Homework3实验报告

2018年12月22日 15:33

- 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。

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

实验步骤

一：读取json文件构建词典

根据text与cluster分别读取，根据text建立词典，然后统计每个单词的词频建立向量空间模型，并将生成的向量写入文件保存

二：划分数据集

按照80%训练集、20%测试集将数据集进行随机划分然后写入文件保存

三：调用sklearn中的聚类算法测试聚类效果

以Affinity Propagation为例

```
from sklearn.cluster import AffinityPropagation
affinityPropagation = AffinityPropagation().fit(trainX)
predictTestY = affinityPropagation.predict(testX)
metrics.normalized_mutual_info_score(testY,predictTestY)#NMI
```

聚类算法	NMI
kmeans	0.7101455860352379
Affinity Propagation	0.7023334906752086
mean-shift	0.22569447521664332
Spectral clustering	0.5046971526915637
Ward hierarchical clustering	0.7544184153493245
Agglomerative Clustering	0.7544184153493245
DBSCAN	-1.8041124150158794e-06
Gaussian Mixture	

Gaussian Mixture算法在电脑上没有跑出结果，试了几次也没找到原因...