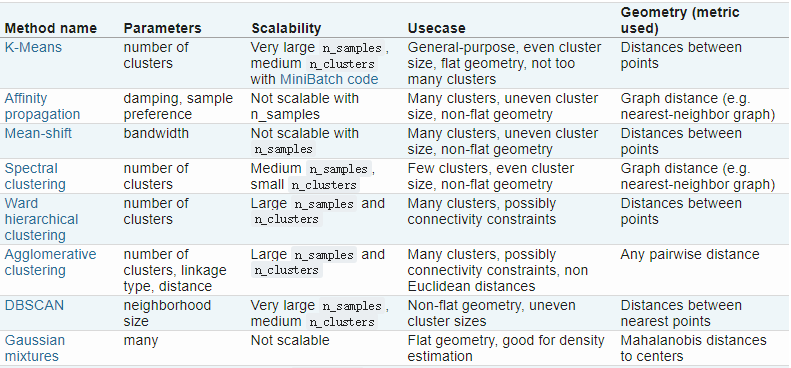
实验报告

1. Clustering with sklearn



* 1. K-Means

(1) 随机选择k个中心;

(2) 遍历所有样本，把样本划分到距离最近的一个中心;

(3) 划分之后就有K个簇，计算每个簇的平均值作为新的 质心;

(4) 重复步骤(2)，直到达到停止条件。

停止条件：

聚类中心不再发生变化；所有的距离最小；迭代次数达到

设定值。

代价函数：误差平方和（SSE）



1.2 Affinity propagation

 AP聚类算法是基于数据点间的"信息传递"的一种聚类算法。与k-均值算法或k中心点算法不同，AP算法不需要在运行算法之前确定聚类的个数。AP算法寻找的"examplars"即聚类中心点是数据集合中实际存在的点，作为每类的代表。

(1) 计算初始的相似度矩阵，将各点之间的吸引度 r(i,k) 和归属度 a(i,k) 初始化为 0；

(2) 更新各点之间的吸引度，随之更新各点之间的归属度

(3) 确定当前样本 i 的代表样本(exemplar)点 k，k就是使

{a(i,k)+r(i,k)}取得最大值的那个 k；

（4）重复步骤 2 和步骤 3，直到所有的样本的所属都不再变化为止。

1.3 Mean-shift

Mean-shift（即：均值迁移）的基本思想：在数据集中选定一个点，然后以这个点为圆心，r为半径，画一个圆(二维下是圆)，求出这个点到所有点的向量的平均值，而圆心与向量均值的和为新的圆心，然后迭代此过程，直到满足一点的条件结束。(Fukunage在1975年提出)。

Mean-shift 算法函数：

       a）核心函数：sklearn.cluster.MeanShift(核函数：RBF核函数)

           由上图可知，圆心(或种子)的确定和半径(或带宽)的选择，是影响算法效率的两个主要因素。所以在sklearn.cluster.MeanShift中重点说明了这两个参数的设定问题。

1.4 Spectral clustering

Spectral Clustering(SC,即谱聚类)，是一种基于图论的聚类方法,它能够识别任意形状的样本空间且收敛于全局最有解，其基本思想是利用样本数据的相似矩阵进行特征分解后得到的特征向量进行聚类.它与样本特征无关而只与样本个数有关。

     基本思路：将样本看作顶点,样本间的相似度看作带权的边,从而将聚类问题转为图分割问题:找到一种图分割的方法使得连接不同组的边的权重尽可能低(这意味着组间相似度要尽可能低),组内的边的权重尽可能高(这意味着组内相似度要尽可能高)

1.5 Ward Hierarchical Clustering

Hierarchical Clustering(层次聚类)：就是按照某种方法进行层次分类，直到满足某种条件为止。

1.6 Agglomerative clustering

Hierarchical Clustering(层次聚类)：就是按照某种方法进行层次分类，直到满足某种条件为止。

算法步骤:

          a）将每个对象归为一类, 共得到N类, 每类仅包含一个对象. 类与类之间的距离就是它们所包含的对象之间的距离.

          b）找到最接近的两个类并合并成一类, 于是总的类数少了一个.

          c）重新计算新的类与所有旧类之间的距离.

          d）重复第2步和第3步, 直到最后合并成一个类为止(此类包含了N个对象).

1.7 DBSCAN

（1）DBSCAN（Density-Based Spatial Clustering of Application with Noise）基于密度的空间聚类算法。

（2）两个参数：

* Eps邻域半径(epsilon,小量，小的值)
* MinPts(minimum number of points required to form a cluster

定义核心点时的阈值。

 （3）DBSCAN核心思想：从某个选定的核心点出发，不断向

密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。

1.8 Gaussian Mixtures

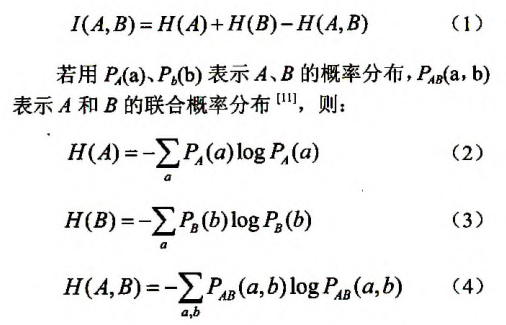
GaussianMixtureModel(混合高斯模型，GMM)。

        GMM的基本思想就是：任意形状的概率分布都可以用多个高斯分布函数去近似，也就是说GMM就是有多个单高斯密度分布（Gaussian）组成的，每个Gaussian叫一个"Component"，这些"Component"线性加成在一起就组成了 GMM 的概率密度函数，也就是下面的函数。

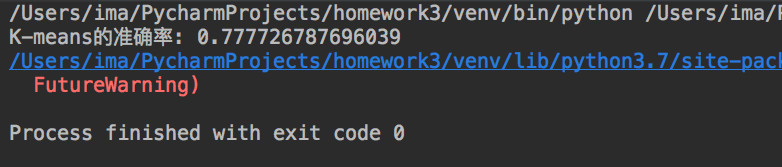
1. NMI(Normalized Mutual Information)

Normalized Mutual Information(NMI)常用在聚类中，度量2个聚

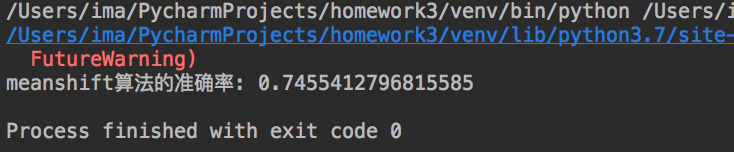
类结果的相近程度。

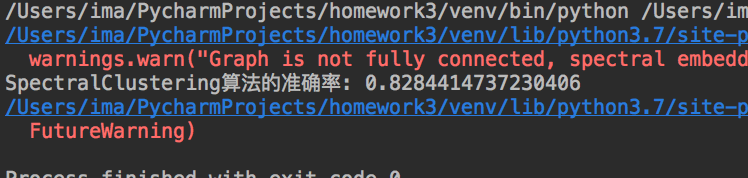


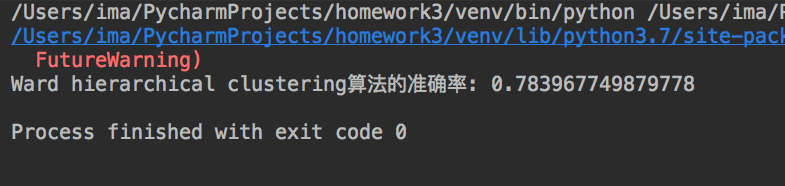
1. 测试在tweets数据集上的聚类效果
   1. k=89

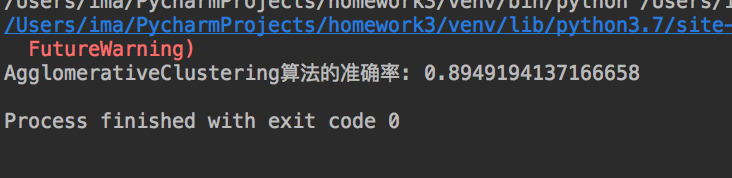


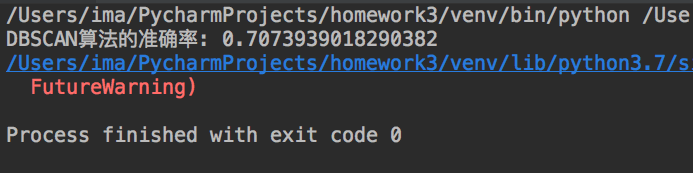


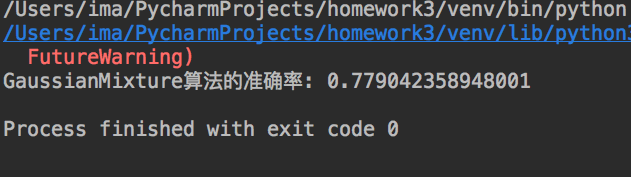




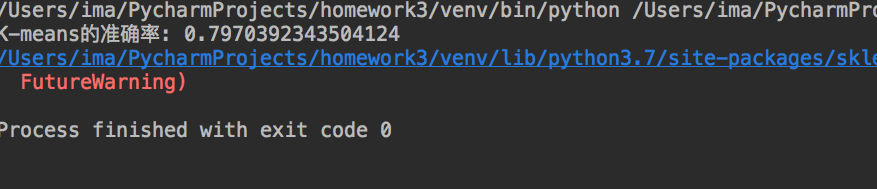




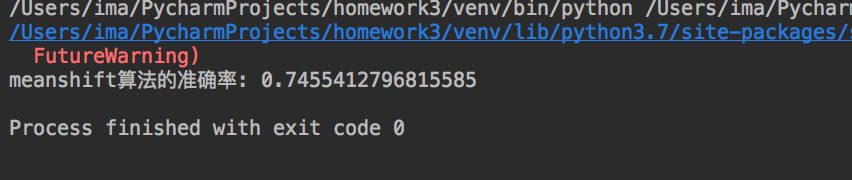


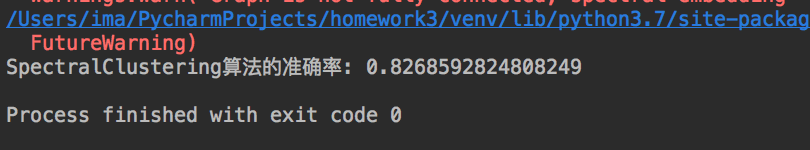


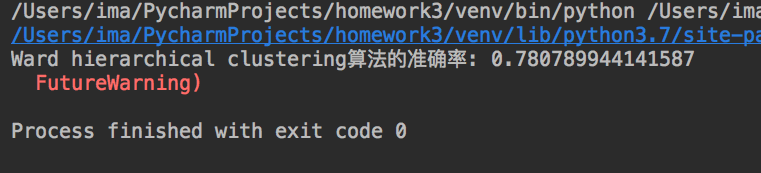
* 1. k=99



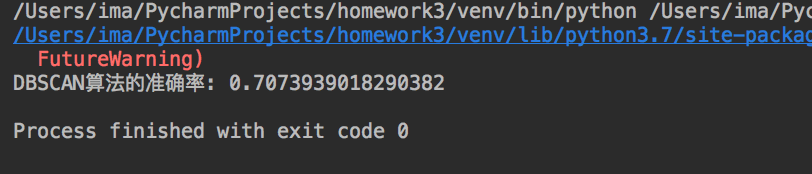


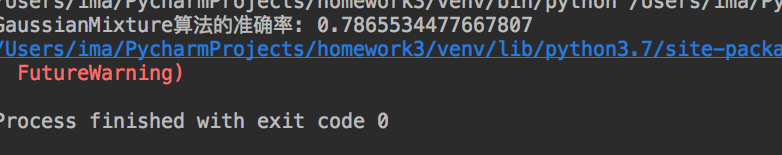




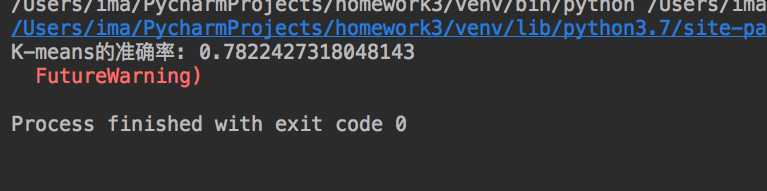


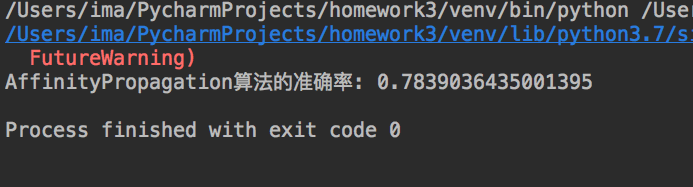


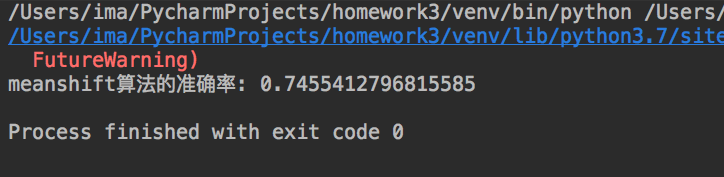


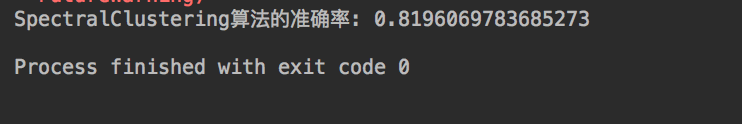


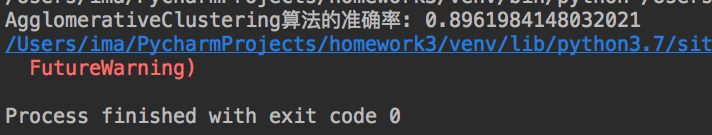
* 1. k=109



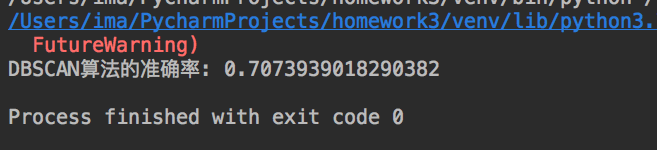


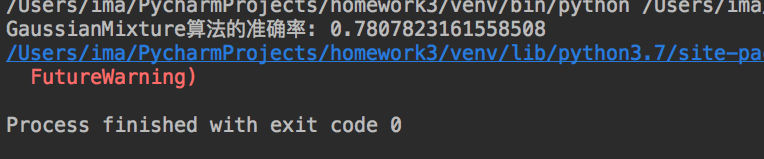












4. 实验总结

在使用sklearn中的聚类算法的相关函数进行聚类时，十分的方便快捷，节省了大量的时间。  
 在测试了各类算法的聚类效果后发现，基于本人这个测试，相较于K=89,K=109来说，K=99时整体都呈现出较好的效果，Agglomerative clustering算法呈现出最好的聚类效果，其次是Spectral clustering算法效果较好，而DBSCAN呈现出了最差的聚类效果。整体这八类算法的聚类效果都还不错。