# Scipy.cluster.hierarchy.linkage

用这个工具可以进行层次化聚类，主要注意这样几个函数：

1. linkage

基本原理是用聚合的思想，先将每一个元素看做一个簇，然后不断合并距离最近的两个簇，直到只有一个簇为止，即对于一个有n个初始节点的数据集，会迭代（n-1）次。

函数原型：

scipy.cluster.hierarchy.linkage(y, method='single', metric='euclidean')

函数说明：

y是输入的 矩阵（其中矩阵类在pandas包中），这个矩阵是通过pdist函数计算出来的每个元素之间的距离；

Method 是设定聚类时所依据的距离算法，具体包括：

1. sigle 
2. complete 
3. average 

等（其中，u，v为当前某一个簇，i,j为簇中某一节点）

Metric则为距离的定义方式，具体包括：欧式距离，曼哈顿距离，切比雪夫距离，闵可夫斯基距离等。

函数返回值为一个 的矩阵，如下所示：

Z =  
 3.0000 4.0000 0.2228 100.0000  
 2.0000 5.0000 0.5401 101.0000  
 1.0000 7.0000 1.0267 102.0000  
 6.0000 9.0000 1.0581 103.0000

其中第i行第一，二列表示第i次迭代中合并的簇的序号，第三列表示这两个簇之间的距离，第四列表示合并后生成的簇序号，并且前两个簇会从数据集中删除。

# fcluster

函数功能是将有linkage函数返回的层次聚类结果转换成扁平的一维数组形式

函数原型：

scipy.cluster.hierarchy.fcluster(Z,t,criterion='inconsistent', depth=2, R=None, monocrit=None)

函数说明：

Z, 有linkage函数输出的矩阵

t，聚类阀值

criterion，聚类参数，可有如下取值：

inconsistent

如果群集节点和它的所有后代具有不一致的值小于或等于t则其所有的叶后代属于同一平面集群。当没有非单群集满足该标准，每个节点被分配给其自己的集群。

其他的看不怎么懂了。。。

Depth 在使用inconsistent时的最大深度，在其他模式下没有意义，默认为2

R，在用inconsistent参数运算时的‘inconsistent’矩阵，如果为空，则自动生成。

其他的函数主要用于数据可视化，检测两个聚类是否重构等。

1. 代码示例

From scipy.cluster.hierarchy

import linkage ,fcluster,fclusterdata

import pandas as pd

data = pd.read\_csv('data',sep='t',index\_col=0)

matData = data.as\_matrix()

distance = spatial.distance.pdist(matData)

linkresult = linkage(distance,method='average',metric='euclidean')

fcluster(linkresult,t=0.99,criterion='inconsistent',depth=2,R=None,monocrit=None)#这个需要先计算linkage，再出结果

dendrogram(linkresult,get\_leaves=False,show\_leaf\_counts=False)#这个可以绘制出树形图

参考链接：http://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html