目录

[1本微博爬虫的设计](#_Toc968)

[1.1本爬虫的模块](#_Toc6189)

[1.1.1基类设计](#_Toc17602)

[1.2模块的功能](#_Toc18596)

[1.3使用的技术和软件配置](#_Toc2891)

[1.4本爬虫项目的使用说明](#_Toc22240)

[2针对课程目标的完成度总结](#_Toc32307)

[3总结与心得](#_Toc23535)

# 引言

本项目的项目背景是基于是大数据环境下人们对信息获取方式的迭代以及对美食信息获取方式产生疑问。

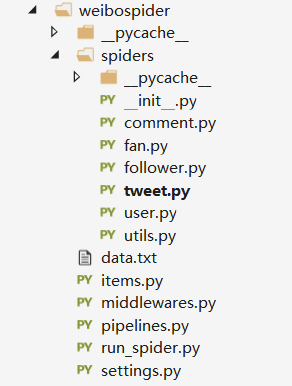
基于python语言进行的微博数据爬取，将杭州地区相关美食微博发博信息进行整理和完善。

主要内容是基于python语言进行的微博数据爬取，将杭州地区相关美食微博发博信息进行整理和完善

组内分工主要有：谢瑶瑶负责：一、代码的编译和任务分析；二、设计问题的解决方案；三、学习相关技术和确定技术路线；四、制定工作计划；五、根据结果完善项目信息、六、实践报告的制作。姚博涵负责：一、对“杭州美食”相关数据和文本的类型进行整理和筛选；二、基于数据和文本的选取编写进行python爬虫的源代码并进行调试；三、针对每次的结果进行整理和分析并反复进行修改和调试。

# 1本微博爬虫的设计

## 1.1本爬虫的模块

1.1.1基类设计

创建关键词搜索类，主要包括爬取微博的微博id,微博URL，微博发表时间，微博点赞数、转发数、评论数，微博内容，发表该微博的用户id和定位的经纬度信息等。

class TweetItem(Item):

"""Tweet information """

\_id = Field() # 微博id

weibo\_url = Field() # 微博URL

created\_at = Field() # 微博发表时间

like\_num = Field() # 点赞数

repost\_num = Field() # 转发数

comment\_num = Field() # 评论数

content = Field() # 微博内容

user\_id = Field() # 发表该微博用户的id

tool = Field() # 发布微博的工具

image\_url = Field() # 图片

video\_url = Field() # 视频

origin\_weibo = Field() # 原始微博，只有转发的微博才有这个字段

location\_map\_info = Field() # 定位的经纬度信息

crawl\_time = Field() # 抓取时间戳

## 1.2模块的功能

本系统主要运用的是关键词搜索爬取模块，所以主要是使用tweet.py模块，其他模块不予表述。



tweet.py文件中，定义TweetSpider(Spider)类，实行具体的爬取。

关键代码如下：

#### 1.2.1创建TweetSpider类，并使用https://weibo.cn站点进行爬取。

class TweetSpider(Spider):

name = "tweet\_spider"

base\_url = "https://weibo.cn"

…………

#### 1.2.2使用关键词”杭州美食“进行搜索，以及date\_start和date\_end进行微博发布时间的筛选，因为项目需要近两三年的微博信息，所以可以把时间设置成2018-01-01和2021-01-01.

def init\_url\_by\_keywords():

keywords = ['杭州美食']

date\_start = datetime.datetime.strptime("2018-01-01", '%Y-%m-%d')

date\_end = datetime.datetime.strptime("2021-01-01", '%Y-%m-%d')

…………

while date\_start < date\_end:

next\_time = date\_start + time\_spread

urls.extend(

[url\_format.format(keyword, date\_start.strftime("%Y%m%d"), next\_time.strftime("%Y%m%d"))

for keyword in keywords]

)

date\_start = next\_time

return urls

…………

#### 1.2.3筛选有经纬度信息的微博，若微博有经纬度信息，则继续爬取此微博的各类信息

map\_node = tweet\_node.xpath('.//a[contains(text(),"显示地图")]')

if map\_node:

if "来自" in create\_time\_info:

tweet\_item['created\_at'] = time\_fix(create\_time\_info.split('来自')[0].strip())

tweet\_item['tool'] = create\_time\_info.split('来自')[1].strip()

else:

tweet\_item['created\_at'] = time\_fix(create\_time\_info.strip())

like\_num = tweet\_node.xpath('.//a[contains(text(),"赞[")]/text()')[-1]

tweet\_item['like\_num'] = int(re.search('\d+', like\_num).group())

repost\_num = tweet\_node.xpath('.//a[contains(text(),"转发[")]/text()')[-1]

tweet\_item['repost\_num'] = int(re.search('\d+', repost\_num).group())

comment\_num = tweet\_node.xpath(

'.//a[contains(text(),"评论[") and not(contains(text(),"原文"))]/text()')[-1]

tweet\_item['comment\_num'] = int(re.search('\d+', comment\_num).group())

images = tweet\_node.xpath('.//img[@alt="图片"]/@src')

if images:

tweet\_item['image\_url'] = images

videos = tweet\_node.xpath('.//a[contains(@href,"https://m.weibo.cn/s/video/show?object\_id=")]/@href')

if videos:

tweet\_item['video\_url'] = videos

map\_node = tweet\_node.xpath('.//a[contains(text(),"显示地图")]')

if map\_node:

map\_node = map\_node[0]

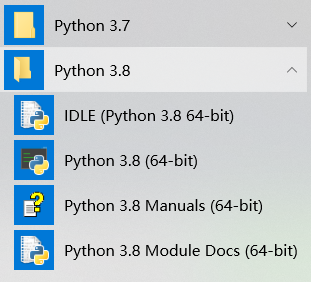
map\_node\_url = map\_node.xpath('./@href')[0]

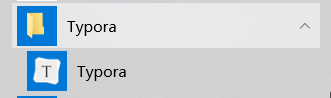
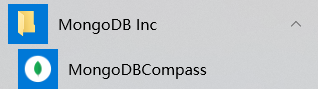
map\_info = re.search(r'xy=(.\*?)&', map\_node\_url).group(1)

tweet\_item['location\_map\_info'] = map\_info

## 1.3使用的技术和软件配置

本项目适用的Python版本为Python3.6及以上

  
 使用的软件有Mongodb和Typora



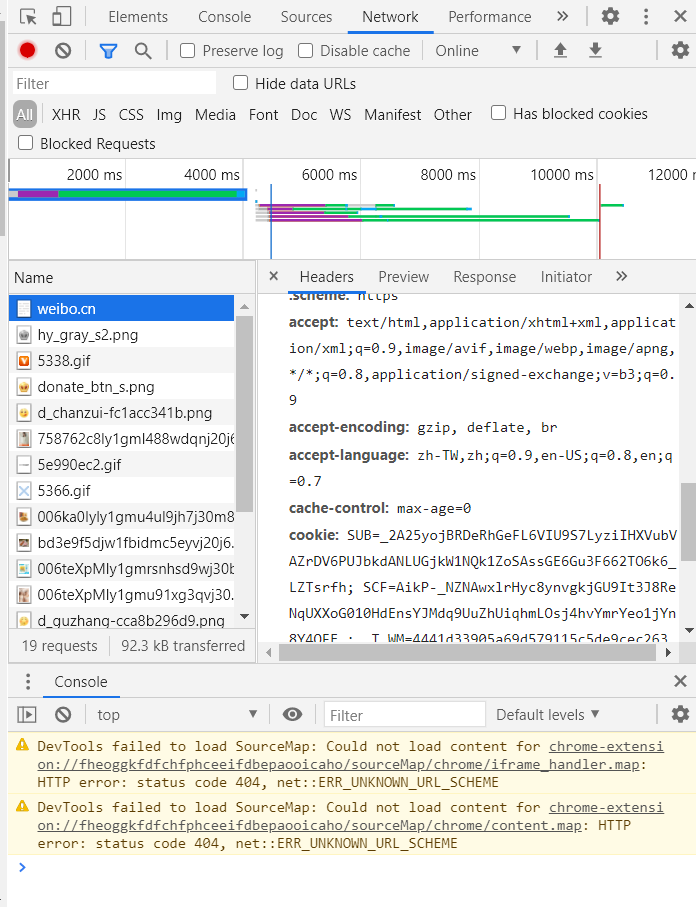
## 1.4本爬虫项目的使用说明

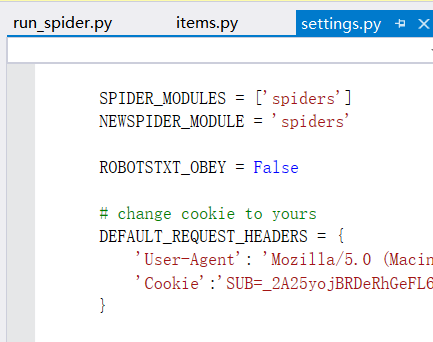
出了安装上述软件之外，如若需要在其他电脑上操作本爬虫，你还需要：

首先，将https://weibo.cn站点的微博账户登录。



然后拿到此页面的cookie，将weibospider的settings.py文件中的Cookie替换。

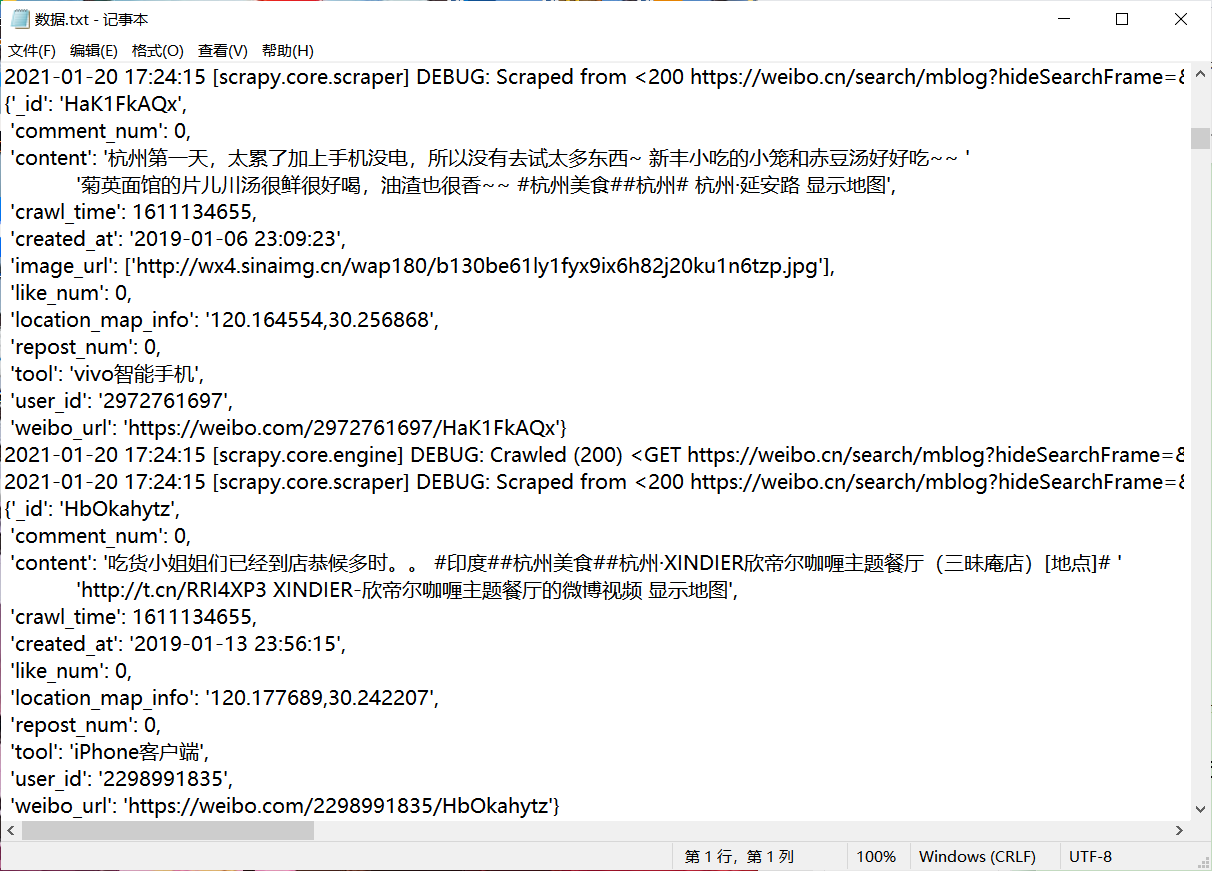




如若爬虫运行出现403或302，则账号被封或cookie失效。

替换好后，执行python run\_spider.py tweet命令即可开始爬虫。





PS：如果一段时间程序没用反应但无报错，请耐心等待。等待5-10分钟无微博内容爬出，可检查自己账号或网络是否出错。

# 2针对课程目标的完成度总结

课程目标1：掌握对实际问题进行系统分析和设计解决方案的能力，具体可以包括：需求分析、数据库设计、交互界面设计、算法设计、模块设计等。

总结：本项目名称为”基于微博数据的杭州美食地图和推荐服务”,而项目内容主要分为：位置微博数据爬取、优质餐饮店识别，餐饮口味分布展示、个性化的实时餐饮推荐。

本小组负责位置微博数据爬取内容。需要面对的主要问题是未接触过python语言和爬虫学习。针对这俩个实际问题，我们小组经过系统分析后制定了学习方案和资料整理方案以及代码设计方案。

针对代码设计方案，由于位置微博数据爬取系统，主要是靠tweet即关键词模块进行，所以设计解决方案时，主要是对关键词模块进行设计，关注爬虫的一些算法和需要注意的站点。

课程目标2：能够带着问题对相关领域进行文献查阅，学习相关技术并确定技术路线。

总结：微博爬虫目前主要有利用API端口爬取和直接使用网页源爬取两种。对于如何实现爬虫项目。我们首先确认了目前爬虫的语言，大部分都是基于python，即需要制定python语言的学习计划。其次，在将API和网页源两种爬虫的优缺点比对后，我们选择了学习并编译网页源爬虫的技术路线。在查阅CSDN、GITHUB等优秀的程序员社区的爬虫有关资料后，我们选择了廖雪峰老师的python课程进行学习，并借阅了有关的优秀代码，确定了技术路线。

课程目标3：能够与队友、导师、以及项目可能相关的其他人员进行沟通。

总结：每周一至周五，本小组成员都会针对各自分工进行学习和问题汇总；而周六周日，组员之间则会对积累下的问题进行商议及解决，并制定下周计划。如若彼此遇到不能解决的问题，则会咨询带项目的导师以及研究生学姐。其次，由于爬虫收集的资料也是其他两组所需的，尤其是第二组，我们还组织了一次组与组之间的交流，协助其他组，共同提高本项目的完成度。

课程目标4：具备完成各类技术和非技术文档的能力。

总结：在整理资料方面，提取爬虫爬取的内容到txt，或是md文件中。md文件是当下十分好用的文件，在IT行业使用率最高，使用的语言是Markdown语言，对于编纂和整理资料有着非常大的帮助。

课程目标5：能够制订详细的工作计划，并能严格按照计划完成每一个小任务。

总结：自项目开始，本小组就制定了详细的工作计划。比如在第一个月内学习python语言，其中周一到周五是组员自我学习的时间，而周六周日则会将自己遇到的难题互相沟通。在第一个月过后，掌握了python语言后，本小组开始进行网络爬虫的学习和环境基础安装。第二个月第一周，组内人员集中对爬虫需要的软件和安装环境进行了统一。第二个月第二周到第四周，开始收集爬虫资料和课程，进行学习。第三个月第一、二周，开始编写爬虫代码。

组内人员都对自己的学习计划进行了详尽的规划，并严格按照规划完成了任务。