

Parse_html

2017年3月25日 14:52

大家正在搜：山东公安

微博

首页 视频 发现 游戏 啊啊啊啊啊啊03...

#-*- coding: utf-8 -*-
import re
from bs4 import BeautifulSoup
soup = BeautifulSoup(open('../docs/html.txt'), 'lxml')

吕利娜
她还没有填写个人简介

+ 关注 私信

她的主页 她的相册

171 关注 235 粉丝 815 微博

二维码
扫描二维码，可以用手机访问本页
下载

勋章信息

等级信息
当前等级：Lv 11 经验值：1140
距离升级经验值：240
查看详情 >

基本信息
昵称：吕利娜
所在地：河北 唐山
性别：女
注册时间：2011-03-13

标签信息
标签：唐山生活

base_info =
soup.select('.WB_innerwrap')
for li in base_info[6].find_all('li'):
for span in li.find_all('span'):
print span.string.strip()

a_pattern = re.compile('.*span>((?:.|\n)*?)(.*)((?:.|\n)*?)', re.M)
for a in base_info[7].find_all('a'):
a_content = a_pattern.findall(str(a))[0]
print a_content[2].strip()

level_info = soup.select('.level_info')[0].select('.info')
for info in level_info:
print info.select('.S_txt1')[0].string.strip()

badge_item_pattern = re.compile('.*alt="(.*?)".*height.*medalcard="(\\d*)".*', flags = 0)
for badge_item in soup.select('.badge_item'):
print badge_item_pattern.findall(str(badge_item.img))[0][0],
badge_item_pattern.findall(str(badge_item.img))[0][1]

微博精彩
热门微博 热门话题
名人堂 微博会员
微相册 微游戏
微指数

手机玩微博
扫描二维码下载，更多版本
点这里

认证&合作
申请认证 开放平台
企业微博 链接网站
微博标识 广告服务
微博商学院

微博帮助
常见问题
自助服务

微博客服 意见反馈 舞弊举报 开放平台 微博招聘 新浪网导航 举报处理大厅 中文(简体)

京ICP证100780号 互联网药品服务许可证 互联网医疗保健许可证 京网文[2014]2046-296号 京ICP备12002058号 增值电信业务经营许可证B2-20140447
Copyright © 2009-2017 WEIBO 北京微梦创科网络技术有限公司 京公网安备11000002000019号