实验报告

201844911　杨子玉

1.相关资料：https://scikit-learn.org/stable/modules/clustering.html#

实验任务：测试 sklearn 中以下聚类算法在 tweets 数据集上的聚类效果。

使用 NMI(Normalized Mutual Information)作为评价指标。

2.相关方法：

scikit-learn 简称 sklearn，支持包括分类、回归、降维和聚类四大机器学习算法。还包含了特征提取、数据处理和模型评估三大模块。

此次作业主要使用以下几种聚类方法：

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

3.处理文本数据集:

1)将实验数据的文本和应属于的类别放入两个向量中

2)调用库函数计算每个文本的 tf-idf 值

4.进行聚类：

调用函数聚类，同时采用 NMI(Normalized Mutual Information) 标准化互信息 评价效果

5.实验结果如下图所示

可以看到大多集中在 0.7 左右范围，AffinityPropagation 的效果最好。

```
In [48]: runfile('F:/anacodaa/123/Lib/site-packages/sklearn/
feature_extraction/untitled11.py', wdir='F:/anacodaa/123/Lib/site-
packages/sklearn/feature_extraction')
start cluster!
K-means: 0.7841980308246572
AffinityPropagation: 0.785654609647782
MeanShift: 0.7468492000608158
SpectralClustering: 0.6740829992908092
AgglomerativeClustering: 0.7843154591464184
DBSCAN: 0.7049439626810924
GaussianMixture:0.775646245521511
end
```