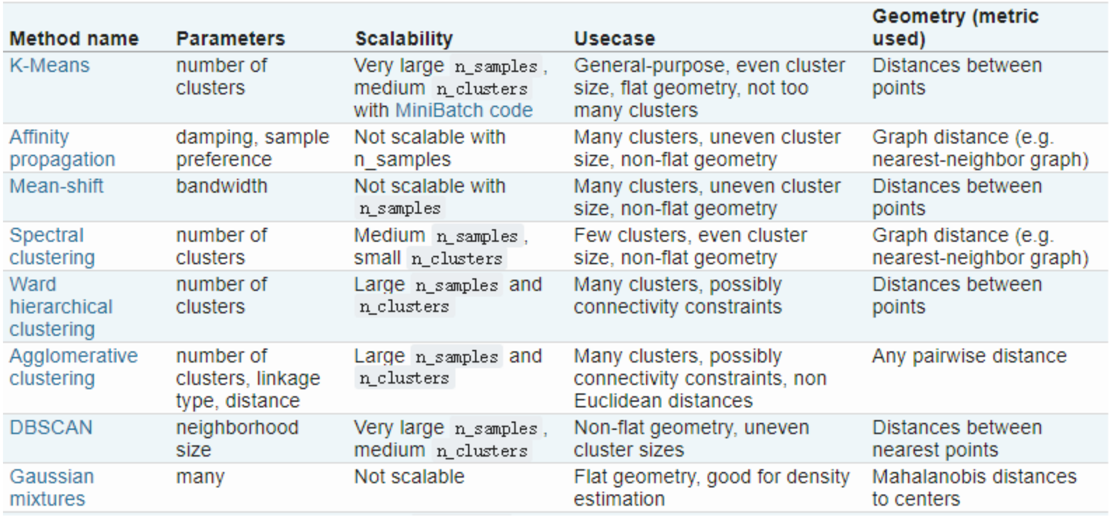
DATA Mining

-Homework2

Task：

1. 测试 sklearn 中以下聚类算法在 tweets 数据集上的聚类效果。
2. 使用 NMI(Normalized Mutual Information)作为评价指标。



Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large n_samples, medium n_clusters with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium n_samples, small n_clusters	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large n_samples and n_clusters	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large n_samples and n_clusters	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large n_samples, medium n_clusters	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers

Work：

利用 sklearn.metrics 中的 normalized_mutual_info_score()函数实现对以下聚类方法的评分。其中权重采用 TfIdf 的加权方式，聚类的方法均采用 sklearn 的标准聚类库

，在此基础上进行调参。具体评分如下：

K-means 的准确率：0.7905588364633279

AffinityPropagation 算法的准确率：0.7859274713522757

meanshift 算法的准确率：0.7468492000608158

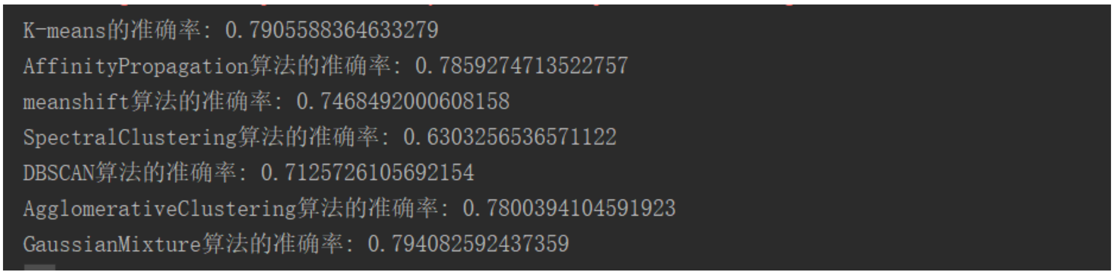
SpectralClustering 算法的准确率：0.6303256536571122

DBSCAN 算法的准确率：0.7125726105692154

AgglomerativeClustering 算法的准确率：0.7800394104591923

GaussianMixture 算法的准确率：0.794082592437359

截图：



```
K-means的准确率：0.7905588364633279
AffinityPropagation算法的准确率：0.7859274713522757
meanshift算法的准确率：0.7468492000608158
SpectralClustering算法的准确率：0.6303256536571122
DBSCAN算法的准确率：0.7125726105692154
AgglomerativeClustering算法的准确率：0.7800394104591923
GaussianMixture算法的准确率：0.794082592437359
```