# 基于用户评论的电影关键词提取工具

丁成
13307130299
sding13@fudan.edu.cn

**摘要**

本工具以豆瓣电影为代表，未经监督训练，抓取用户评论，再通过用户的评论文章，使用 TF-IDF 来提取一部电影的关键词信息，并将关键字评分排序，展现为可视化数据图表。从数据抓取到数据展现，全程均实现了自动化。

## 1. 介绍

目前，所有主流的电影资讯、评论、分享网站，为一部电影提供的类别区分仅有数十种左右，且相当模糊。电影作为数百分钟的音乐、画面、剧情故事载体，题材与形式的组合成千上万，大部分情况下并不能被粗暴地归为几类。并且，这样的区分方式十分不适合用户来整理、挑选电影。

以世界上最大的电影资料库 IMDB 为例，其中用户评分最高[1]的电影《肖申克的救赎》（The Shawshank Redemption），只有“犯罪（crime）”和“剧情（drama）”两个类别[2]。而在最大的华语电影资料库豆瓣电影上，《肖申克的救赎》同样也只有这两个类别[3]。而且豆瓣上，官方提供的标准化主题关键词数据则有更加限。而用户自行添加的标签多为“经典”、“励志”，有时又为电影年代如“1949”，有时又为演员昵称，都未经过标准整理。

本工具以豆瓣电影为代表，未经监督训练，通过用户的评论文章来提取一部电影的关键词信息，并将关键字评分排序，展现为数据图表。从数据抓取到数据展现的整个过程，均实现了自动化。例如同一部电影，本工具得到的结果如下：此工具的主要特点有：

- 通过电影名，自动搜索并抓取所有评论信息，存入数据库；

[1] Top Rated Movies - IMDB
[2] The Shawshank Redemption - IMDB
[3] 肖申克的救赎 - 豆瓣

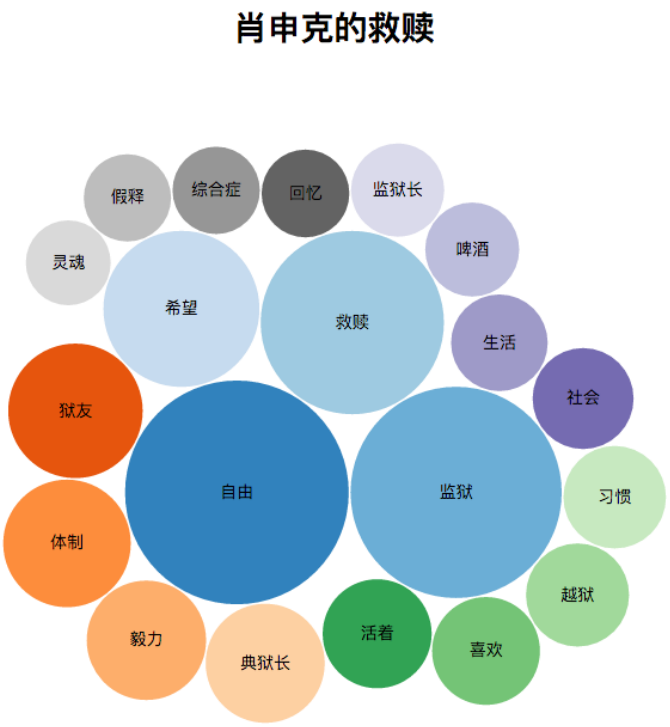


**Figure 1.** 《肖申克的救赎》自动提取关键词

- 不同关键词的大小表示其权重，权重越大表示关键词越重要；
- 提取剧情、题材相关的关键性名词、动词，例如上图的“自由”、“监狱”、“希望”、“毅力”、“活着”等。这些词构成了一部电影的骨干；
- 提取形容电影节奏、画面、音乐、演员表演水平等等部分的形容词（上例中未出现）；
- 过滤了非电影内容本身相关的词语，例如“出品年代”、“影院”、“IMAX”这些经常出现的标签；
- 整理出出现关键词的句子，可交互查看关键词的详情。


**Figure 2.** 短评

# 2. 细节实现

## 2.1. 搜索、抓取、规范化数据

本项目独立实现了一个爬虫，并接入了豆瓣电影的各个 API。运行步骤如下：

1. 第一步，爬虫通过电影名称搜索到其唯一编号（search 接口）；
2. 在豆瓣电影页面获取其基本信息（details 接口），如演员、类型、短评入口、长评入口；
3. 分别抓取一定数量的短评（comments 接口）、长评（reviews 接口）；
4. 将各个评论的评分（一至五星）、内容、用户赞同数量等数据清理成可用的格式（去掉空格，转换成评分百分比等等）

爬虫脚本使用 Node.js 实现，所有接口都支持并发、缓存等特性，代码位于 `core/api.js`。


肖申克的救赎的影评 · · · · · · ( 全部3586 )
我来评论这部电影
十年·肖申克的救赎
大头绿豆 2005-05-12 20:44:13
{原文}: http://www.bighead.cn/?p=34 这些天按时上下班，衣冠楚楚，与时俱进，过得颇麻木。于是夜里心情便有些低落，寻了肖申克的救赎来看。 距离 Frank Darabont 们缔造这部伟大的作品已经有十年了。我知道美好的东西想必大家都能感受，但是很抱歉，我的聒噪仍将一如既往。 今夜在我眼里，The Shawshank Redemption 与信念、自由和友谊有关。 [1]、信念 Red .....
.....
6385/6551 有用 532回复
《肖申克的救赎》与斯德哥尔摩综合症－－你我都...
中原 2007-09-15 22:59:08
斯德哥尔摩综合症（Stockholm syndrome），斯德哥尔摩效应，又称斯德哥尔摩症候群或者称为人质情结或人质综合症，是指犯罪的被害者对于犯罪者产生情感，甚至反过来帮助犯罪者的一种情结。这个情感造成被害人对加害人产生好感、依赖心、甚至协助加害人。 1973年8月23日，两名有前科的罪犯Olsson与Olofsson，在意图抢劫瑞典首都斯德哥尔摩市内最大的一家银行失败后，挟持.....
.....

**Figure 3.** 长评

## 2.2. 分析

### 2.2.1. 短评

根据粗略观察，短评限于其字数，没有办法围绕一个重点来展开讨论。以《肖申克的救赎》为例，排名第一的短评为

> 忒经典的东西，我要带去我的坟墓

其中关键词可能有 “经典” 和 “坟墓”，而 TF-IDF 算法会将 “坟墓” 视为第一关键词。而根据语义，“坟墓” 和电影主题并不相关。

大多数短评围绕的内容比较五花八门，因此其中名词并无太多利用价值。而大部分用户写一句话短评的时候，其中形容词几乎都用于形容电影，例如这个短评：

> “这是一部男人必看的电影。” 人人都这么说。但单纯从性别区分，就会让这电影变狭隘。《肖申克的救赎》突破了男人电影的局限，通篇几乎充满令人难以置信的温馨基调，而电影里最伟大的主题是 “希望”。当我们无奈地遇到了如同肖申克一般囚禁了心灵自由的那种囹圄，我们是无奈的老布鲁克，灰心的瑞德，还是智慧的安迪？运用智慧，信任希望，并且勇敢面对恐惧心理，去打败它？经典的电影之所以经典，因为他们都在做同一件事——让你从不同的角度来欣赏希望的美好。

其中大多数形容词，例如 “温馨”、“无奈”、“灰心”、“智慧”、“勇敢”、……都是电影中重要的主题。

另一类短评，会围绕电影本身形式的一些方面进行描述，例如 “摄影构图精美”、“配乐恢宏”、“剧情波折” 等等。

### 2.2.2. 长评

长评论因为没有了字数限制，可以通篇围绕一个或多个主题来展开详细评论。这样带来的好处是，其主题会在通篇文字中多次出现。例如这部电影下排名第一的评论[4]《十年 • 肖申克的救赎》，作者围绕信念、自由、友谊三个主题来进行详细讨论。这类文章往往会引导其下评论的方向，也为这三个主题。因此整个页面中 “自由” 一词出现近 40 次。

[4] 十年 • 肖申克的救赎

于是我们首先使用 TF-IDF 算法提取出全文关键词。接着，根据每篇长评之后的用户赞同数（例如本篇是：有用 6386 没用 166），进行评分。赞同数越多文章，其中关键字整体权重越大。

最后，我们合并所有长评的关键词与其评分，进行排序。总体公式如下：

$$s(word) = \sum_{\text{所有长评论}} \log\left(word\text{在此篇中的 TF-IDF 评分}\right) * \text{此篇权重} \quad (1)$$

**Note**. 权重定义为赞同数占所有长评赞同数之和的比。

## 2.3. 实现

在具体处理时，我们首先抓取了大量电影相关术语（共约 400 个），避免被视作关键词。例如：

镜头
奥斯卡
暗箱
借位
功夫片
电影人
院线
戏院
近景
特写
调色
...

这些词语从各个电影论坛抓起，文件位于 `core/movie.dict.utf8` 中。另外还有常用的 stopwords，位于 `core/stopwords.utf8` 中。

接着我们排除了一些标签，比如时间、地点、人物这类常常出现的、但与电影剧情等无关的词 [3]。

分词和 TF-IDF 使用了 jieba 中文分词库 [2]。这部分代码主要位于 `core/main.js` 中。

最后将所有数据传至前端，以 d3.js 渲染出来。

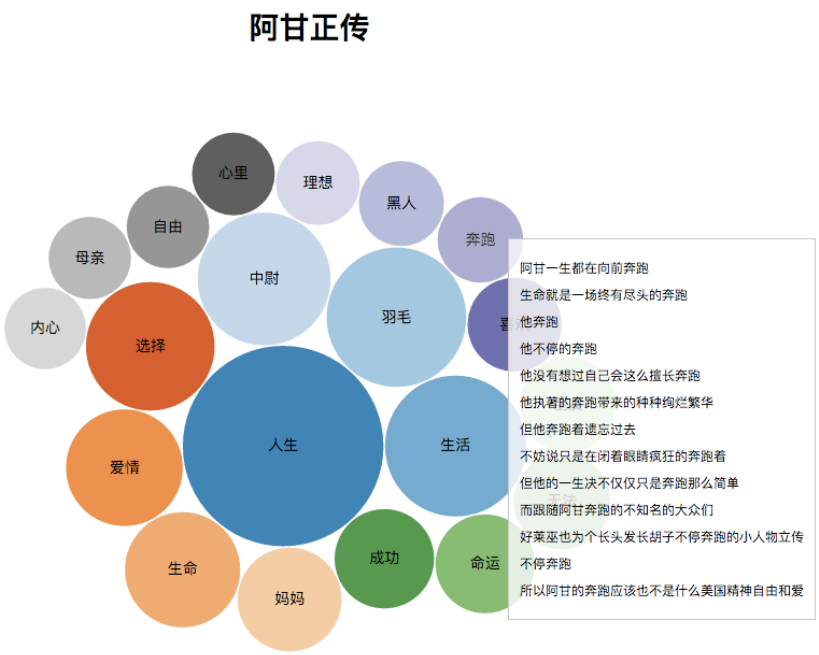


**Figure 4.** 阿甘正传

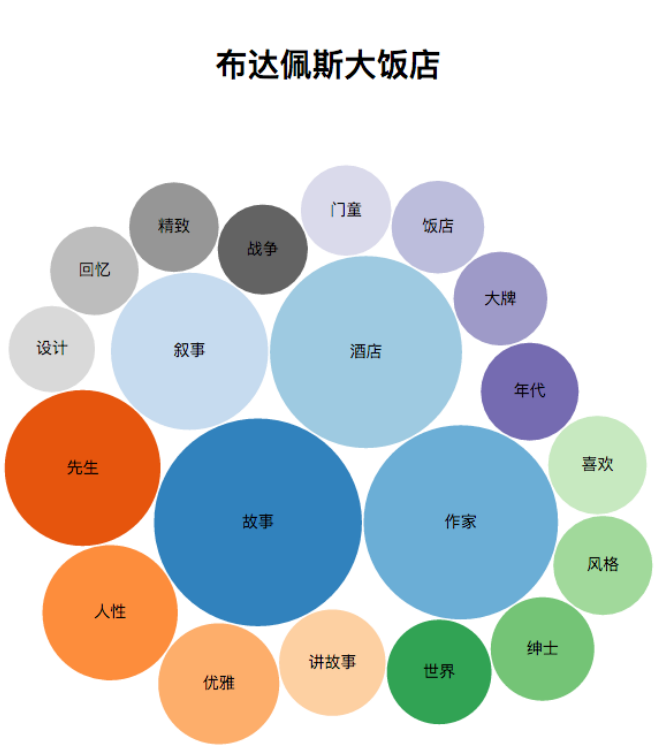


**Figure 5.** 布达佩斯大饭店


**Figure 6.** 星球大战

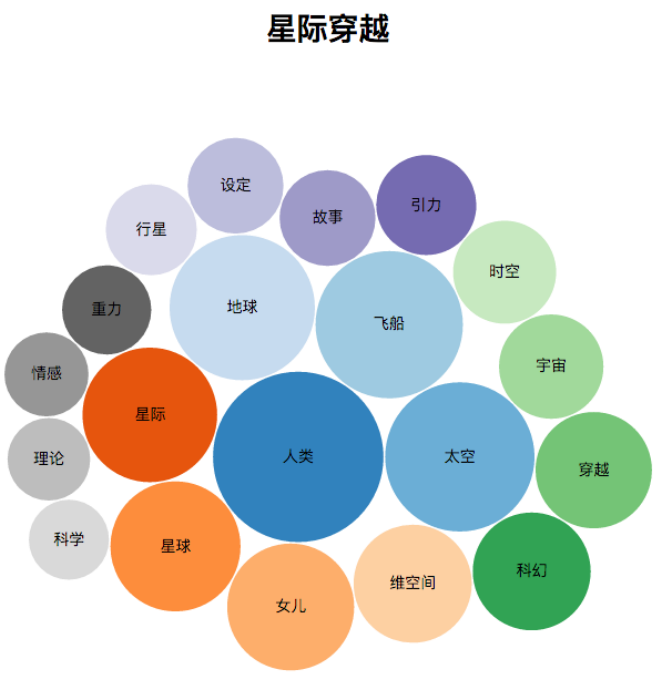


**Figure 7.** 星际穿越

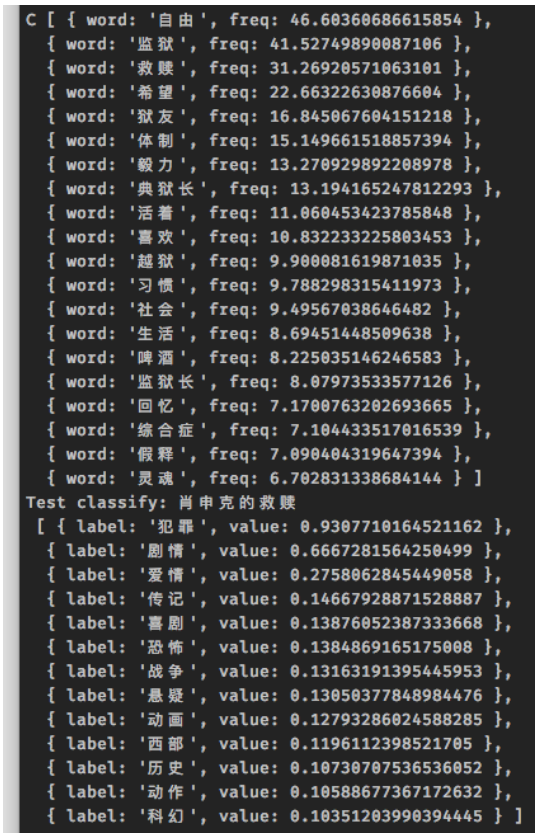


**Figure 8.** 《肖申克的救赎》电影类别判断

### 2.4. 效果展示

## 3. 实验

可以发现，根据网络用户自发的评论，我们确实可以总结出许多官方未提供的关键词信息。在数据量足够多的情况下，我们可以做许多有趣的分析。

比如通过比对大量科幻片，我们就能发现科幻电影中“飞船”出现的几率特别高。因此可以反过来，在只给出影评的情况下来判断一部电影的类型。

我们首先采用上文的方法，提取出若干关键词。再分别采用朴素贝叶斯和线性回归分类器做试验，由于数据量不够多，结果并不如期望那么好。图 8 是对《肖申克的救赎》的关键词统计，以及根据关键词和线性回归分类器所做的电影成分分析。

分类器的代码位于 `core/basic.js`，用于训练的电影列表位于 `index.js`，分类器目前训练出的结果在 `basic/classifier.json` 中。

## 4. 展望和总结

目前大部分人观看电影是根据档期（当前上映）。这种方式很难与自己喜爱的电影“不期而遇”。而如果数据量足够多，即使重口难调，我们也可以采用题材、元素的方式来挑选我们想看的电影。例如我们可以寻找“有梦境、枪战等元素的科幻片”（盗梦空间）。

通过一学期自然信息处理的学习，我从零开始接触到了文本信息处理的很多细节，学到了分句分词、语义、标注、分割、分类器等等概念，并自己动手实现、使用了这样一套工具流程。

本文想法和实现上参考了论文 [1]、[4]。最后感谢黄萱菁老师的辛勤付出！

## References

[1] Noah A. Smith David Bamman, Brendan O’ Connor. Learning latent personas of film characters. Master’s thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA, 2012.

[2] fxsjy. 结巴中文分词. URL https://github.com/fxsjy/jieba.

[3] luw2007. Ictpos3.0 词性标记集. URL https://gist.github.com/luw2007/6016931.

[4] Eduard Hovy Michael Fleischman. Recommendations without user preferences: A natural language processing approach. Master’s thesis, USC Information Science Institute, 4676 Admiralty Way, Marina del Rey, CA 90292-6695.