从豆瓣影评中提取电影标签

——中文信息处理课程项目

梁晓涛

13307130319

January 9, 2016

Contents

1 问题的背景

在豆瓣的网站上对一部电影进行评价时，页面上会显示若干常用标签供用户参考（图 1）。并且每篇影评除了文章内容外，还有一个有用数和无用数（图 2），表示其他用户赞同或反对这篇影评的数目，以及文章的发布时间等其他关于影评的数据。于是就想能不能利用这些数据，使用自然语言处理的技术，从用户的影评中自动提取电影的标签。

图 1: 电影《神探夏洛克》的常用标签

图 2: 某篇影评的有用数和无用数

2 语料的收集和整理

2.1 爬取影评

写了一个爬虫，自动以固定的频率从豆瓣上爬取电影下的长评，利用正则表达式从 HTML 页面中提取影评的数据，并把数据以 CSV 格式保存下来，每一行对应一篇影评（表 1，图 3）。最终收集了 20 部电影的一共 2000 篇影评，CSV 文件的总大小为 8.4M。

2.2 处理文本

1. 把文章中的 HTML 标签去掉，只保留标签之间的内容

subject	movie	title	star	time	useful	useless	content
电影在豆瓣的编号	电影名字	标题	评分	发布时间	有用数	无用数	文章内容

表 1: 数据描述

	A	B	C	D	E	F	G	
1	subject	movie	title	star	time	useful	useless	content
2	1292720	阿甘正传	阿甘正传	5	2007-11-29 13:58:31	2405	124	94年电影界诞生了很多经典之作，阳光灿烂的日子、活着、低俗小说。这个
3	1292720	阿甘正传	飘飞的羽毛	5	2005-06-13 18:18:47	1449	124	常常和朋友们谈起阿甘。阿甘太运气了！朋友们总是这样说。我哑然。我很
4	1292720	阿甘正传	一羽人生	5	2005-11-25 19:04:16	1017	30	第一次看《阿甘正传》，我为片头和片尾的那片羽毛而困惑，导演以这片羽毛
5	10463953	模仿游戏	技术基佬什么的太有爱了！（附真人八卦）	5	2015-01-09 11:50:41	4778	75	多年前我大一的时候，有一门全系公选课叫“信息技术导论”，每一节课讲一
6	10463953	模仿游戏	他和他的Christopher	4	2014-11-27 06:23:00	1887	23	先分享一些电影背后的小八卦（随手直译自IMDB，包含剧透）1.丘吉尔认为
7	10463953	模仿游戏	把图灵拍成Sheldon Cooper真的好么。。	3	2015-01-10 16:13:00	1160	454	1，我觉得这件事简直侮辱人。编剧的意思是，一个人IQ高了EQ必然低成屎
8	1307914	无间道	再回首，已是百年身	5	2011-09-12 13:52:30	1649	45	一直很喜欢陈永仁（梁朝伟）和他的前女友May（萧亚轩）巧遇的那个桥段，
9	1307914	无间道	【无间道系列】留下的三句话	5	2005-09-02 08:46:26	853	101	面对“无间道III终极无间”狂轰滥炸的宣传攻势，我有些按耐不住，I和II
10	1307914	无间道	Day 4 - IMDB 247 - 你的孤独一文不值	None	2009-04-12 09:13:01	601	24	十一年前，在青松观的大殿上，琛哥对着七名新人说，出来行走江湖的，是生
11	10533913	头脑特工队	一种超越星盘和八字的性格分析法	5	2015-06-19 18:25:26	2946	67	为什么我们在面对一个人时心里会问：TA到底在想什么？ 为什么有
12	10533913	头脑特工队	《头脑特工队》彩蛋大搜集（10.7更新）	5	2015-07-08 15:10:35	1081	18	飞屋环游记莱莉的大脑中有成千上万个记忆球，其中就包括《飞屋环游记》开
13	10533913	头脑特工队	甩那些卖萌咯吱人笑的动画几条街	5	2015-06-22 11:33:21	895	81	当时看完电影哭到不能自已然后写下了下面这些文字（我室友全场冷静导致从
14	1889243	星际穿越	当你想描写一个触手可及的未来，然而却.....	5	2014-11-07 00:31:01	7394	196	先推一篇视角独特的影评：被忽略的阿弗莱克 http://www.douban.com/note/4
15	1889243	星际穿越	Interstellar 观影感+全剧透+2刷发现果然没漏洞。。	5	2014-11-05 15:13:13	3316	475	11.4号70mm IMAX场，提前接近一个月就订好了票，连基友都没有带就准备
16	1889243	星际穿越	诸兰的维度	3	2014-11-06 22:35:00	2223	348	Spoiler Alert在我们进入对“Interstellar”的具体讨论之前，也许需要对诸兰导
17	24405378	王牌特工：特工学院	【有重大剧透】关于本片的fun facts	5	2015-02-18 12:25:37	3586	118	【严禁未授权转载或商用】关于导演：http://www.douban.com/note/5118127
18	24405378	王牌特工：特工学院	你们只要看柯林叔静静地耍帅就可以了（无剧透）	4	2015-02-01 06:11:12	1274	126	该评论无任何实质性内容，因为这就是一部很有趣的动作喜剧，没什么可说的
19	24405378	王牌特工：特工学院	商业电影的胜利	5	2015-01-28 13:07:57	664	48	我其实一直不喜欢“商业电影”和“艺术电影”两种区分方法，好像“商业”
20	1292656	心灵捕手	天才为什么要堕落	5	2009-03-27 17:14:43	6670	181	麻省理工学院（MIT）的蓝勃教授是数学界中大名鼎鼎的人物，他获得过被誉
21	1292656	心灵捕手	一段让我恐惧的台词	4	2009-01-31 12:42:26	1674	48	《心灵捕手》这部片子，有一段台词让我印象深刻，那是sean对will说的一段
22	1292656	心灵捕手	你根本不知道你在说什么.....	4	2006-08-30 11:29:44	1363	102	一位学心理学的朋友几个月前推荐我看这部电影，今天我才“履行”了我的任
23	1292052	肖申克的救赎	十年.肖申克的救赎	5	2005-05-12 20:44:13	6370	165	{原文}：http://www.bighead.cn/?p=34这些天按时上下班，衣冠楚楚，与
24	1292052	肖申克的救赎	《肖申克的救赎》与斯德哥尔摩综合症 - - 你我都是患者！	5	2007-09-15 22:59:08	2610	294	斯德哥尔摩综合症（Stockholm syndrome），斯德哥尔摩效应，又称斯德哥
25	1292052	肖申克的救赎	终于找到了郁闷人生的原因 _观《肖申克的救赎》有感	5	2005-07-12 11:23:39	2672	157	周末看了一部美国影片《肖申克的救赎》（《The Shawshank Redemption》
26	3793023	三傻大闹宝莱坞	那一场关于理想的美梦	4	2010-08-20 09:05:38	2572	152	我最喜欢的一位单口相声表演者叫Russell Peters，他是加拿大第二代印度移
27	3793023	三傻大闹宝莱坞	Aal izz well------《三傻大闹宝莱坞》VS《人在囧途》	5	2010-08-20 23:48:25	2040	124	有识之士路过请原谅我把这两部电影放在一起比较，事情的起因是前几天我向
28	3793023	三傻大闹宝莱坞	三傻大闹宝莱坞3 Idiots 经典台词	5	2010-08-05 17:41:13	1692	79	“他的人和名字一样不同寻常” “一出生就有人告诉我们，生活是场赛跑，不
29	24751756	老炮儿	我该谈论什么，死亡还是信仰	4	2015-12-16 02:04:33	5431	357	（文/杨时旸）严格来讲，这其实是一部方言电影，它其中的对白并不是真正
30	24751756	老炮儿	赤诚的火焰	4	2015-11-14 12:17:16	4635	541	1.故事发生在北京的冬天。肮脏的雾，污浊的霾，干冷的风，和独属于北方的
31	24751756	老炮儿	马小军老了	4	2015-11-12 15:55:41	1687	191	我们骂冯小刚是个烂导演，骂了好多年，一不留神忘了他曾是个好演员。在19
32	2027909	12怒汉：十宾判	俄罗斯民族的自我审视	5	2009-04-15 22:30:38	459	20	去年看的片子，许多细节记不清楚了，在其论坛上跟W大讨论本片时忍不住写

Sheet1

图 3: 一小部分数据的截图

2. 对文章分词[1]
3. 只保留匹配`[a-zA-Z0-9\u4E00-\u9FFF]+`的词，从而把中英文的标点符号和空白字符去掉
4. 英文字母全部转为小写，以确保同一个单词不会因形式不同而被错认为是不同单词
5. 利用中英文停用词表，将停用词去掉

3 关键词提取算法

3.1 TF-IDF

TF-IDF 是一种用于评估一个词对一个文档集中其中一份文档的重要性的统计方法，以下作简单介绍，具体可参考这篇文献[2]。它的主要思想是：词频 TF，表示一个词在一篇文档中出现的频率，如果该词在文档中频繁出现，则说明该词很好地代表了这篇文档的文本特征；逆向文件频率 IDF，用来衡量一个词对文档的区分能力，如果文档集中包含这个词的文档越少，则说明这个词的类别区分能力越强，IDF 也就越小。而 TF-IDF 将两者综合起来，词 w_i 对于文档 d_j 的权重为

$$TFIDF_{i,j} = TF_{i,j} * IDF_i$$

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

[1]使用“结巴”中文分词 http://github.com/fxsjy/jieba

[2]H.C. Wu, R.W.P Luk, K.F. Wong, K.L. Kwok ”Interpreting TF-IDF term weights as making relevance decisions”. *ACM Transactions on Information Systems* 26(3):1-37, 2008.

其中 $n_{i,j}$ 表示词 w_i 在文档 d_j 中的出现次数，分母表示文档 d_j 中所有词的出现次数之和。

$$IDF_i = \log \frac{|D|}{|j : w_i \in d_j|}$$

$|D|$ 表示文档集中的文档总数，$|\{j : w_i \in d_j\}|$ 表示包含词 w_i 的文档数目。

3.2 基于 TF-IDF 的关键词提取算法

对于从影评中提取关键词的问题，可以重新描述如下：

有 K 类文档 $C_1, C_2, \ldots, C_K$，C_i 包含第 i 部电影的影评 $d_{i,j}$，$C_i = \{d_{i,1}, d_{i,2}, \ldots\}$，每篇文档 $d_{i,j}$ 都有一个属性集 $attr_{i,j}$，属性集中包含这篇影评的有用数、无用数、发布时间等数据。而我们的问题就是要给每类文档 C_i 提取关键词，要求关键词既能表示这类文档的文本特征，又能区分不同类别的文档。

一个简单的想法是，对于文档类 C_i，直接把其中的文档 $d_{i,j} \in C_i$ 合成一个文档 d_i，文档集 $D = \{d_1, d_2, \ldots, d_K\}$，利用上节介绍的 TF-IDF，求出合成文档 d_i 每个词的 TF-IDF，取 TF-IDF 最大的若干个词作为 C_i 的关键词。

这个简单想法的主要问题是没有利用影评的属性信息。例如，对于同一类文档 C_i，显然有用数多的文档比有用数少的文档更能表示 C_i。于是，我们考虑通过修改 TF-IDF 来提取关键词，除了使用影评的文章内容外，还使用影评的属性数据，从而提高关键词的相关性。

具体的做法如下：

对于每一篇文档 $d_{i,j}$，假设如果一个用户认为这篇文档有用，则相当于他写了一篇一模一样的文档，而如果一个用户认为这篇文档无用，则抵消一篇。如果无用数比有用数还多，则认为这篇文档对于表示 C_i 毫无贡献，所以计算文档 $d_{i,j}$ 的得分为

$$t_{i,j} = \max\{usefull_{i,j} - useless_{i,j} + 1, 0\}$$

利用每篇文档的得分 $t_{i,j}$，计算词 w_l 在 C_i 内文档的词频的加权平均

$$\overline{TF}_l^i = \frac{\sum_j t_{i,j} * TF_{l,j}^i}{\sum_j t_{i,j}}$$

其中 $TF_{l,j}^i$ 表示词 w_l 在 $d_{i,j}$ 中的词频。

IDF 描述词 w_l 的区分能力，计算 IDF 的时候并不区分一个词属于哪一个文档类，即相当于把不同类别的文档组成一个文档集

$$IDF_l = \log \frac{\sum_i |C_i|}{|\{(i,j) : w_l \in d_{i,j}\}|}$$

其中分子表示包括不同类别的所有文档数，分母表示包含词 w_l 的文档数目。

于是词 w_l 对于文档类 C_i 的权重为

$$TFIDF_l^i = \overline{TF}_l^i * IDF_l$$

取修改版 TF-IDF 最大的若干个词作为 C_i 的关键词。

4 实验结果

在语料库和对数据的预处理都相同的情况下，分别使用简单版 TF-IDF (TF-IDF) 和修改版 TF-IDF (TF-IDF*) 对选取的电影提取标签。表 2 中显示了部分实验结果，其中每部电影包含三行标签，依次是豆瓣上的常用标签、TF-IDF 提取的标签、TF-IDF* 提取的标签，每行包含 10 个标签，并且后两行的标签按权重递减的顺序排列。

提取方式	标签
	12 怒汉：大审判 12　(2027899)
豆瓣	米哈尔科夫 俄罗斯 剧情 人性 俄罗斯电影 翻拍 Mikhalkov 2007 电影 法律
TF-IDF	法律 陪审团 男孩 俄罗斯 无罪 12 怒汉 车臣 有罪 审判
TF-IDF*	俄罗斯 车臣 美版 12 陪审团 战争 法律 男孩儿 男孩 案件
	饮食男女　(1291818)
豆瓣	台湾电影 张艾嘉 李安 饮食男女 家庭 剧情 台湾 伦理 中国 电影
TF-IDF	女儿 父亲 李安 闷骚 家庭 老朱 朱 朱师傅 同期 饮食男女
TF-IDF*	父亲 情结 埃勒克 特拉 女儿 家倩 一个 李安 朱师傅 艺术品
	肖申克的救赎　(1292052)
豆瓣	信念 剧情 人生 经典 美国 人性 自由 励志 1994 犯罪
TF-IDF	安迪 Andy 自由 监狱 瑞德 体制 肖申克 希望 Red 救赎
TF-IDF*	斯德哥尔摩 自由 安迪 Andy 监狱 体制 瑞德 习惯 Red 综合症
	星际穿越　(1889243)
豆瓣	宇宙 人性 美国 星际 冒险 亲情 2014 科幻 剧情 悬疑
TF-IDF	Cooper 黑洞 人类 诺兰 库珀 星际 洞 穿越 虫 星球
TF-IDF*	黑洞 Cooper Matthew 人类 洞 虫 诺兰 星球 女儿 地球
	三傻大闹宝莱坞　(3793023)
豆瓣	搞笑 剧情 励志 印度 人生 宝莱坞 喜剧 经典 爱情 2009
TF-IDF	na 印度 兰彻 Rancho 理工 电影 梦想 祖碧 杜比 学生
TF-IDF*	印度 Rancho na 途 兰彻 理工 病毒 兰乔 傻瓜 也许

表 2: 对比选取的标签

4.1　豆瓣与 TF-IDF* 的标签的比较

对于电影《12 怒汉: 大审判 12》，豆瓣和 TF-IDF* 的公共标签有“俄罗斯”、“法律”，豆瓣的标签还有表示电影类型的“剧情”、“翻拍”，导演的中英文名字“米哈尔科夫”、“Mikhalkov”，以及跟具体电影无关的词“电影”。TF-IDF* 则提取了几个特别的词“车臣”、“美版”、“战争”，“车臣”和“战争”反映了电影故事的背景是俄罗斯和车臣之间的战争，而“美版”反映了用户把这部翻拍的电影与 1957 年美版的电影进行比较。可以注意到 TF-IDF* 的标签中同时包含了“男孩儿”和“男孩”这两个表达相同含义的词。

对于电影《饮食男女》，豆瓣和 TF-IDF* 的公共标签有“李安”，豆瓣的标签还有表示电影类型的“剧情”、“伦理”，演员的名字“张艾嘉”。TF-IDF* 则提取了几个特别的词“情结”、“埃勒克”、“特拉”、“父亲”、“女儿”，“埃勒克特拉情结”概括的是一种“恋父”情节，这几个词是对电影情节的更深层次的挖掘。但 TF-IDF* 的标签中也包含了“一个”等无用的词。

对于电影《星际穿越》，豆瓣的标签更多的是概括电影的剧情：“人性”、“冒险”、“亲情”、“科幻”、“剧情”、“悬疑”。TF-IDF* 则提取了跟物理现象有关的词“黑洞”、“洞”、“虫”、“星球”、“地球”，这反映了很多影评在解释这些物理概念以帮助分析剧情。注意到“虫洞”一词被拆分成了两个词，这是分词的问题，但 TF-IDF* 还是能把“洞”和“虫”两个词排在一起。

总的来说，豆瓣的标签主要的用途是帮助用户给电影分类，所以往往出现导演名字、演员名字、电影的年份、电影的类型，以及一些概括电影剧情的词，同时还有一些用处不大的词，如“电影”。

而 TF-IDF* 提取的词，如果包含了导演或演员的名字，往往是因为他们对这部电影贡献特别大或者表现特别突出。TF-IDF* 能够提取一些表达电影更深层次的内涵的词，如电影的主题和背景。

4.2 TF-IDF 与 TF-IDF* 的标签的比较

对于电影《饮食男女》，TF-IDF 提取的词大部分只是一些在电影中频繁出现的词：“女儿”、“父亲”、“家庭”、“老朱”、“朱师傅”。而 TF-IDF* 则能提取出一些总结性、表示观众观点的词：“情结”、“埃勒克”、“特拉”、“艺术品”。并且可以注意到，TF-IDF 提取的“朱”、“老朱”、“朱师傅”三个词都表示相同的意思，而 TF-IDF* 中的只出现了“朱师傅”。这主要是不同用户对同一个对象的叫法不同导致，而 TF-IDF* 相比简单 TF-IDF 更集中于某些影评，从而在一定程度上减少同义词的出现。

对于电影《三傻大闹宝莱坞》，“na”这个词在两种方法中都有出现。通过检查语料库，发现这个词之所以出现是因为有一篇影评引用了电影中的歌曲，歌词中大量出现“na”这个词，虽然“na”的 IDF 很小，但它的 TF 实在太大，导致总的权重大。但 TF-IDF* 并没有像简单 TF-IDF 那样把这个词排在了第一，只是排在了第三，说明 TF-IDF* 相比于简单 TF-IDF 更能降低“噪声”。

总的来说，简单 TF-IDF 倾向于提取电影中频繁出现的词，而 TF-IDF* 能提取表达观众观点的词，并且能相对地降低一些同义词、噪声词的权重。

5 总结

本项目提出了一个从豆瓣影评中提取电影标签的方法，该方法从经典的 TF-IDF 方法出发，利用影评的元数据得到了效果更好的关键词。相比于理论知识，在实际应用中需要考虑更多的东西：怎么获取语料库，怎么处理杂乱的数据，怎么利用实际应用中数据的各种信息改进理论的方法。

本项目还有许多可以改进的地方。提取的关键词的效果跟分词的效果直接相关，限于影评语料库的大小，分词时只利用了结巴分词的语料库，如果利用影评语料库来分词，是否可以更好地发现未登陆词，从而提高关键词的质量。提出的算法只基于简单的统计，相信如果考虑语义、词性等，能得到更好的效果。除了利用影评的有用数、无用数，还可以利用发布时间等其它信息，甚至可以把用户的账号信息利用起来。

A 代码的说明

目录结构：

```
code/
├── subject_list.txt          # 记录电影在豆瓣上的ID
├── douban_big.csv            # 影评数据
├── chinese_stopwords.txt     # 中文停止词
├── scratch.py                # 爬虫
└── extract.py                # 关键词提取
```

依赖：

- Python3 (http://www.python.org/)
- Jieba (http://github.com/fxsjy/jieba)

- NLTK (http://www.nltk.org/)

A.1 scratch.py

该脚本用于从豆瓣上爬取影评数据：

- 从文件 `subject_list.txt` 中读入需要爬取数据的电影的 ID
- 为了防止被封 IP，两次网络请求之间至少相隔 1.5s
- 通过修改脚本中的 `MaxLimit` 变量，控制每部电影最多爬取的影评数目
- 爬取的数据保存在文件 `douban.csv` 中

使用命令：`python3 scratch.py`

A.2 extract.py

该脚本用于从爬取的影评数据中提取关键词。
使用命令：

```
Usage: python3 extract.py [options] filename
filename: 影评数据的文件名
optinos:
    -k [NUM]:    指定提取关键词的数目，默认是10
    -w:          要求输出关键词的权重
```