TF-IDF的实现及利用TF-IDF文本分类

苏智龙

一、TF-IDF计算过程

假设有语料库一共只要2篇文档：d1和d2，其中d1=(A,B,C,D,D)一共有5个单词组成；d2=(B,E,C,B)一共有4个单词组成。

1. TF

TF即词频(Term Frequency)，每篇文档中关键词的频率

TF=某个词在文章中出现的次数。

考虑到文章有长短之分，为了便于不同文章的比较，进行”词频”标准化：

TF=某个词在文章中出现的次数/文章总词数 (1)

计算过程如表1.

	d1	d2
A	1/5	0/5
B	1/5	2/5
C	1/5	1/5
D	2/5	0/5
E	0/5	1/5

表1 TF值

2. IDF

IDF即逆文档频率(Inverse Document Frequency)，为了防止分母为零，分子分母同时加1。表2为IDF计算过程。

IDF=ln(文档总数+1/包含该词的文档数+1) (2)

A	ln(3/2)
B	ln(3/3)
C	ln(3/3)
D	ln(3/2)
E	ln(3/2)

表2 IDF值

3. TF-IDF

TF-IDF=TF * IDF (3)

表3为TF-IDF计算过程。

	d1	d2
A	1/5 * ln(3/2)	0/5 * ln(3/2)
B	1/5 * ln(3/3)	2/5 * ln(3/3)
C	1/5 * ln(3/3)	1/5 * ln(3/3)
D	2/5 * ln(3/2)	0/5 * ln(3/2)
E	0/5 * ln(3/2)	1/5 * ln(3/2)

表3 TF-IDF值

二、利用TF-IDF的分类过程

1. 从新浪新闻找了4篇不同类型的文章，分别是科技、军事、财经和体育，和1篇科技类测试文档，分别存为txt文档，文档名为keji.txt, junshi.txt, caijing.txt, tiyu.txt, keji_test.txt。txt文档见附件。
2. TF-IDF计算

1）用jieba分词库对文章进行分词；

2）用stopwords.txt除去停用词；

3）计算TF值：

文章一共5篇，利用公式（1）得到图1结果。其中只取了降序排列的前20。

```
------------------keji.txt文章tf前20词汇-----------------
:0.181102362205 nova:0.0866141732283 机型:0.0236220472441 采用:0.0236220472441
快充:0.0236220472441 40W:0.0236220472441 华为:0.0236220472441 支持:0.0236220472441
Pro:0.0236220472441 鲁:0.0157480314961 万:0.0157480314961 大师:0.0157480314961
手机:0.0157480314961 方案:0.0157480314961 系列:0.0157480314961 第一款:0.0157480314961
5i:0.0157480314961 屏幕:0.0157480314961 首款:0.0157480314961 1080:0.00787401574803
------------------junshi.txt文章tf前20词汇-----------------
核弹头:0.0544554455446 枚:0.039603960396 核武器:0.029702970297 美国:0.029702970297
报告:0.0247524752475 国家:0.019801980198 核武库:0.019801980198 中国:0.019801980198
核:0.019801980198 削减:0.019801980198 称:0.019801980198 估计:0.0148514851485
全球:0.0148514851485 低:0.0148514851485 年:0.0148514851485 巴基斯坦:0.0148514851485
当量:0.0148514851485 下降:0.00990099009901 2019:0.00990099009901 情况:0.00990099009901
------------------caijing.txt文章tf前20词汇-----------------
中:0.05 中美:0.05 陆慷:0.0333333333333 找到:0.0333333333333 解决办法:0.0333333333333
协商:0.0333333333333 对话:0.0333333333333 :0.0166666666667 协议:0.0166666666667
出路:0.0166666666667 双赢:0.0166666666667 原则:0.0166666666667 月:0.0166666666667
平等互利:0.0166666666667 两国人民:0.0166666666667 美:0.0166666666667 通话:0.0166666666667
两国:0.0166666666667 重申:0.0166666666667 经贸:0.0166666666667
------------------tiyu.txt文章tf前20词汇-----------------
卡帅:0.036496350365 中卫:0.029197080292 恒大:0.029197080292 面对:0.014598540146
朴志洙:0.014598540146 防线:0.014598540146 体系:0.014598540146 U23:0.014598540146
昔日:0.014598540146 战术:0.014598540146 冯潇霆:0.014598540146 防守:0.014598540146
换上:0.014598540146 上港:0.014598540146 球迷:0.014598540146 政策:0.014598540146
:0.00729927007299 取得胜利:0.00729927007299 球员:0.00729927007299 布朗宁:0.00729927007299
------------------keji_test.txt文章tf前20词汇-----------------
5G:0.0861538461538 手机:0.0584615384615 4G:0.0307692307692 :0.02 套餐:0.0153846153846
资费:0.0138461538462 年:0.0123076923077 直播:0.0123076923077 价格:0.0123076923077
视频:0.0123076923077 提供:0.00923076923077 互联:0.00923076923077 万物:0.00923076923077
价值:0.00769230769231 中:0.00769230769231 买:0.00769230769231 1080p:0.00615384615385
服务:0.00615384615385 平台:0.00615384615385 网速:0.00615384615385
------------------------------------------------------------
```

图1 5篇文章的tf值

4）计算IDF：

利用公式（2）得到图2结果。其中只取了降序排列的前20。

5）计算TF-IDF：

利用公式（3）求的5篇文章的TF-IDF值，如图3。其中只取了降序排列的前20。

```
-----------------keji.txt文章idf前20词汇----------------
鲁:1.09861228867 万:1.09861228867 1080:1.09861228867 值得注意:1.09861228867
980:1.09861228867 大师:1.09861228867 像素:1.09861228867 Mate:1.09861228867
内存:1.09861228867 四摄:1.09861228867 一枚:1.09861228867 曝光:1.09861228867
珍珠:1.09861228867 4800:1.09861228867 方案:1.09861228867 配备:1.09861228867
系列:1.09861228867 2340:1.09861228867 后置:1.09861228867 TOF:1.09861228867
-----------------junshi.txt文章idf前20词汇----------------
国家:1.09861228867 估算:1.09861228867 各有:1.09861228867 130:1.09861228867
共有:1.09861228867 现有:1.09861228867 研究所:1.09861228867 日前:1.09861228867
推测:1.09861228867 年鉴:1.09861228867 下降:1.09861228867 2019:1.09861228867
全球:1.09861228867 希望:1.09861228867 结构:1.09861228867 公开:1.09861228867
条约:1.09861228867 巡航导弹:1.09861228867 核打击:1.09861228867 情况:1.09861228867
-----------------caijing.txt文章idf前20词汇----------------
出路:1.09861228867 双赢:1.09861228867 原则:1.09861228867 平等互利:1.09861228867
两国人民:1.09861228867 通话:1.09861228867 两国:1.09861228867 重申:1.09861228867
经贸:1.09861228867 违背:1.09861228867 关切:1.09861228867 互利:1.09861228867
总体:1.09861228867 陆慷:1.09861228867 达成:1.09861228867 基础:1.09861228867
例行:1.09861228867 照顾:1.09861228867 期盼:1.09861228867 全世界:1.09861228867
-----------------tiyu.txt文章idf前20词汇----------------
取得胜利:1.09861228867 球员:1.09861228867 布朗宁:1.09861228867 变得:1.09861228867
看好:1.09861228867 解围:1.09861228867 替换:1.09861228867 机会:1.09861228867
开场:1.09861228867 场:1.09861228867 上周末:1.09861228867 并不需要:1.09861228867
钟义:1.09861228867 缺兵:1.09861228867 面对:1.09861228867 强敌:1.09861228867
朴志洙:1.09861228867 李学鹏:1.09861228867 鲁能:1.09861228867 调教:1.09861228867
-----------------keji_test.txt文章idf前20词汇----------------
特性:1.09861228867 发放:1.09861228867 条件:1.09861228867 购入:1.09861228867
经验:1.09861228867 通信:1.09861228867 国军:1.09861228867 国内:1.09861228867
服务器:1.09861228867 打开:1.09861228867 1080p:1.09861228867 服务:1.09861228867
何出:1.09861228867 GB:1.09861228867 话:1.09861228867 难以:1.09861228867
日程:1.09861228867 平台:1.09861228867 何种:1.09861228867 运营商:1.09861228867
------------------------------------------------------------
```

图2　5篇文章的IDF值

```
-----------------keji.txt文章tfidf前20词汇----------------
nova:0.0951553950815 快充:0.0259514713859 40W:0.0259514713859 华为:0.0259514713859
Pro:0.0259514713859 鲁:0.0173009809239 万:0.0173009809239 大师:0.0173009809239
方案:0.0173009809239 系列:0.0173009809239 第一款:0.0173009809239 5i:0.0173009809239
屏幕:0.0173009809239 首款:0.0173009809239 机型:0.0163735554463 采用:0.0163735554463
支持:0.0163735554463 手机:0.0109157036309 1080:0.00865049046195
值得注意:0.00865049046195
-----------------junshi.txt文章tfidf前20词汇----------------
核弹头:0.0598254216601 枚:0.043509397571 核武器:0.0326320481783 美国:0.0326320481783
报告:0.0271933734819 国家:0.0217546987855 核武库:0.0217546987855 核:0.0217546987855
削减:0.0217546987855 全球:0.0163160240891 巴基斯坦:0.0163160240891 当量:0.0163160240891
中国:0.0137256867438 称:0.0137256867438 下降:0.0108773493928 2019:0.0108773493928
情况:0.0108773493928 拥有:0.0108773493928 英国:0.0108773493928 以色列:0.0108773493928
-----------------caijing.txt文章tfidf前20词汇----------------
中美:0.0549306144334 陆慷:0.0366204096223 解决办法:0.0366204096223 协商:0.0366204096223
对话:0.0366204096223 找到:0.0231049060187 出路:0.0183102048111 双赢:0.0183102048111
原则:0.0183102048111 平等互利:0.0183102048111 两国人民:0.0183102048111
通话:0.0183102048111 两国:0.0183102048111 重申:0.0183102048111 经贸:0.0183102048111
违背:0.0183102048111 关切:0.0183102048111 互利:0.0183102048111 总体:0.0183102048111
达成:0.0183102048111
-----------------tiyu.txt文章tfidf前20词汇----------------
卡帅:0.0400953390025 中卫:0.032076271202 恒大:0.032076271202 面对:0.016038135601
朴志洙:0.016038135601 防线:0.016038135601 体系:0.016038135601 U23:0.016038135601
昔日:0.016038135601 冯潇霆:0.016038135601 防守:0.016038135601 换上:0.016038135601
上港:0.016038135601 球迷:0.016038135601 政策:0.016038135601 战术:0.0101189369425
取得胜利:0.0080190678005 球员:0.0080190678005 布朗宁:0.0080190678005
变得:0.0080190678005
-----------------keji_test.txt文章tfidf前20词汇----------------
5G:0.0946496741006 手机:0.0405224505558 4G:0.0338034550359 套餐:0.016901727518
资费:0.0152115547662 直播:0.0135213820144 价格:0.0135213820144 视频:0.0135213820144
提供:0.0101410365108 互联:0.0101410365108 万物:0.0101410365108 年:0.00853104222228
价值:0.00845086375899 买:0.00845086375899 1080p:0.00676069100719 服务:0.00676069100719
平台:0.00676069100719 网速:0.00676069100719 笔者:0.00676069100719 网络:0.00676069100719
------------------------------------------------------------
```

图3　5篇文章的TF-IDF值

3. 文本分类

5篇文章中，keji_test.txt是待分类的文章。通过上述过程，已经求出它的TF-IDF值，降序排列取前20个作为它的特征值。

本文利用文章中前20个特征值中，与前面4篇文章通过TF-IDF值求出的前20个词相同个数作为文章分类标准。

其中，与keji.txt相同的词数为1，与其他文章相同词数为0。如图4。

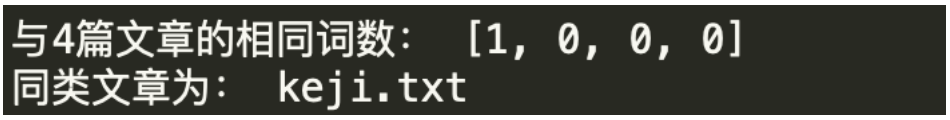

图4 相同词数和分类结果

综上：keji_test.txt和keji.txt同属于科技类文章。