实验报告

Clustering

一、 向量空间模型

实验步骤;

1. 读取 Tweets.json，并将结果单词向量存放在 vectors 中，将标签存放在 label 中。
2. 生成词汇表：遍历所有文档向量，即 vectors。将所有单词无重复的存入词汇表 wordtable 中。由于单词数比较少，所以不进行词频过滤。
3. 计算 tf-idf：循环遍历所有文档，对于其中的一篇文档：tf：首先，计算这篇文档的单词在这篇文档中出现的次数作为 tf；idf：首先计算每个单词出现在不同文档中的文档数，并利用公式 idf=log((N+1)/(df+1))表示 idf。其中 N 为文档总数。
4. 将计算好的 tf-idf 向量保存在文件中，并把类别号保存在 tf-idf 向量的最后一维。

二、 聚类

实验步骤:

1. 首先加载之前 vsm 计算好的 tf-idf 向量，将其保存在 vectors 中，类别保存在 label 中。
2. 导入 sklearn 中的聚类算法，包括 k-means、AffinityPropagation、DBSCAN 等。
3. 对每个聚类算法，若需要指定 k 个类，则 k 取数据集类别总数。对每个聚类算法进行适配、预测样本标签，计算 NMI 并返回 NMI 最终结果。

三、 实验结果

实验利用 sklearn 写好的几个聚类算法对数据集进行聚类，并计算 NMI:

```
K-Means: 0.662567540573726
AffinityPropagation: 0.6594122330661669
Mean-shift: 0.21091318304744194
Spectral clustering: 0.050504861582973 83
Ward hierarchical clustering: 0.7977523619202832
DBSCAN: 0.09149604705084 69
Gaussian mixtures: 0.7154829884491714
```