# 实验报告三

**——sklean各种聚类方法对比**

201814814 赖文锋

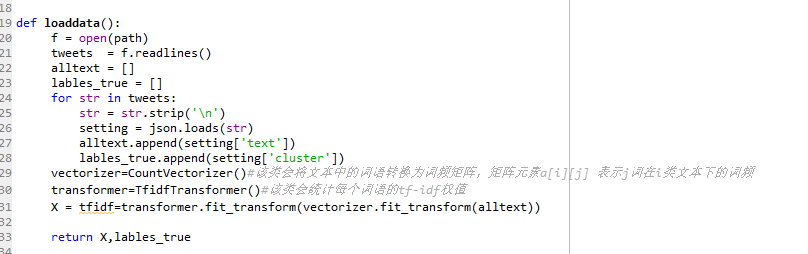
1. 概述

这是实验三的实验报告，该实验的主要内容是使用scikitlearn里自带的各种方法进行聚类，并且比较实验结果。其中包括七种聚类方法。

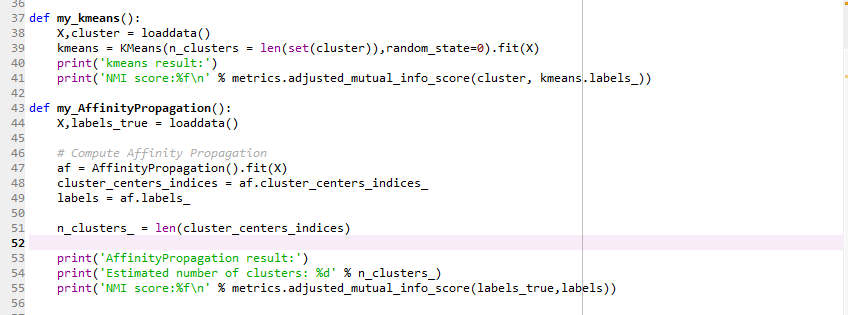
1. 主要思想及理论

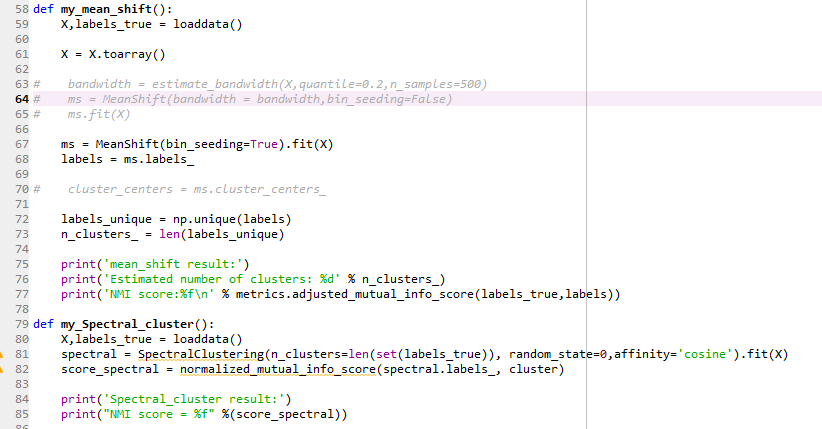
本着学习使用scikitlearn的态度，本次实验包括预处理全部使用了框架自带的方法。但从与同学之间的对比来看，其生成VSM的方法并不如自己使用NLTK编写的效果好。程序主要分为两部分，第一部分是预处理，首先读取tweets.txt，这个文件每一行是一个json格式的数据，可以用python自带的库解析，然后提取出文本的内容和标签存下。然后调用scikitlearn的tf-idf方法把文档列表转化为VSM。第二步就是调用各种聚类方法对文档进行聚类，并使用NMI作为评价指标，其中设置参数较为繁琐不一一赘述，

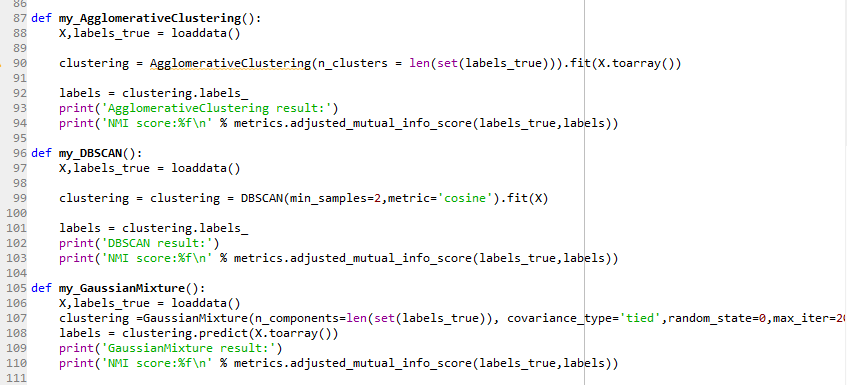
1. 主要方法介绍
2. 预处理方法，调用库里的方法把文档转化为VSM：



1. 各种聚类方法的使用：

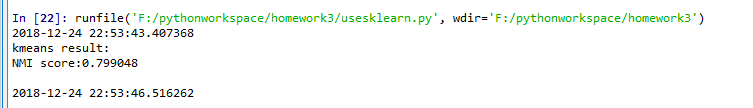






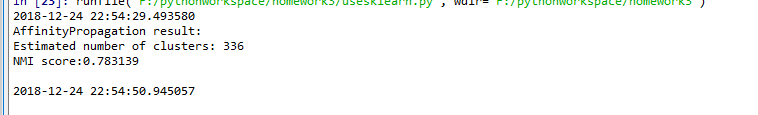
1. 结果展示

1.kmeans：



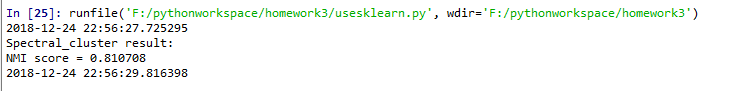
执行时间极为迅速，包括预处理竟然只用时3秒钟，可谓效率极高，评分上有接近0.80，效果在各个聚类算法里也位列前茅，加上算法易于理解，可谓性价比极高。另外观察发现，参数里能指定聚类中心个数的算法往往表现较好。

2.AffinityPropagation：



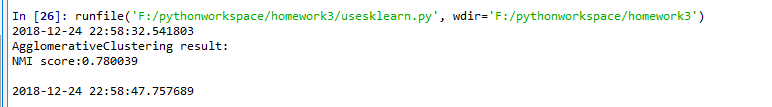
执行时间20秒，速度不错；但是由于不需要指定聚类的类的个数，导致真实分类中心不足100个，算法缺给出了336个，因此分数依然有0.78，在这个数据集上表现稍微劣于kmeans。考虑到不需要指定聚类的个数，其实是一个不错的选择。

1. Spectral\_cluster()：



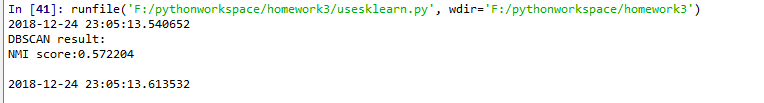
运算时间快至的2.1s，分数也达到了0.81，超过了kmeans，在这个数据集上的表现可谓是全方位超越kmeans，十分亮眼。

4. AgglomerativeClustering()



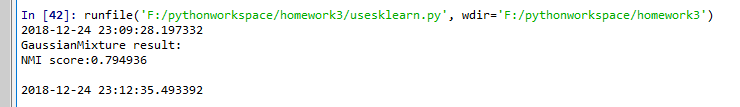
运行时间15秒，分数0.78，需要指定聚类个数，总而言之，表现十分中规中矩的一个算法。

5.DBSCAN



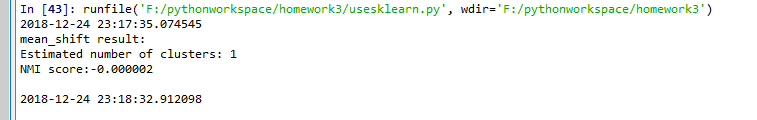
速度可谓无与伦比，达到了0.07秒，是别的算法的几百上千倍速度，但可惜的是预测结果不尽人意，只有0.57的低分，由此可见这个算法适用于数据量极大，但对精度要求不那么高的场合中。

6. GaussianMixture：



运行时间是最慢的，用时3分07秒，然而精度也不见得是最高的，分数0.79，属于好但是依然有更好的。推测在更大的数据集上有更好的结果，但这意味着要花出更多的时间成本。

7. mean\_shift：



很令人费解的一个算法，我调整了很多次参数都无法获得更好的结果，暂时不清楚问题出在哪，有可能是数据集预处理不好导致的，该算法把所有的文档分到了一个类中，分数只有-0.000002，耗时1分钟，真可谓陪了夫人又折兵……

1. 结果分析与总结。

通过这个实验，我熟悉了scikitlearn的基本用法，对于其

语法和简便性有了体会。并且对各个算法的性能和执行时间也有了直观的认识。感谢教导我的老师和与我交流的同学们。让我能顺利完成此次实验。