实验报告：Clustering with sklearn

姓名： 刘兵 学号： 201814819

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。

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

一、 工具平台

python3.7.0，Eclipse.Version: 2018-09 (4.9.0)

二、 实现方法

1.准备数据集

当前路径下的 Tweets.txt，通过 loadDataSet(dataFile)函数加载。

2.文件预处理。

（1）载入文档：

将文档以换行为划分，读为 list。通过 json 库读取文档中 text 键值对应的内容存为 corpus 作为文本数据集，读取 cluster 键值所对应的内容存为 clsNoLst 作为每个文本类别号。

（2）文本预处理，计算 TfIDF 值

通过 TextLemmatization 函数对单个文本进行处理词形还原。通过 Tfidf 函数进行去停用词、去字母小于 2 的词、平滑、正则化。计算出

词频稀疏矩阵。

3. Clustering with sklearn

（1）通过自定义函数Kmeans(tfidf_csr_mat, clsNoLst)计算出该方式聚类的NMI值。

（2）通过自定义函数affinityPropagation(tfidf_csr_mat, clsNoLst) 计算出该方式聚类的NMI值。

（3）通过自定义函数meanShift(tfidf_csr_mat, clsNoLst) 计算出该方式聚类的NMI值。

（4）通过自定义函数WardHierarchicalClustering(tfidf_csr_mat, clsNoLst) 计算出该方式聚类的NMI值。

（5）通过自定义函数spectralClustering(tfidf_csr_mat, clsNoLst) 计算出该方式聚类的NMI值。

（6）通过自定义函数agglomerativeClustering(tfidf_csr_mat, clsNoLst) 计算出该方式聚类的NMI值。

（7）通过自定义函数dBSCAN(tfidf_csr_mat, clsNoLst) 计算出该方式聚类的NMI值。

（8）通过自定义函数gaussianMixture(tfidf_csr_mat, clsNoLst) 计算出该方式聚类的NMI值。

三、 实验结果

```
**********Kmeans**********
NMI: 0.783920517820315
Kmeans耗时： 5.788652658462524
**********affinityPropagation**********
NMI: 0.7812722737896983
affinityPropagation耗时： 34.77854561805725
**********meanShift**********
NMI: -0.7265625
```

meanShift耗时： 601.7775073051453
**********WardHierarchicalClustering**********
NMI: 0.7741746117936851
WardHierarchicalClustering耗时： 12.872848987579346
**********spectralClustering**********
NMI: 0.6714722557858377
spectralClustering耗时： 5.748222351074219
**********agglomerativeClustering**********
NMI: 0.7741746117936851
agglomerativeClustering耗时： 13.191979885101318
**********dBSCAN**********
NMI: 0.5683136787031786
dBSCAN耗时： 63.17505955696106
**********gaussianMixture**********
NMI: 0.8095986364758593
gaussianMixture耗时： 8.236756324768066
总耗时： 746.6577022075653 秒
结束...

四、调试总结

通过用八种方法对文档进行聚类处理，不只了解了这些方法效果上的差异，更重要的收获是学会了如何利用官方文档（sklearn），如何从中查找资料，查看用法，学会了查看官方文档说明，在将来遇到新的问题能够自行解决，这是最大的收获。