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