VSM+KNN 实验报告

姓名：周延　　学号：201814853

本实验使用 python 版本为 python3.6 操作系统为 ubuntu16.04

实验使用数据文件夹绝对路径为 pwd/../20news-18828

本实验操作分别在三个文件：readfile.py vsm.py knn.py

1.readfile.py 用于将所有数据绝对路径存放于列表 listfile 中

2.vsm.py 用于计算文档的 vsm，用字典 file_word_dict 表示，最后存储在 vsm.json 文件中

3.knn.py 用于通过 KNN 分类器计算测试集属于哪个类，计算正确率

VSM：

1.通过函数 get_word_nubmber 用来获取每个文件中单词数目存入字典 file_word_dict，该函数会将单词小写化，使用 nltk 库进行词性还原，去掉不属于单词的符号，file_word_dict 结构为{file-{word-numbers}}，结果写入 file_word_dict.json 中；

2.通过 file_word_dict 计算出所有文件内单词和单词出现个数存入字典 word_numbers，再对 file_word_dict 进行一些处理去掉不再 10-1000 之间的单词；

3.通过函数 com_VSM 计算出每个文件内每个单词打 TF-IDF，同样存入 vsm-word 中，结构为{file-{word-tf*idf}},结果写入 vsm.json 文件中。

0.7285145712449033, "atheist": 3.8820950628977595, "resource":
1.8686841678952038, "december": 0.9589888728156154, "version":
1.3728862691556472, "address": 1.138030192584587,
"organization": 1.0044581010149467, "usa": 1.6211008528363156,
"freedom": 1.0995315351758075, "religion": 2.19392220374364,
"foundation": 1.3645203888246091, "darwin": 3.1982551765274403,
"fish": 3.448801988011216, "bumper": 1.1427472111460681,
"sticker": 1.062974980489618, "assorted": 2.572509598511201,
"write": 2.0826713119920086, "box": 1.1329558273363958,
"madison": 1.205852639800365, "wi": 1.0954569621369095,
"telephone": 2.0263097030550057, "evolution":
1.9688853754246372, "design": 1.3507730278270795, "sell":
0.9307068720825785, "symbol": 1.055001782660802, "stick":
0.7833298389235881, "foot": 0.7518293105034667, "written":
1.0044581010149467, "inside": 0.7154370502426797, "deluxe":
1.2177315076268185, "plastic": 0.9621556127878057, "laurel":
1.4145889746473 06, "canyon": 1.2663588487128865, "north":
0.7643733197120476, "hollywood": 1.3107456704373726, "bay":
0.8977702521198128, "area": 0.8269825463666388, "gold":
0.982351597931234, "mailing": 0.8361297114798554, "net":
0.8932664644869521, "directly": 0.6897939182862077, "price":
0.5625259569894053, "american": 2.2419503380774444, "press":
2.935355754742116, "publish": 2.60058500382752, "various":
1.3182006885175102, "critique": 1.8438010820598743, "bible":
2.6187657633603965, "biblical": 0.8683191603518844,
"contradiction": 1.4780506190948952, "handbook":

KNN:

1.将所有数据分为 20%测试数据和 80%训练数据

2.如计算 vsm 时所做，分别计算出 test_data 和 train_data 的 vsm，使用 tf-idf 表示

3.对于每一个测试数据遍历训练集计算距离 cos 值

4.排序后取出 K 个最大的训练集文档，K 个文档中类最多的即为测试文档的类

5.验证分类是否正确，计算完所有测试数据后再计算分类成功率

6.以上步骤重复 5 次

7.调参，改变参数 K，找出成功率最大的 K

结果:

K 统计 3-8 之间打参数，平均成功率分别为:

K=3　成功率 70.34%

K=4　成功率 74.91%

K=5　成功率 77.23%

K=6　成功率 78.42%

K=7　成功率 78.65%

K=8　成功率 78.72%

K=6 后趋于稳定　K=6 比较好