

DATA Mining

-Homework1

Task :

- 1.预处理文本数据集，并且得到每个文本的 VSM 表示。
- 2.实现 KNN 分类器，测试其在 20Newsgroups 上的效果。
- 3.20%为测试集，其余的为训练集，保证各个文档的均匀性

处理步骤：

1. data download

下载地址：<http://qwone.com/~jason/20Newsgroups/>

2. data process

- 1) `pre_process()` 对源文件进行预处理
- 2) `at_dict()` 生成目标词典
- 3) `VSM()` 对照词典生成特征词

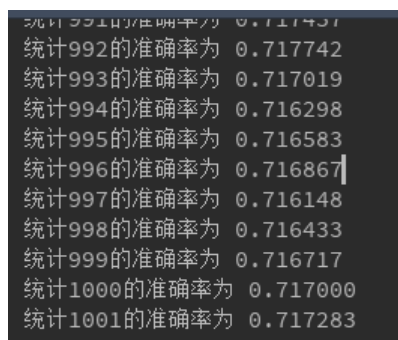
3. KNN

- 1) `IDF()` 计算特征词 IDf 值
- 2) `TF_IDF()` 计算 TF_IDf 值
- 3) `Knn()` 利用 Knn 算法计算文件文件相似度

结果展示：

在 K=20 的情况下的 acc 为（设备有限，只测试了前 test 数据集的前 1000 文档）：0.71

如下截图：



A screenshot of a terminal window showing accuracy results for K=20. The text is as follows:

```
统计992的准确率为 0.717742  
统计993的准确率为 0.717019  
统计994的准确率为 0.716298  
统计995的准确率为 0.716583  
统计996的准确率为 0.716867  
统计997的准确率为 0.716148  
统计998的准确率为 0.716433  
统计999的准确率为 0.716717  
统计1000的准确率为 0.717000  
统计1001的准确率为 0.717283
```

