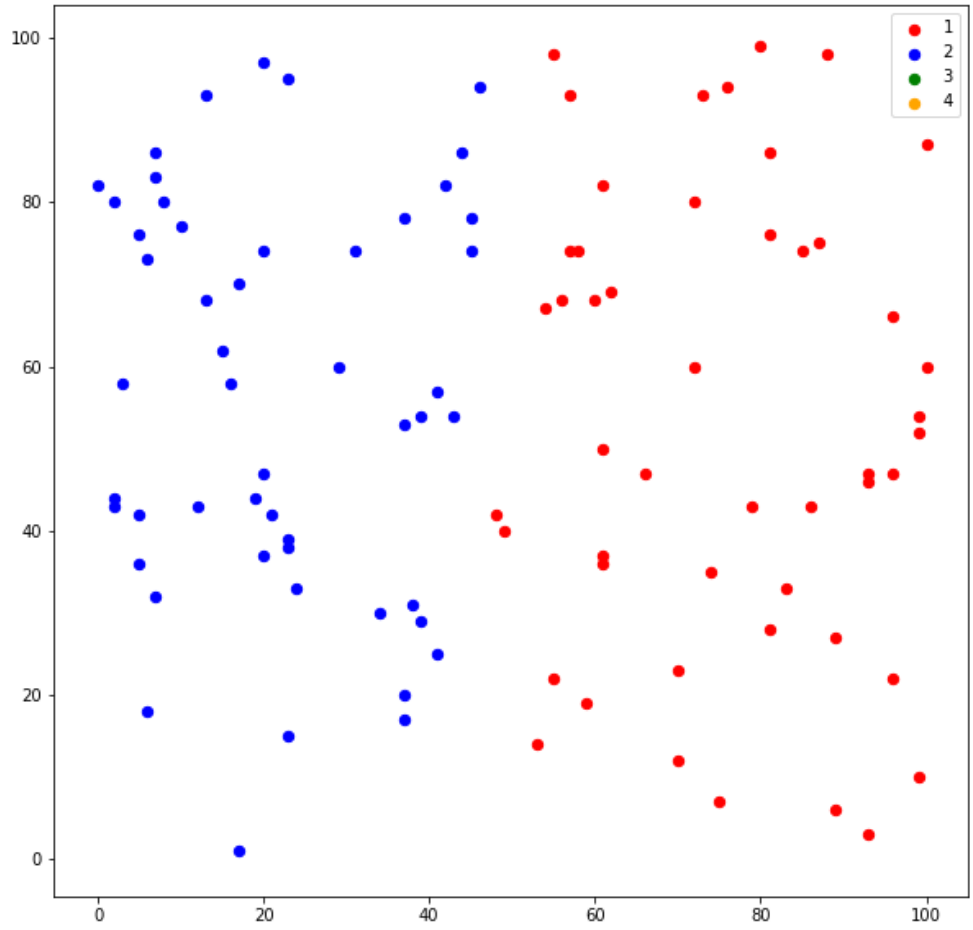
代码如下

```
import matplotlib.pyplot as plt
x0=[]
y0=[]
x=[[],[],[],[]]
y=[[],[],[],[]]
c=[]
with open("res2,20.txt") as f:
    s = [i[:-1].split(',') for i in f.readlines()]
    for i in range(0,len(s)):
        s[i][1]=s[i][1].split('\t')
        #x0.append(s[i][0])
        #y0.append(s[i][1][0])
        x[int(s[i][1][1])-1].append(int(s[i][0]))
        y[int(s[i][1][1])-1].append(int(s[i][1][0]))
color=['red','blue','green','orange']
plt.figure(figsize=(10,10))
#plt.scatter(x0,y0,color = 'red', s = 40 ,label = '1')
plt.scatter(x[0],y[0],color = 'red', s = 40 ,label = '1')
plt.scatter(x[1],y[1],color = 'blue', s = 40 ,label = '2')
plt.scatter(x[2],y[2],color = 'green', s = 40 ,label = '3')
plt.scatter(x[3],y[3],color = 'orange', s = 40 ,label = '4')
plt.legend(loc = 'best')    # 设置 图例所在的位置 使用推荐位置
plt.savefig('2,20.png')
plt.show()
```

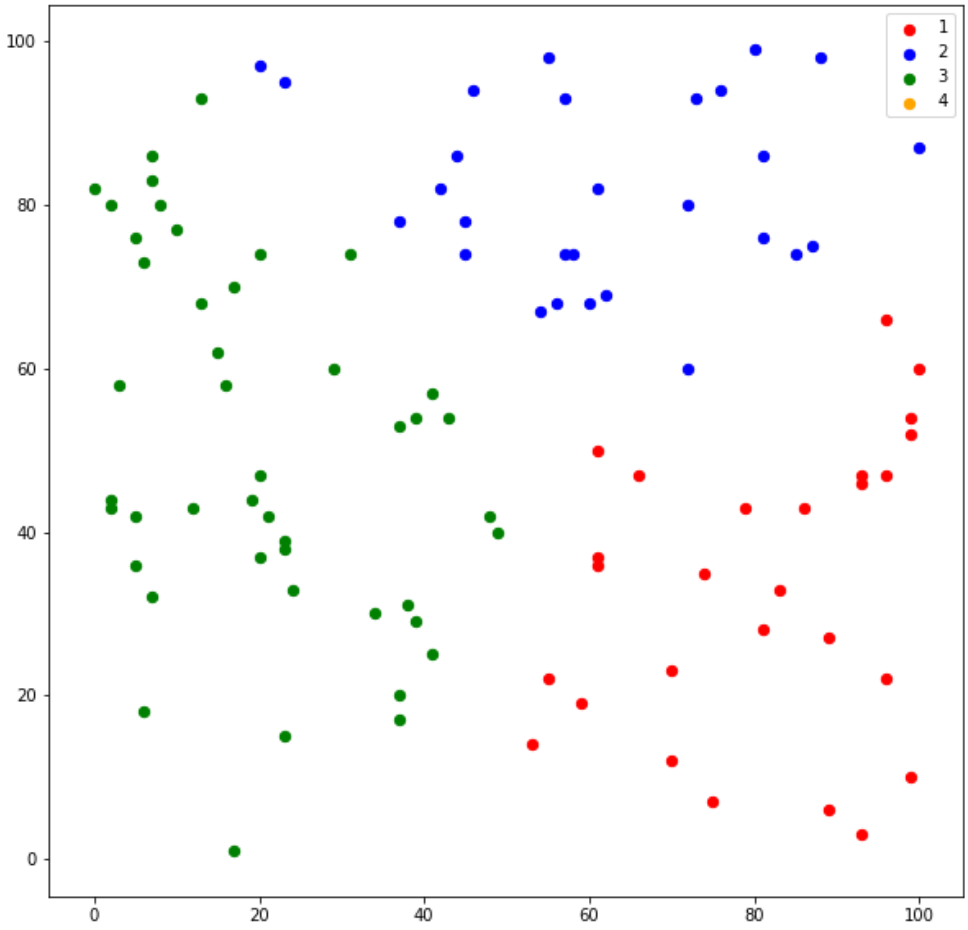
结果：



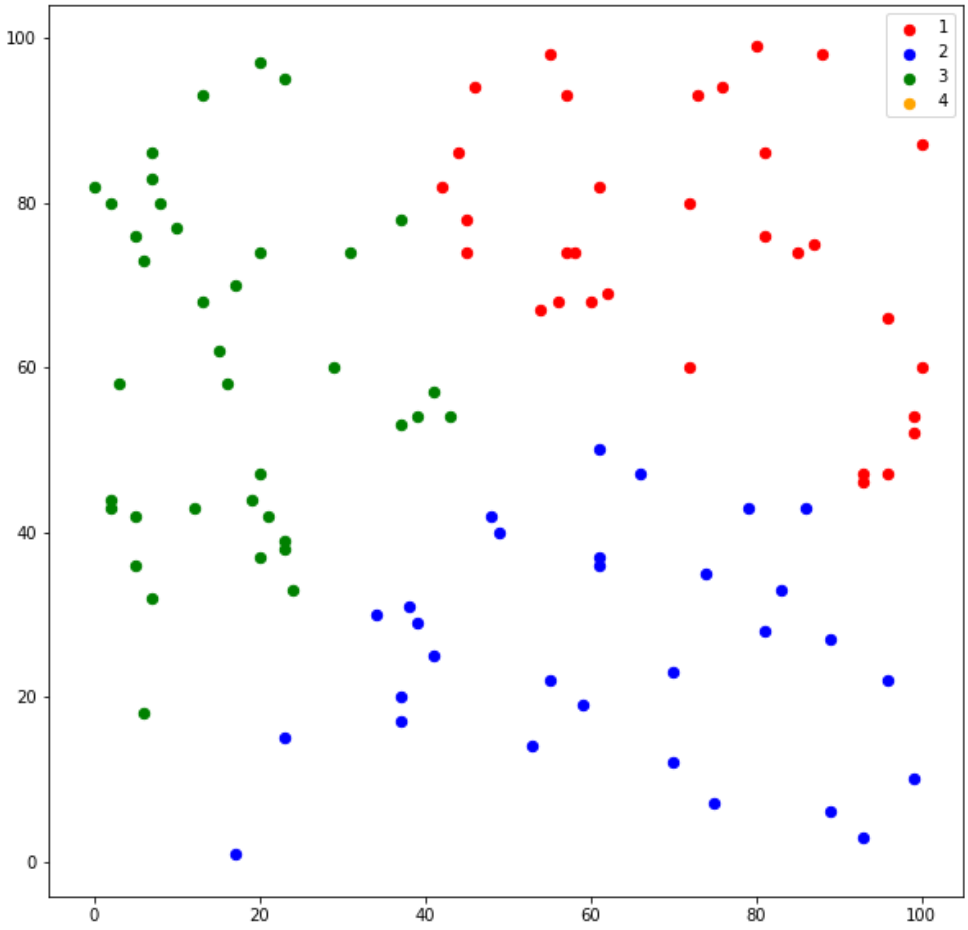
k=2，迭代次数10



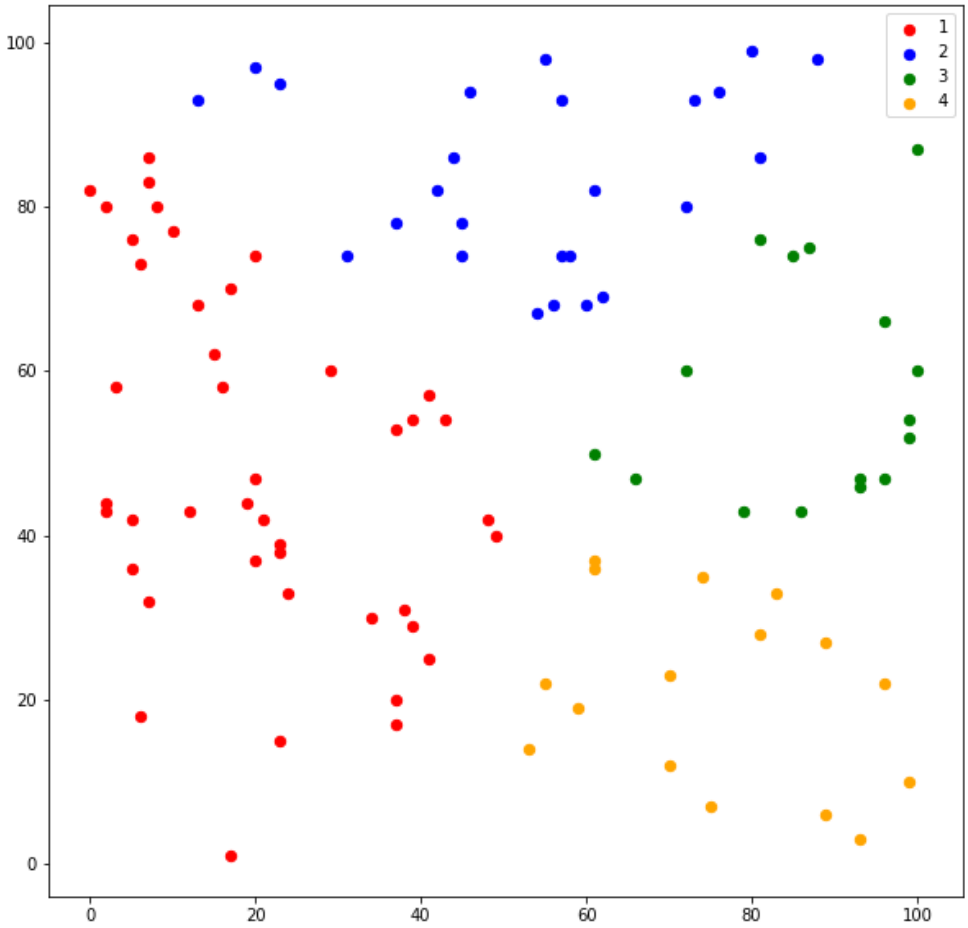
k=2，迭代次数20



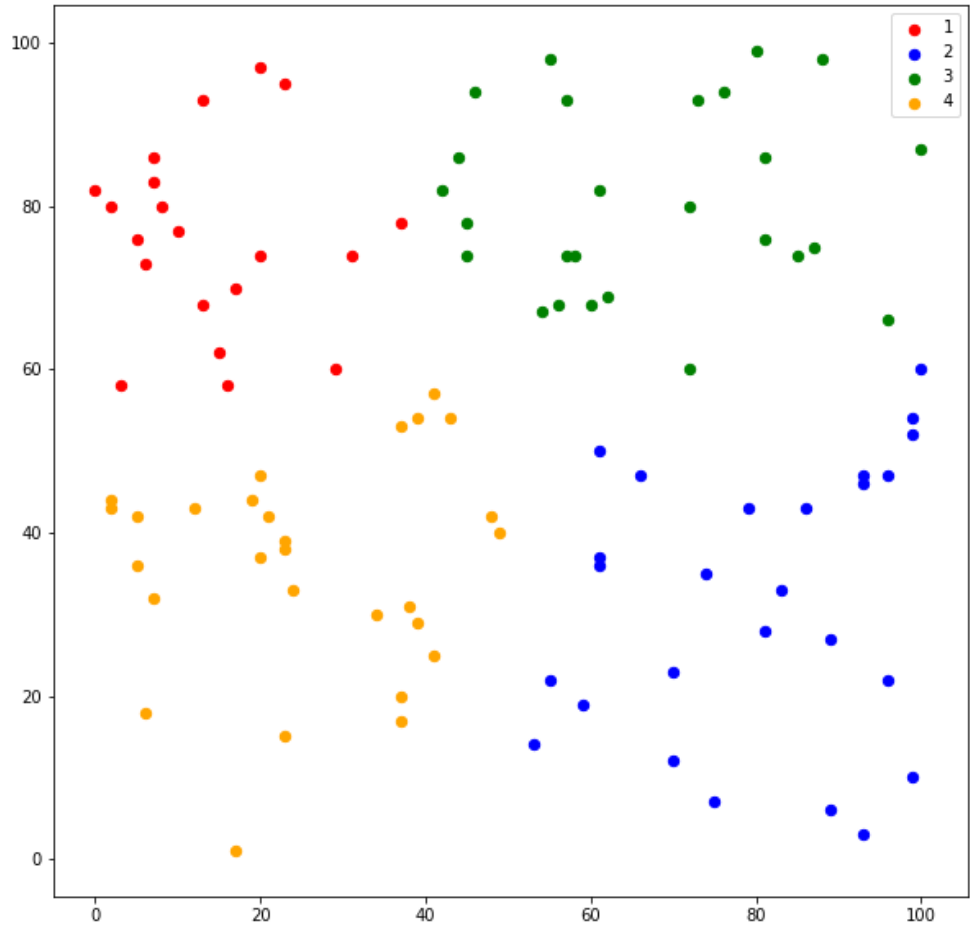
k=3，迭代次数10



k=3，迭代次数20



k=4，迭代次数10



k=4，迭代次数20

web截图