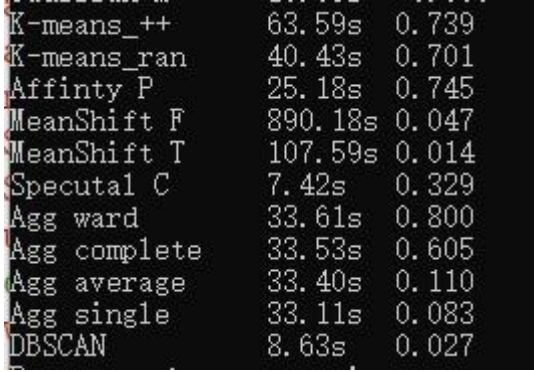


信息检索与数据挖掘 课程实验报告

学号：201600130032	姓名： 陆禹承	班级： 智能 16
实验题目： 聚类算法实现		
实验内容： 分析比较不同聚类算法在文本分类上的效果		
实验过程中遇到和解决的问题： 使用 python 建立单词表，并将每个文档处理成一个向量，表示每个词在文档中出现的频率，然后调用 sklearn 的方法进行处理。 文本数据特点是类别多，每个类别的文档较少，一般 1~10 个 (部分运行结果截图)  经过多次测试， MeanShift 因为文本维度太多不适用于文本聚类，直接 PASS K-Means 表现不错，需要指定类别数，初始化方式影响较大，多次测试，初始化会对分数有 15%浮动的影响 对于 AgglomerativeClustering，显然 ward 方式（组内 L2 最小，组间 L2 最大）最好 linkage（两类间的距离=最远点距离）还行 其他距离函数效果不太好。 需要提前指定类别数 AffinityPropagation 比较出色，其特点在于不需要指定类别数，且初始值影响不敏感。需要指定参考度。		

DBSCAN 需要较高的 `eps` 参数值（因为太稀疏了），不过相应的时间会大幅度增加，测试效果没有前面的 AP,AC,K-MEANS 好，达不到同等效果，最高得分 0.6，时间大约在 80 秒，`eps` 增加分类效果并不是正关系，`eps` 太高会降低效果。总之比其他算法效果差有点多

GM（高斯混合）没有跑出结果，也许因为数据太稀疏了。

综上，对于原始的 VSM

AgglomerativeClustering+ward 方式比较好，并且调参少

AffinityPropagation 适合在类别数无法确定时使用

对于经过 tf-idf 处理的数据，处理结果：

Agg ward	35.76s	0.775
Agg complete	25.65s	0.762
Agg average	23.68s	0.900
Agg single	23.60s	0.130
Specutal C1	4.67s	0.676
Specutal C1.5	5.30s	0.615
Specutal C0.5	6.63s	0.726
DBSCAN2	0.29s	-0.000
DBSCAN1.5	0.29s	-0.000
DBSCAN1	0.12s	0.425
DBSCAN0.5	0.13s	0.036
K-means_++	9.84s	0.784
K-means_ran	8.60s	0.754
Affinty P	40.33s	0.769

SpectralClustering 的优势在这里被很好地体现出来，速度非常快，经过调参后分数可以达到较高的水平，DBSCAN 非常快但是分数依旧比较低

结论分析与体会：

对不同的聚合算法和评价函数有了大体了解。