Data Mining Homework1 总结报告

一、VSM

1、数据预处理

将文档分析和小写化，去掉数字、单字母项，使用 NLTK 库进行词性还原并去除非单词项，删除停用词。

2、统计词频

对 18828 个文档分别统计词频，结果存放在字典类型中并以 josn 文件格式保存于“word_dict.json”文件中。


```
{".\\\\20news-18828\\\\alt.atheism\\49960": {"subject": 2, "faq": 1, "atheist": 10,
"resource": 4, "december": 1, "version": 4, "address": 3, "organization": 2, "usa":
4, "freedom": 2, "religion": 6, "foundation": 2, "darwin": 4, "fish": 6, "bumper":
1, "sticker": 1, "assorted": 3, "paraphernalia": 1, "available": 2, "write": 6,
"box": 3, "madison": 1, "wi": 1, "telephone": 4, "evolution": 3, "design": 3,
"sell": 2, "symbol": 1, "like": 1, "one": 4, "christian": 6, "stick": 1, "car": 1,
"foot": 1, "word": 1, "written": 2, "inside": 1, "deluxe": 1, "moulded": 1,
"plastic": 1, "postpaid": 1, "laurel": 1, "canyon": 1, "north": 1, "hollywood": 1,
"ca": 1, "people": 6, "bay": 1, "area": 2, "get": 3, "gold": 1, "try": 2, "mailing":
1, "net": 2, "go": 2, "directly": 1, "price": 1, "american": 7, "press": 8,
"publish": 4, "various": 3, "book": 15, "critique": 2, "bible": 7, "list": 1,
"biblical": 1, "contradiction": 2, "handbook": 1, "ball": 2, "edition": 3,
"absurdity": 1, "atrocity": 1, "immorality": 2, "contains": 3, "contradicts": 1,
"based": 2, "king": 1, "james": 3, "austin": 2, "tx": 2, "road": 2, "fax": 2,
"prometheus": 5, "including": 2, "holy": 2, "horror": 2, "see": 2, "east": 1,
"street": 2, "buffalo": 3, "new": 4, "york": 2, "alternate": 1, "may": 1, "newer":
1, "older": 1, "glenn": 1, "drive": 1, "ny": 2, "african-american": 1, "humanism":
5, "promoting": 1, "black": 2, "secular": 3, "uncovering": 1, "history": 6,
"quarterly": 1, "newsletter": 1, "aah": 1, "examiner": 1, "norm": 2, "allen": 2,
"jr.": 2, "african": 3, "united": 1, "kingdom": 1, "rationalist": 1, "association":
2, "national": 2, "society": 3, "high": 1, "london": 4, "british": 1, "humanist": 1,
"south": 1, "place": 1, "ethical": 1, "lamb": 1, "conduit": 1, "passage": 1, "hall":
1, "red": 1, "lion": 1, "square": 1, "freethinker": 1, "monthly": 1, "magazine": 1,
"founded": 1, "germany": 4, "internationaler": 2, "der": 3, "berlin": 2, "journal":
2, "hannover": 1, "fiction": 1, "thomas": 1, "santa": 2, "compromise": 1, "short":
```


3、计算 VSM

在第 2 步中统计词频基础上计算 VSM，利用 Maximum TF scaling 公式计算 TF，α 值定为 0.1，公式如下：

$$tf(t,d) = \alpha + (1-\alpha)\frac{c(t,d)}{\max_t c(t,d)}$$

计算 IDF：

$$IDF(t) = \log(\frac{N}{df(t)})$$

结果以 josn 文件格式保存于“ALL_files_VSM.json”文件中。

{".\\\\20news-18828\\\\alt.atheism\\49960": {"faq": 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": 1.414588974647306, "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": 1.1122821076317841, "ball": 1.1498278694117157, "edition": 1.7929786076400596, "absurdity": 1.1269047054476484, "atrocity": 1.0576225138659459, "immorality": 2.2112132709704015, "contains": 1.5505261522333815, "contradicts": 1.1741123637143636, "based": 0.8467693744287433, "king": 0.7555601456616698, "james": 1.2781794288808992, "austin": 1.4019091787594755, "tx": 1.3180445503318161, "road": 0.9503790823447692, "fax": 0.8356699029643069, "including": 0.889676671726815, "holy": 1.268370438574189,

二、KNN

1、基本步骤

STEP1：随机将 20 类文档每类分出 20%作为分类测试集，其余为训练集。

STEP2：分别计算训练集和测试集 VSM，计算测试集 VSM 时 IDF 使用训练集计算结果。

STEP3：遍历测试集，计算测试文档与所有训练集文档的 cos 值，排序提取前 K 个训练集文档。

STEP4：统计 K 个训练及文档确定类型，与测试文档类型进行比较判断是否正确，记录结果。

重复以上步骤取得多次试验结果。

2、词频范围确定

通过观察单词统计结果，最后确定取单词词频在 5 和 5000 之间的单词。

3、K 值确定

试验过程总测试了 K 值取 3-7 值时效果，随着 K 增长成功率增长不明显，最终取 K 为 7。

4、实验结果

通过 10 次测试成功率最高为 80.66%，最低为 79.12%。结果如下：

测试序号	分类成功次数	测试文档总数	成功率	备注
0	3031	3759	80.63%	
1	3032	3759	80.66%	最高
2	3000	3759	79.81%	
3	2988	3759	79.49%	
4	2974	3759	79.12%	最低
5	2991	3759	79.57%	
6	2993	3759	79.62%	
7	3012	3759	80.13%	
8	3013	3759	80.15%	
9	2998	3759	79.76%	

吉晓辉

2018 年 11 月 04 日