Homework1--VSM/KNN（2018.11.3）

姓名：陈潇琳 学号：201834857

任务

1. 对原始数据进行预处理，计算单词的 TF-IDF，得到每个文档的 VSM 表示。
2. 实现 KNN 分类器，测试其在 20Newsgroups 数据集上的效果。

流程

1. **预处理**

(1) Tokenization。把文档中的文本进行分句、分词。

(2) 过滤单词：a) 去标点符号;

b) 去停用词；

c) 使用 stem 寻找词根；

经过过滤，构建大小为 21686 的词库。

2. **计算每个单词的 TF 和 IDF**

以单词命名建立 json 文件，以字典格式存放在此单词在每个文件中的 TF 以及 IDF。

3. **对每个文档构建与字典同等大小的向量空间表示。**
4. **KNN 分类**

使用 cosine 计算距离，按分数从大到小排序，选取前 k 个 (k=5)，这 k 个文档所属类别占多数的那个类别就是这个 test 文档的类别。

实验结果

最终结果可以达到 0.75

总结

1. 本次作业 11 月初才开始写，使得整个过程比较被动。下次作业一定要早做。
2. 读写操作比较耗时，还需要进一步改进。
3. KNN 中 k 的选取，并不是越大越好。