实验报告

一、实验步骤

首先处理给定文本，把其中”text”对应的内容放到 Data 这个列表中，同时把”cluster”对应的数据提取出来，放到 Labels 这个列表中，然后，使用 TfidfVectorizer 把 Data 中的元素转为向量形式表示，最后依次调用 sklearn 中的聚类方法，用 NMI 评价聚类结果。

表 1.1

聚类方法	评价指标 NMI
K-Means	0.7830779694677665
Affinity propagation	0.7836988975391975
Mean-shift	-0.7265625
Spectral clustering	0.6528537755966615
Ward hierarchical clustering	0.7823244114906182
Agglomerative	0.6198448865824127
DBSCAN	0.8962440814686098

二、遇到的问题

这些聚类方法大部分只需要指定分类的数量 n_cluster 这个参数，其余使用默认值就可以。但是 DBSCAN 使用默认参数值分类效果就很差，默认值为：eps=0.5，min_simples=5，将这两个参数分别调整为 1.13，6 后效果最佳。