实验五 多种聚类方法对twitter数据集进行聚类

数据集：

老师发布的tweets的json文件的数据集，共109个分类标签

实验方法：

1. 先将数据集读入，分别将tweet的正文和分类标签存入列表
2. 将数据集向量化，处理成tf-idf加权的形式
3. 分别使用sklearn的库函数进行聚类，分别是：kmeans, ，affinity propagation，mean-shift，spectral，ward，agglomerative，DBSCAN和Gauss-mixture方法
4. 将聚类得出的分类标签和真实结果对比进行NMI检验，并求出得分

结果展示：

由于Gauss-Mixture方法的速度太慢，所以减少了cluster的数量至10，但是也导致结果很不准确。

|  |  |
| --- | --- |
| clustering method | NMI score |
| Kmeans | 0.7869117087548917 |
| affinity propagation | 0.6946841230646935 |
| mean-shift | 0.6796756366401875 |
| spectral | 0.6796756366401875 |
| ward | 0.7777727544931091 |
| agglomerative | 0.8999329168527519 |
| DBscan | 0.7009526046894612 |
| gauss-mixture | 0.1253467393244675 |