DataMining

聚类方法对比

姓名：董潇
学号：201834861
班级：2018 级计算机专硕班
指导老师：尹建华
时间：2018 年 12 月 26 日

一、 实验要求

用 scikit-learn 中的各种聚类方法对 Tweets 数据集进行聚类。

二、 实验内容

1. 准备好实验所用数据集：Tweets.txt
2. 对数据集进行向量化处理，方便作为输入传入要调用的函数中。
3. 调用所需的聚类函数，需要注意的是参数的设置。
4. 输出各个聚类函数的正确率进行比较。

三、 实验结果

```
In [8]: runfile('D:/DataMining/HomeWork3/cluster.py', wdir='D:/DataMining/HomeWork3')
K-means的准确率: 0.7910405358188654
AffinityPropagation算法的准确率: 0.7842866176251021
meanshift算法的准确率: 0.7455412796815585
D:\Anaconda3\lib\site-packages\sklearn\manifold\spectral_embedding_.py:234: UserWarning: Graph is not
fully connected, spectral embedding may not work as expected.
  warnings.warn("Graph is not fully connected, spectral embedding"
SpectralClustering算法的准确率: 0.6277834214648518
DBSCAN算法的准确率: 0.7073939018290382
AgglomerativeClustering算法的准确率: 0.783967749879778
GaussianMixture算法的准确率: 0.779042358948001
```

实验结果截图

四、 实验总结

通过本次实验掌握了 sklearn 库里面这 7 种聚类方法的调用与使用，收获很多。