Python 网络爬虫实习报告

Python 网络爬虫实习（报告）

目录

一、选题背景**.................................................................................... - 2 -**

二、爬虫原理**.................................................................................... - 2 -**

三、爬虫历史和分类**......................................................................... - 2 -**

四、常用爬虫框架比较**..................................................................... - 2 -**

五、数据爬取实战（豆瓣网爬取电影数据）**................................... - 3 -**

1 分析网页 ......................................................................................... - 3 -

2 爬取数据 ......................................................................................... - 3 -

3 数据整理、转换.............................................................................. - 4 -

4 数据保存、展示.............................................................................. - 9 -

5 技术难点关键点............................................................................ - 10 -

六、总结 **......................................................................................... - 13 -**

- 1 -



Python 网络爬虫实习（报告）

一、 选题背景

二、 爬虫原理

三、 爬虫历史和分类

四、 常用爬虫框架比较

Scrapy 框架:Scrapy 框架是一套比较成熟的 Python 爬虫框架，是使用

Python 开发的快速、高层次的信息爬取框架，可以高效的爬取 web 页

面并提取出结构化数据。Scrapy 应用范围很广，爬虫开发、数据挖掘、

数据监测、自动化测试等。

Crawley 框架:Crawley 也是 Python 开发出的爬虫框架，该框架致力于

改变人们从互联网中提取数据的方式。

Portia 框架:Portia 框架是一款允许没有任何编程基础的用户可视化

地爬取网页的爬虫框架。

newspaper 框架:newspaper 框架是一个用来提取新闻、文章以及内容

分析的 Python 爬虫框架。

Python-goose 框架：Python-goose 框架可提取的信息包括：<1>文章

主体内容;<2>文章主要图片;<3>文章中嵌入的任 heYoutube/Vimeo 视

频;<4>元描述;<5>元标签

- 2 -



Python 网络爬虫实习（报告）

五、数据爬取实战（豆瓣网爬取电影数据）

1 分析网页

# headers = {'User-Agent':'Mozilla/5.0(Windows NT

6.1) AppleWebKit/537.11 (KHTML, like Gecko)

Chrome/23.0.1271.64 Safari/537.11',

# 'Referer':None注#意如果依然不能抓取的话，这里

可以设置抓取网站的host

# opener = urllib.request.build\_opener()

# opener.addheaders = [headers]

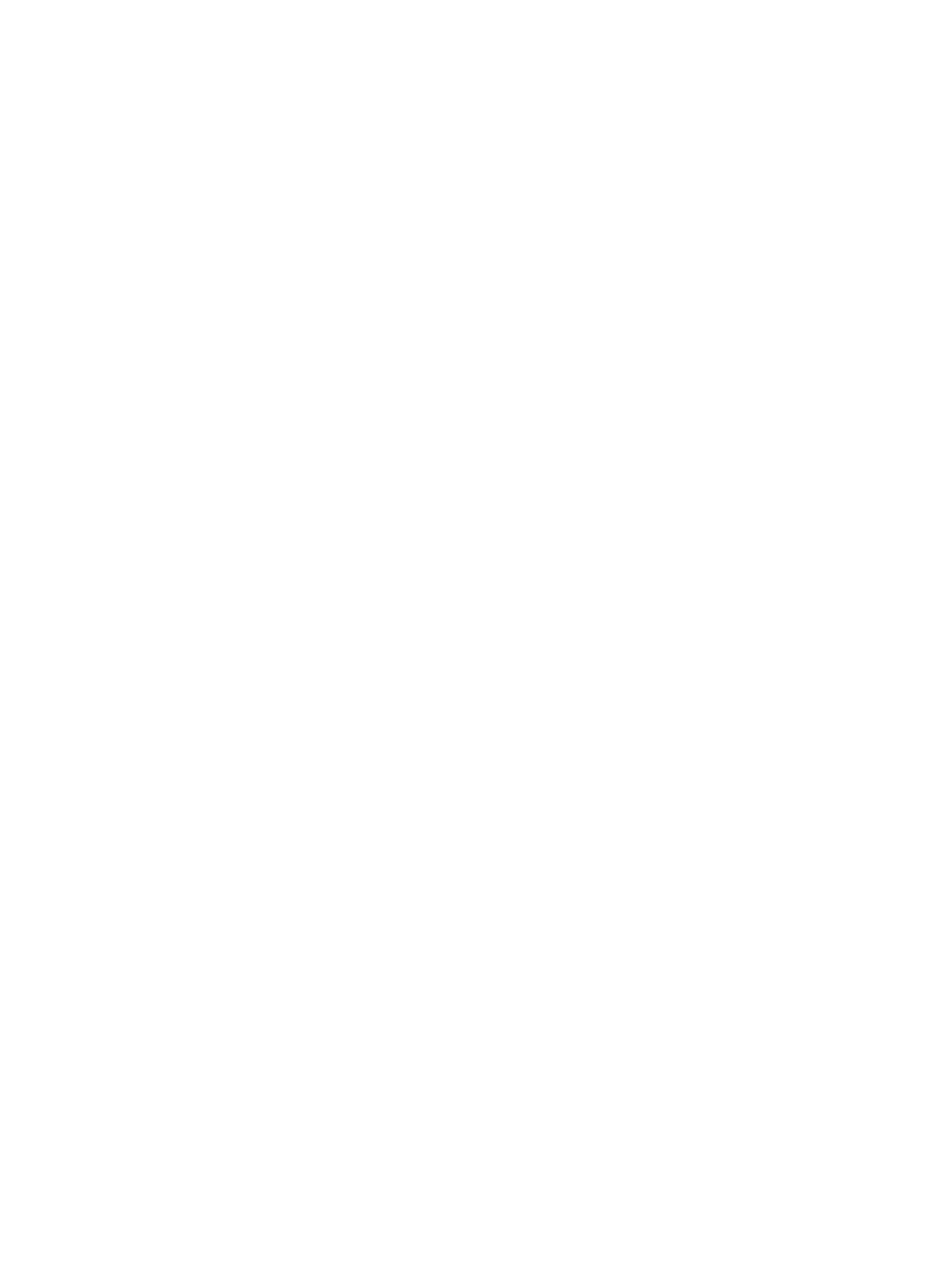
url = "https://movie.douban.com/top250?start=" +

str(pageSize) + "&filter=" + str(pageNum)

# data['html%s' %

i ]=urllib.request.urlopen(url).read().decode("utf-8")

- 3 -



Python 网络爬虫实习（报告）

data.append(urllib.request.urlopen(url).read().decode("utf-

return data

2 爬取数据

soup = BeautifulSoup(html, "html.parser")

for li in soup.find("ol", attrs={'class':

'grid\_view'}).find\_all("li"):

- 4 -



Python 网络爬虫实习（报告）

title.append(li.find("span", class\_="title").text)

#rating\_num.append(li.find("div",

class\_='star').find("span", class\_='rating\_num').text)

range\_num.append(li.find("div",

class\_='pic').find("em").text)

#spans = li.find("div",

#

rating\_people\_num.append(spans[x].string[-len(spans[x].stri

ng):-3])

class\_='pic').find("em").text)

if (li.find("div", class\_='pic').find("em").text

- 5 -



Python 网络爬虫实习（报告）

# print("aaa")

movie\_author.append(str[4:index])

#data['rating\_num'] = rating\_num

data['range\_num'] = range\_num

#data['rating\_people\_num'] = rating\_people\_num

data['movie\_author'] = movie\_author

return data

3 数据整理、转换

f = open('F://douban\_movie.html', 'w',encoding='utf-8')

f.write("<html>")

f.write("<head><meta charset='UTF-8'><title>Insert title

here</title></head>")

f.write("<h1>爬取豆瓣电影</h1>")

f.write("<h4> 作者：刘文斌</h4>")

- 6 -



Python 网络爬虫实习（报告）

f.write("<h4> 时间：" + nowtime + "</h4>")

f.write("<hr>")

f.write("<table width='800px' border='1' align=center>")

f.write("<thead>")

#f.write("<th width='50px'><font size='5' color=green>

评分</font></th>")

f.write("<th width='50px'><font size='5' color=green>排

名</font></th>")

#f.write("<th width='100px'><font size='5' color=green>

评价人数</font></th>")

f.write("<tr>")

f.write("<td

- 7 -



Python 网络爬虫实习（报告）

# f.write("<td

f.write("</thead>")

- 8 -



Python 网络爬虫实习（报告）

结果如后图所示：

- 9 -



Python 网络爬虫实习（报告）

5 技术难点关键点

数据爬取实战（搜房网爬取房屋数据）

# 设置编码方式

soup = BeautifulSoup(html, 'html.parser')

- 10 -



Python 网络爬虫实习（报告）

f.write("<head><meta charset='UTF-8'><title>Insert title

here</title></head>")

f.write("<center><h1>新房成交 TOP3</h1></center>")

f.write("<table border='1px' width='1000px' height='800px'

align=center><tr>")

f.write("<th><h2>房址</h2></th>")

f.write("<th><h2>成交量</h2></th>")

f.write("<th><h2>均价</h2></th></tr>")

for li in soup.find("ul",class\_="ul02").find\_all("li"):

name=li.find("div",class\_="pbtext").find("p").text

chengjiaoliang=li.find("span",class\_="red-f3").text

try:

junjia=li.find("div",class\_="ohter").find("p",class\_="gray-9

")#.text.replace('�O', '平方米')

junjia=li.find("div",class\_="gray-9")#.text.replace('�O', '

- 11 -



Python网络爬虫实习（报告）

平方米')

f.write("<td align=center><font size='5px'

color=blue>%s</font></td>" % chengjiaoliang)

f.write("<td align=center><font size='5px'

color=green>%s</font></td></tr>" % junjia)

print(name)

f.write("</table>")

f.write("</body>")

- 12 -



Python 网络爬虫实习（报告）

六、总结

教师评语：

成绩：

指导教师：

- 13 -



Python 网络爬虫实习（报告）

A 按年度计划分配率分配制造费用的方法适用于(季节性生产企业)

- 14 -

