Central University of Finance and Economics

中央财经大学

非结构化大数据分析

爬虫练习

刘亭筱

吴宇翀

email@wuyuchong.com

指导老师：刘苗

2022 年 3 月 10 日

摘要

在此次作业中，我们爬取了豆瓣读书中的评论，我们选取《哈利波特》一书，爬取了 20 页评论，并将其进行词云绘制。

目录

1 可视化

图 1: 词云图

2 代码

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

# --------------------> 准备工作 ------------------------------------------

import time
import random
import requests
import jieba
import matplotlib.pyplot as plt
from bs4 import BeautifulSoup
```

```
from imageio import imread
from wordcloud import WordCloud
from wordcloud import ImageColorGenerator
# ------------------------------------------------------------------------------



# --------------------> 爬取评论 ------------------------------------------------
def getHtml(url, headers):

    try:
        r = requests.get(url, timeout=30, headers=headers)
        r.raise_for_status()
        return r.text
    except:
        return ''

# 获取评论
def getComment(html):
    soup = BeautifulSoup(html, 'html.parser')
    comments_list = []  # 评论列表
    comment_nodes = soup.select('.comment > p')
    for node in comment_nodes:
        comments_list.append(node.get_text().strip().replace("\n", "") + u'\n')
    return comments_list

# 获取并将评论保存到文件中
def saveCommentText(fpath, headers, pre_url, depth):
    with open(fpath, 'w', encoding='utf-8') as f:
        for i in range(1, depth):
            print(' 开始爬取第{}页评论...'.format(i))
            url = pre_url + 'start=' + str(20 * i) + '&limit=20&status=P&sort=new_score'
            html = getHtml(url, headers)
            f.writelines(getComment(html))
            # 设置随机休眠防止 IP 被封
            time.sleep(1 + float(random.randint(1, 20)) / 20)
    print(' 成功完成爬取任务')

# 书籍评论网址
pre_url = "https://book.douban.com/subject/2295163/comments/?"
```

```
# 浏览器信息 - 依据特定电脑信息（https://blog.csdn.net/ysblogs/article/details/88530124）
headers = {'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTM

# 输出文件位置
fpath = './outcome.txt'

# 爬取页数
depth = 20

saveCommentText(fpath, headers, pre_url, depth)
# ------------------------------------------------------------------------------


# ------------------->  可视化 ---------------------------------------------


# 切词
def cutWords(fpath):
    text = ''
    with open(fpath, 'r', encoding='utf-8') as fin:
        for line in fin.readlines():
            line = line.strip('\n')
            text += ' '.join(jieba.cut(line))
            text += ' '
    with open('cut_word.txt', 'w', encoding='utf-8') as f:
        f.write(text)
    print("\n分词完成，文件保存成功！")

# 绘制词云
def drawWordcloud():
    with open('cut_word.txt', encoding='utf-8') as f:
        comment_text = f.read()
    color_mask = imread("backgroud.png")  # 读取背景图片
    Stopwords = [u' 就是', u' 作者', u' 你们', u' 这么', u' 不过', u' 但是', u' 什么', u' 没有',
                 u' 这个', u' 那个', u' 大家', u' 比较', u' 看到', u' 真是', u' 就', u' 在', u'
                 u' 都', u' 但', u' 是', u' 多', u' 这', u' 看', u' 完', u' 了', u' 我', u' 也',
                 u' 除了', u' 时候', u' 已经', u' 可以', u', 'u'。']
    cloud = WordCloud(font_path="SimHei.ttf",  # 中文字体，否则无法显示
```

```
                    background_color='white',
                    max_words=200,
                    max_font_size=200,
                    min_font_size=4,
                    mask=color_mask,
                    stopwords=Stopwords)
    word_cloud = cloud.generate(comment_text)  # 产生词云
    image_colors = ImageColorGenerator(color_mask)
    # 保存图片
    word_cloud.to_file("comment_cloud.jpg")
    print('\n词云图保存成功')

cutWords(fpath)
drawWordcloud()
# ----------------------------------------------------------------------------
```

3 爬取评论展示

赶在去大阪的前一天重温完了这套书。为了把大布局收尾，罗琳硬生生编出两个“N大”设定，这是让线索合理化的最简单技巧
我用六个小时读完了《哈利波特与死圣》。中文版的。中间只站起来喝了一口水。合上书时人都要虚脱了。
我所看过的最杰出的一部通俗文学。
卢平，教授，一路走好
一个时代的完结
舍不得告别。
就这么结束了，书拿到手的时候，马不停蹄的读，一方面又害怕看完，是在是很矛盾。看到斯内普那段的时候流泪了……我就
一？难道小时候没读第七本？对作为成年人的我，这个幻想故事显得力道不足了。现实比幻想艰难多了，现实里读哈利波特长
2020年重读08：故事结束了，难过的时候我会拿出来读读，因为还相信某些就像书里所说早已被耻笑的东西，是因为这些才想
想写的太多来不及交代完的感觉，还有还我的双胞胎！
终于看完了，斯内普是最大的剧情了，总体来说，越来越烦哈利波特本人。。
这本稀碎情节太多，完全都是没必要的。比如哈利怀疑邓布利多的过去、罗恩的出走、哈利让罗恩摧毁魂器、赫敏再见到罗恩
呜呜呜哭了好几次。
在哈利三人组开始毫无头绪的逃亡之后，作者又试图不断用圣器魂器这类名词混淆读者的思路，但最终，还是以哈利在决斗中
反正总是要看完的嘛！
毕竟是大结局……
打眼一看评论，基本都是在沉痛悼念斯内普。
这两天重读的，比以前读懂得多了一点。也许，以前还不能体会到其中的一种温存的暖意。邓布利多的身世真的很令人惊讶，
很高兴我们能和全世界的孩子一起见证我们童年的快乐和回忆，以及一起画上的还算圆满的句号。
要是所有的书我都能读得那么快就好了：去大马出差了四天，等飞机看，吃饭看，坐车看，等人看，工作完了回宾馆继续看，

最后知道真相的我眼泪掉下来
对不起，我实在是无法接受这个结局。
2009.02.12.和这套书里的主人公一起长大了。在小说全部读完之后，有一种终于长大和童年告别的不舍感。故事里的大家都
不得不说对斯内普的处理实在对不起前面的无数铺垫，还不如让丫活着呢!
结束了
我怎么会没标过已读……（我永远爱哈利波特）
总感觉伏地魔太蠢了，也罢，谁让他是儿童文学里的坏人呢。。。7本哈利波特，到此，终于全部读完了
每看一遍都为斯内普哭得稀里哗啦的....小天狼星那次已经被虐惨了,没想到最后一本又被狠狠虐了一次!!!
啊居然就完了就完了=- -!
扣一分是因为赫敏居然和罗恩在一起了，为什么不是克鲁姆
结束了青春童话的一部份。
十一月重读哈利波特计划，顺利完成！看完只想说：哈利波特万岁！五年后再见!
终于，终于。
7
这本就俗套了呢
罗琳忽悠了全球读者
二刷。听到快大决战的部分已经半夜，想着不看完今晚是睡不着了，就拿起手机开始接着看。真的很感谢罗琳给的这个结局，
All was fine. And it's my final final.
纠结的Snape啊。。。
Severus
阿姨您太同人了
超帅超漂亮超好看..
结尾有点仓促..
終于結束了……
哎
陪伴我成长的故事……真的完结了……
惯性让人读完。
TXT甚嘛的还是很方便的~——为什么要那么多人领便当啊啊啊！！！
一代人无法磨灭的记忆。
早就读过了啊……
七本都看鸟。。
！！！还我弗雷特！！！……
故事结束时总会感到悲伤……
十年后才看到大结局。失望。
我的猫头鹰竟然晚了十年
狗尾~
呀。终于把7本哈利波特看完了。
再见 波特
7部Harry Potter，超乎儿童文学的构造庞大，虽然有bug，但是也辛苦罗琳了。伴随我和很多人一起长大的一部书，看完后

结局不喜欢
从第一部到第七部，从小看到大，陪了我这么多年了……
大概算是一个俗套而……美好的结局
没有原先疯狂的感觉了，只是顺着原先的惯性吧
2014.7.13～7.20
小时候读到废寝忘食的哇
星云雨果合集（下）
人物形象终于完整，可惜情节安排凌乱。
情节峰回路转，让人摸不着头脑，总体感觉罗琳的逻辑开始秀逗了
看过英文版之后就对中文版不那么感兴趣了……
悲情的斯内普。“那我的灵魂呢？邓布利多？我的呢”当属全书最有分量的话，沉重地压在我心上。
闹闹哄哄的，终于收场了，如释重负。
闭着目承认故事看完
2010年我最期盼的一部电影就是哈7，因为这么多年，他们长大了，我也长大了，我只想看看他们三个现在怎样了，我去到电
终于结束了。。
我读得太快了，完全是忽略细节在追剧情。这本书是该系列信息量最大的一本，罗琳阿姨写这本书算是一种大爆发，把这么多
益

全家是情圣
原作文字的力量远远超过电影的影像。爱与勇气能够战胜死亡，这种仿佛说腻了的陈腔滥调在经历了生离死别之后，随着少年
大结局
这本哈利波特是我唯一一本花了一年时间才读完的……因为中途是高考，加上我对哈利波特的感情在坠机般的下降
@20110802
从哈一到哈七。。。水平依次下降。。越写越不如前，作家都免不了这样的杯具么？
←哈利波特脑残粉！
每到暑假就忍不住重温一下。。
书确实比电影精彩多了，不过偶们不能拿书要求电影不是。。。后三部很好，链接成一个宏大的故事～
这就结束了啊...太感慨了。。。呃看完英文版再看中文就忍不住一直在挑刺。。。
补标。十年前吧。当时特别喜欢霍格沃茨保卫战，里面有太多太多的细节了，韦斯莱家真的是珍宝一样的存在。唯一的不解就
那是一个时代
比前六部烂多了
无声地夸大了一些东西的影响。
十九年后。
邓布列多是个计划通啊 和斯内普部署的计划一步一步实现 甚至牺牲自己去打败伏地魔 而哈利在这个过程中也着实很痛苦 因
15岁以前每年最开心事的就是等待哈利波特！
斯内普你这个傲娇怪.
就这么 就这么 完结了...
小学运动会，YZF同意周末借给我回家看，我超开心。午后读完，不舍合卷。
意犹未尽。几个比较意外的点：与德思礼一家分别时刻，达力意外的感恩；赫敏的S.P.E.W终于在最后一刻发挥了作用，多比

终于看完了，弥补青春的回忆，做一个勇敢正直的人。
终于重读完了，啊，小时候记忆深刻的点（which is哭点），仍然让人热泪盈眶啊，就是纳威，我是真的很高兴看到平凡的小
2000年出版，七年后完结。我却故意把整个故事的结局又拖了六年。不过，一切的冒险还是成长都终于结束了。Blessing u
补上这堂魔法课
终于看完了
出一本，买一本，每年暑假拿出来重温一遍。儿童文学的老师说，它不具有文学性，但那又怎样呢？斯内普真是大逆转啊~
作为一个结局，罗琳阿姨有一点用力过度了……
童年结束了。虽然第七部构思很奇妙，布局也很优秀，但是总觉得很遗憾，哪里缺了一点，反而没有小时候的感动。可能变的
补，我读书很少哭，但第一遍读到海格抱着哈利走在草地上的情节，不禁流泪。书上还留着十几年前的泪痕。
伟大之处在于西弗勒斯的塑造，魂器的呼应，魔幻界最精彩的战争之一，不输魔戒
=v=谢谢帮我买到它的人。熬夜也要看完的，很舍不得的故事。
死亡圣器的故事是完满的结局，那么多鲜活可爱的角色，那么多九曲回肠的传奇，伴随着自高自大丑态尽显的伏地魔找到了自
尼玛再吼一次，罗琳你TM少弄死个人，你书会卖不出去么！！
感谢罗琳女士带来的美妙幻想。
宏大的多卷魔咒书终于落幕了，每个人物都有详细且合理的交待安排，哈利也成家生子了。。。不知道会不会有续集呢？
|4497:2456|
我居然没有标记这本我已经读过?!
花了7年终于完结　JK罗琳把一个童话故事变成了现实小说...顺便缅怀我的青春小鸟一去不复还
大爱
2008-4-14 终于下决心去完结这个系列，其实已经没多大兴趣了，意义多点
2008-5-8 第十二章。看不下去了，等电影吧
这套书根本就是在宣扬扭曲的价值观吧
过了这么多年，终于读了。
陪伴着我长大的一个系列,看完之后一时间有点惆怅
“伤疤已经十九年没有疼过了，一切太平。”
某日晚,通宵.第二天微积分课闭上一会眼睛,被和蔼亲切的女老师询问"睡着了?"...
我一直喜欢快乐结局.但这样的结局没有心疼.
严歌苓的小说总是让我心疼.
ximeng说那么"罗嗦"的书你也看?
突然认识到这点.也许这就是为什么少了很多共鸣的原因.
但结局始终是结局,一定要看看的.
终于狠下心鼓起勇气把最后的大决战看了，果然电影还是只能自己在屋里下下来看，看书时候哭得那个造孽哟~
用重读缅怀流逝的十年
好神奇的书～～～看了书再看电影更有感觉～～～哈利系列都不错，可惜我这次又是慢热型的，看得也慢。。。
The end。
故事的最后爱上的是西弗勒斯·斯内普
翻译太烂。和上一本一样，还比不上网译版本。
LOOK AT ME kao!@#@$$
十年光阴

糟糕无比的翻译，真可谓每况愈下
和青春一起的回忆划上句号～～
故事的终结，哈利的十七岁和我的十七岁，一起结束了。我还记得中译版没出来我不自量力啃了几页原文，因为当时英文不济
跟了这么多年，哈利.波特系列总算尘埃落定。最后一本少了些悬念，最后的决战也比较仓促，不知是期望过高还是怎地。反正
非常喜欢的魔幻小说!
厄里斯魔镜——献给霍格沃兹最伟大的校长。
比混血王子让我满意多了 赫赫 还是值得读的
不希望结束
从童年一直延续至今的追踪终于有了一个答案～
最后的最后，好大团圆哪。那么多人不在了，但他们三个还是团圆了。
结局个人认为不好…不过作为流行青少小说，这样就够了。
直到最后的最后
少年时期读本
这个结局在侮辱我的智商
读过了之后有一点小失望。也许原版的能给人更美好的阅读感? 所以还是需要好好修炼英语==粉丝们猜对了结尾的大半，而过
结束
我不得不承认，那一只牝鹿把我感动了。
邪恶终究没有办法战胜正义
（T）一塌糊涂，无论读者抑或作者还是都需放平心态，将其当做儿童文学吧
这本书，它可以更好；但作者是有心无力还是心血都舍不得投了呢；除了斯内普教授谜底揭晓，没有亮点，结构和逻辑水准..
尽管结局你不能免俗哇
地铁上啃完，俨然没了十年前初见的marvelous 的感觉。
谁来和我聊聊这个故事……看完没人交流很痛苦
翻完。
还好，最后有点拖拉，在自圆其说。
终于完结了，看的心头一揪一揪。我心爱的双胞胎和斯内普。。。人物众多，看的有点晕，想构架起庞大的人物框架但是对人
终于让我给看到了..
Never end
喜欢。
向罗琳致敬。
糟透了
结束了。
情节很对得起读者。霍格沃茨之战激动人心。永远爱小天狼星，西弗勒斯，邓布利多。
唯独这个结尾觉得很莫名……
英文版看得真爽!
斯内普
受不了那谁和那谁还有那谁的死..
虽然颇有非议，终于是这样结束了。
不太满意的架构,不太满意的结局，可是我还是喜欢的。

为什么非要写点评，疼讯好烦。
为电影而创作...
终于读完最后一部了,可以说是一个了结.值.
什么破结局!
对于赫敏没有和马尔福在一起，我耿耿于怀。而且最后对马尔福德归宿寥寥带过，让我有点伤心哪~
读了英文的，不准备读中文的了
再折磨一次要疯了
结局..我我我不算太喜欢
Mark高三时偷偷看? 还有个同学看哭了
结局有点仓促，不过还好
8年时间，终于终结
一切回归平静。
果然没有看错斯内普。像亮司一样的男人。
什么都不用说了吧
7的结构果然跟之前很不相同~话说到了最后一本告诉我们其实这是一个毁灭7个龙珠消除一个愿望的故事是怎样=-=~以及不受
曾看了三遍 （虽然还是记不住名字= =
结束了结束了，不管怎么说终于结束了
高进同学给背回来的原版.....我还是没有等到小天狼星复活.....sigh
看完了，終於看完了。真的看完了，有點不相信。
哈利波特系列的最终升华!
这是一个读过青春的童话啊~
哈利都谢顶了~
高三看完，结局还是很温暖，导向正确。能写出来是多么幸福的事情。
结尾颠覆了儿童读物的感觉……
最惊险最传奇的一本
第7部，最后一部。
精彩，动人，忠于哈利的人会得到特别的感动
最后的一部，最爱的一部，最光明的一部，最不舍的一部。
我对它有执念，严重的执念。
从高中一直陪我到工作的第五年
终于写完了~很好看~我觉得后面写的比电影的好看~
任何结局对于哈利fan来说都不可能完美……
有点别扭的结局……
那个十九年。。。
伏地魔死的很简单啊。。大概是罗琳想就此停手吧 其实还有点为之惋惜呢~