中国科学院大学
University of Chinese Academy of Sciences

# Python结课论文

| | |
|---|---|
| 论文题目： | 基于Python-网络爬虫的新闻分析 |
| 学　　校： | 长春光学精密机械与物理研究所 |
| 年　　级： | 2019级 |
| 组　　长： | 董明远 |
| 组　　员： | 崔旭；李鑫鹏；李瑞龙；徐飞 |
| 邮　　箱： | 328948975@qq.com |
| 指导教师： | 韩霄松 |
| 时　　间： | 2019 年 11 月 20 日 |

# 基于Python-网络爬虫的新闻分析

**摘要**：在如今这个信息爆炸时代，困难的不是获取信息，而是从网络上的海量信息中快速、准确的找到自己想要的有效信息 ，网络爬虫就可以便捷有效的帮助到我们。使用Python编写网络爬虫，爬取网页中自己想要的信息，再通过文本分析提取、整理自己所获得的信息，是我们本次工作想要做到的。该工作实现了对长春光机所官网综合新闻中新闻的的提取、分析和显示，将新闻按日期分成了多个目录，并统计其新闻数量，以一周为一个时间单元统计了每个部门的平均新闻数量以及各个部门在总新闻数量中的占比，并做了箱线图，方便读者观察和提取信息。

**关键词**：Python；网络爬虫；箱线图；词云

# Campus Network News Analysis Based on Python-WebCrawler

**Abstract**:In today's information explosion era, what is difficult is not to obtain information, but to quickly and accurately find the effective information we want from the massive information on the Internet, so that the web crawler can help us conveniently and effectively.Using Python to write web crawler, crawl to get the information you want in the web page, and then extract and sort out the information you get through text analysis, is what we want to do in this work.The job is realized with changchun light machine to a comprehensive news in the news website of extraction, analysis and display, will be divided into multiple directories news by date, and statistics of the number of news, in a week for a time unit statistics the average number of news of each department and each department in the proportion in the total number of news, and do the boxplot, for easy observation and information extraction.

**Key words**: Python; Web crawler; Box line diagram; Word Cloud

# 1. 引言

随着高新技术的普及,互联网成为人们获取信息的主要来源。从海量文字中迅速提取新闻的关键信息，可以帮助人们更快的得到某个特定时间段发生的重要事件，通过高频词加快人们对某条特定新闻的查找速度，并将不同年份、地区、专业部门等的信息进行横向比较，便利人们生活并协助各部门的办公工作。实现这样一个系统需要网络爬虫技术以及文本分析技术。

网络爬虫技术是一种按照一定的规则模拟用户浏览的方式自动的抓取网站信息的程序或者脚本，相比人工统计有更高的效率、更高的准确度和很好的规范性，避免错误的发生，在各个方面减少人为工作量，为之后的数据处理做准备。另外一些不常使用的名字还有蚂蚁，自动索引，模拟程序或者蠕虫。文本分析是指对文本的表示及其特征项的选取，把从文本中抽取出的特征词进行量化的表示。词云是一种典型的用于文本分析的可视化形式，区别于传统意义上的统计图表，具有更加美观的可视化效果和更好的实用性。

此系统将网络爬虫技术与文本分析技术相结合，提取长春光机所综合新闻网站的部分新闻的关键信息、关键词，并通过词云技术实现数据的可视化。

# 2. 网络爬虫模块

## 2.1 概述

网络爬虫（又称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。另外一些不常使用的名字有蚂蚁、自动索引、模拟程序或者蠕虫。期是搜索引擎抓取系统的重要组成部分，主要目的是将互联网上的网页下载到本地形成一个或联网内容的镜像备份。其结构分为三个部分：控制器、解析器和资源库。

## 2.2 网络爬虫基本流程

本组网络爬虫包括信息爬取子模块（ciompSpider.py）和信息处理子模块（textProcess.py），由main.py作为入口调用子模块完成新闻爬取的任务；

首先，通过信息爬取子模块对新闻列表进行爬取，获得全部新闻标题、新闻时间和新闻url信息，爬取过程中对新闻url的不同情况进行处理；

然后，根据新闻url，对新闻内容和发布部门进行爬取，并调用信息处理子模块对新闻标题、新闻部门进行处理；

最后，将新闻内容以txt格式按日期存放在不同文件夹中，将新闻总表存放在xlsx文件中。

## 2.3 信息爬取子模块

此模块利用python的urllib模块，根据url提取网页html，利用python的bs4模块对html中所需信息进行提取。将实现信息爬取的方法抽象成ciompSpider类，该类包括__init__函数，get_html函数，parse_mainlist函数，get_mainlist函数，get_news函数和process函数。

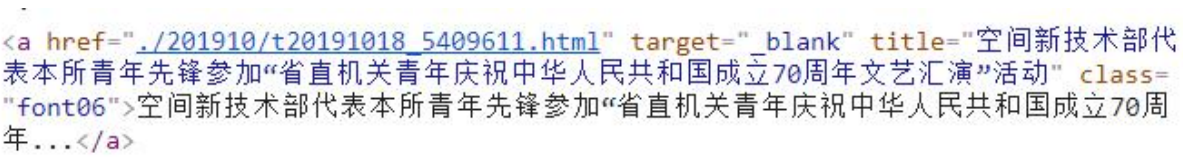
a）__init__函数：包含网页源url和新闻信息存储根目录

b）get_html函数：获取网页html

c）parse_mainlist函数：爬取单页新闻列表

经过对网页的分析，确定利用bs4中select方法，以‘.font06’为条件对html文件包含新闻标题和新闻链接的标签进行提取，以‘.riqi’为条件对新闻发布日期进行提取。

在提取新闻标题的过程中，发现以取标签内容的方式提取标题，长标题会出现带省略号的现象，如下图：




所以，选择以取标签‘title’属性的方式来提取标题，避免了标题不完整的情况。

在提取新闻链接的过程中，发现新闻链接的格式并不统一，以单一方式进行拼接得出的绝对链接会存在错误。经过对初次爬取结果进行分析后，发现新闻链接格式存在4种格式，具体见下表：

| **Href属性格式举例** |
| --- |
| ./201907/t20190708_5336957.html |
| ../../zt/kjcg/zonghexinwen/zonghexinwen_son/201906/t20190624_5327844.html |
| ../yw/201907/t20190702_5331011.html |
| http://www.ccb.cas.cn/xwzx2015/zhxw2015/201902/t20190225_5244377.html |

在对4种方式分别处理后，可得到全部新闻的绝对链接，用于具体新闻内容的爬取。

a）get_mainlist函数：爬取新闻总表

由于长春光机所网站网页的页码部分是通过JavaScript动态生成，使用常规方法无法从html中直接获取，在对新闻列表不同列表的url后，除第一页外，各页链接符合如下格式：

http://www.ciomp.cas.cn/xwdt/zhxw/index_(页码-1).html

同时发现JavaScript中var countPage变量用于存放总页数，利用正则串：

var countPage = \d+

则可获得网页总页数，以此作为迭代器，拼接新闻列表各页url，循环调用get_html函数和parse_mainlist函数，爬取获得新闻总表

b）get_news函数：依据获取的新闻总表中各条新闻的绝对url获得新闻内容和新闻发布部门

经过对新闻页面的分析，确定利用bs4中select方法，以‘td[width="22%"]’为条件对部门进行提取，以‘p’为条件对新闻内容进行提取。

在以获取的文章标题作为文件名存储过程中，发现部分标题存在文件名规定的非法字符，同时爬取的原始部门数据形式非常复杂，不利于后续数据的分析。所以，在此函数中引入信息处理子模块，对文件名和新闻部门进行处理，具体方式于2.4节进行说明

c）process函数：信息爬取子模块执行入口

实现对长春光机所网页的信息爬取后，我们发现长春光机所新闻综合网站只容纳1000条新闻，共38页，每页最多容纳27条，最老的新闻会因新闻更新而在列表中消失，但若提前爬取得到过该新闻链接，依旧可以访问该新闻的具体页面。

## 2.4 信息处理子模块

在爬取长春光机所综合新闻网站的原始数据后，经初步分析，少量新闻标题存在‘?\<>’等非法字符，导致以标题字符串命名txt文件名时报错。从新闻中爬取的部门信息非常混乱，根本无法达到按部门统计新闻的要求，所以，我组对标题及部门信息进行了深入分析，编写了信息处理子模块，完成了标题和部门信息的数据清洗，成功存储了全部1000条新闻内容，为后续数据可视化提供了较为准确的基础数据。

此模块主要完成对字符串的处理，将信息处理的方法抽象为textProcess类，该类包括__init__函数，title_process函数，dep_norm_process函数，dep_process函数。

a) __init__函数：定义字典变量，关键词删除、增补列表

b) title_process函数：标题处理函数

b.1:处理标题内含有 标签的情况

利用正则串：

regex = r’</?P>’

因某些标题内含有两个 标签，所以标题进行两次匹配去除，第一次直接删除，第二次用空格替代，这样既可以去掉标签，还可以保证标题被正常分隔。此方法对只含有一个 标签的标题，第二次替代不会生效，不影响处理结果。

b.2:处理其他非法字符

标题内含有的非法字符包括‘/’和‘?’两种，分别以‘、’和‘？’（中文格式）替代。

c) dep_norm_process函数：部门标准化函数

将爬取部门中的部门缩写、英文大小写等情况进行标准化，防止出现同一部门，两种名称的情况，标准化列表见下表。

| 原始部门名称 | 标准化部门名称 |
|---|---|
| 所办 | 所长办公室 |
| 党办 | 党委办公室 |
| light | Light |
| 国合处 | 国际合作处 |
| 人力处 | 人力资源处 |
| 质量处 | 质量管理处 |
| 监审处 | 监察审计处 |
| 成果处 | 成果转化处 |
| 保密处 | 保密管理处 |
| 图像室 | 图像部 |
| 无人飞行器 | 无人飞行器部 |
| 孵化器公司 | 孵化器 |
| 希达 | 长春希达 |
| 科宇公司 | 科宇物业 |
| 奥普质管部 | 奥普公司 |

d)dep_process函数：部门匹配函数

此函数是清洗部门数据信息的核心函数，具体功能分为匹配字典生成和部门匹配两部分。

d.1:匹配字典生成

step1：

读取部门原始数据，使用str.split()函数进行划分，取划分后结果列表长度为2的生成字典。即将‘研究生部 陈天宝’转化为字典{‘研究生部’：[‘陈天宝’]}。key值相同的则扩充value列表，key值不同则建立新的字典元素。

Step2：

分析字典后，去掉姓名在前，部门在后的情况。删除列表定义在__init__函数中，具体如下。

```
self.del_list = ['张译心', '张凌童', '王启东', '周哲', '张财华', '荆雷', '常唯', '王浩泽', '刘艳']
```

step3：

初步分析结果，在字典中加入已知缺少的部门。增补列表定义在__init__函数中，具体如下：

```
self.add_dic = {'航测部':'', '空间三部':'', '无人飞行器':'', '希达':'', '长光智欧':'', '科宇物业':''}
```

最后，生成匹配字典，字典格式为{‘部门名称’：[成员列表]}

d.2:部门匹配

在生成匹配字典后，即可针对混乱的原始部门数据进行匹配清洗。

Step1：特殊情况处理

在对名单进行分析后，发现含有‘信息中心’和‘朱立禄’的情况非常复杂，对于这两种情况进行单独处理，即匹配到‘信息中心’则将部门改为‘信息中心’，匹配到‘朱立禄’则将部门改为‘所长办公室’，且‘朱立禄’前属于信息中心，后属于‘所长办公室’，则在if分支中，将‘信息中心’的匹配更改置于‘朱立禄’之前，则可避免匹配结果发生错误。

Step2：常规情况处理

在对每个部门字符串进行split()分隔后，针对分隔列表与匹配字典的key进行比对匹配，若匹配到，则用key值替换部门信息。

若未匹配到，一般是出现了以下四种情况：

- 部门与姓名间无空格

  在字符串中查找匹配字典key值，若找到，则用key值替换部门信息。

- 只有姓名

  在字典value值中查找姓名，若找到，则用对于key值替换部门信息。

- 部门信息为空

  在标题信息中查找匹配字典key值，若找到，则将key值写入部门信息。

- 其他情况

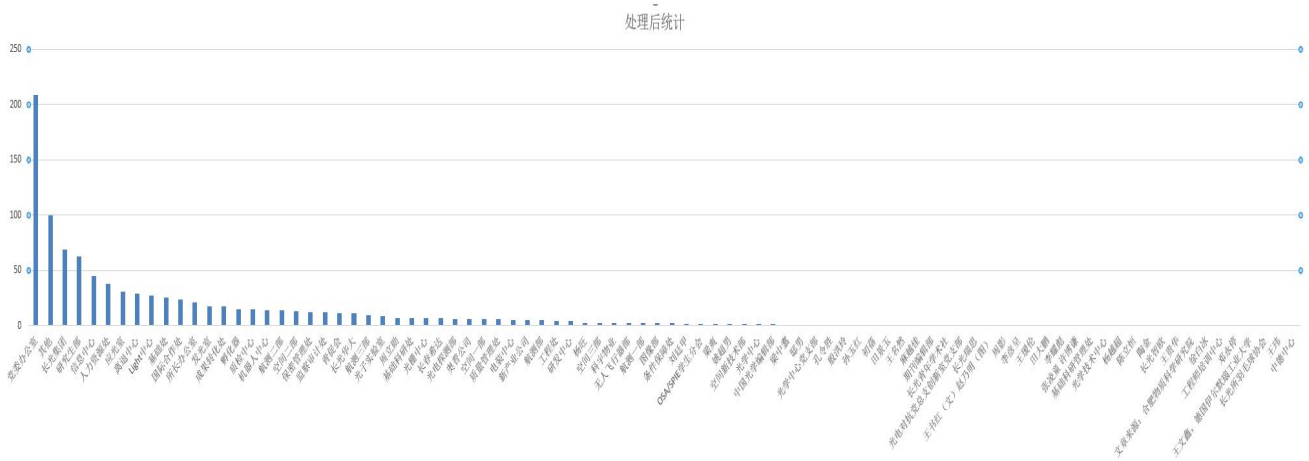
  无法处理，保持原部门信息。

Step3：部门名称标准化及空值处理

将dataframe中的空值填充为‘其他’并对所有部门名称进行标准化处理。

标准化处理放于最后，而不在匹配字典生成时就进行标准化，是因为step2中在标题信息中查找匹配字典key值的情况下，标题中也可能出现部门缩写等情况，为能匹配到更多的部门信息，所以将标准化工作放于最后。

处理前对部门进行统计，除空项以外，部门总数为350个，根本无法依据原始数据进行部门分类。处理后部门总数88个，其中包含未识别的姓名29个，且绝大多数姓名对应文章数不超过3篇，根据此特点，在进一步数据分析时，还可做进一步处理，使按部门分类更为准确。处理后按部门分类的新闻总数柱状图如下图：




## 2.5 程序运行结果

程序运行结果如下图：


生成的文件包括news_index_origin.xlsx，news_index.xlsx，按日期归类的新闻内容txt文件，统一存放在news文件夹中。

爬取后的原始新闻总表（news_index_origin.xlsx）如下图：

| | title | department | date | url |
|---|---|---|---|---|
| 0 | 宾伯格教授荣获德国物理学会最高实验物理奖章 | 中德中心 秦宇 | 2019-11-19 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191119_5438737.html |
| 1 | 光学中心学生党支部开展环卫工人送温暖活动 | 研究生部 陈天宝 | 2019-11-19 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191119_5438731.html |
| 2 | 九三学社长春光机所委员会获九三中央表彰 | 党委办公室 刘丽玫 | 2019-11-19 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191119_5438647.html |
| 3 | 航测部学生党支部开展义务扫雪活动 | 研究生部 马铭阳 | 2019-11-18 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191118_5437982.html |
| 4 | 长春光机所举行2019年研究生应急疏散及灭火演练 | 党委办公室 陆海龙 | 2019-11-18 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191118_5437891.html |
| 5 | Light中心党支部开展“献衣物，送爱心”活动 | Light中心 丁帅 | 2019-11-16 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191116_5437393.html |
| 6 | 国家光栅工程中心学术论坛成功举办 | 王玮 | 2019-11-13 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191113_5430808.html |
| 7 | 研究生学生党支部举办致敬大疖先生活动 | 研究生部 杨依凡 | 2019-11-11 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191112_5430226.html |
| 8 | 中国科学院长春软件测评中心顺利通过国家认可委员会实验室能力验证 | 质检中心 哈清华 | 2019-11-11 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191111_5428504.html |
| 9 | 研究生部举办人文系列讲座——当前国际政治格局分析 | 研究生部 韩爽 | 2019-11-07 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191107_5424679.html |
| 10 | 应用光学国家重点实验室第45期应光论坛成功举办 | 曹乃亮 | 2019-11-07 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191107_5423362.html |
| 11 | 航测三部举办非线性控制专题学术交流活动 | 航测三部 | 2019-11-04 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191104_5420323.html |
| 12 | 长春光机所参加全国科学院联盟理事会第八次全体会议 | 长光集团 高晶 | 2019-11-02 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191102_5419562.html |
| 13 | 吉林省光学学会第八次会员代表大会暨2019年学术年会胜利召开 | 应光室 徐姝睿 | 2019-11-02 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191102_5419561.html |
| 14 | 长春光机所举办机械设计精品培训班 | 人力资源处 王欢 | 2019-10-31 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191031_5416146.html |
| 15 | 2018级硕士生召开开题考核经验分享会 | 研究生部 陈伟帅 | 2019-10-30 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191030_5414749.html |
| 16 | 长春光机所管理系统年轻干部第三理论学习小组正式成立 | 唐奇 | 2019-10-29 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191029_5413726.html |
| 17 | 长春光机所举办“国家重大科研仪器研制项目验收与档案管理”培训 | 基础处 陈惠颖 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413320.html |
| 18 | 长春光机所举办“国际专利申请”专题培训 | 发光室 宋悦 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413200.html |
| 19 | 航测一部年轻干部理论学习小组成立 | 航测一部 史文欣 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413199.html |
| 20 | 光栅中心党支部上党课深入学习“八个明确”和“十四个坚持” | 光栅中心 糜小涛 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413197.html |
| 21 | 长春光机所“光明前行”年轻干部联合理论学习小组成立 | 发光室 陈星 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413193.html |
| 22 | 保密处党支部开展党课学习和主题党日活动 | 保密管理处 李阳 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413185.html |
| 23 | 应光室党支部召开年轻干部理论学习小组成立大会 | 应光室 姚舜 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5412923.html |
| 24 | Light中心召开年轻干部理论学习小组成立大会 | Light中心 袁培江 | 2019-10-26 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191026_5412790.html |

处理后的新闻总表（news_index.xlsx）如图：

| | | title | department | date | url |
|---|---|---|---|---|---|
| 2 | 0 | 宾伯格教授荣获德国物理学会最高实验物理奖章 | 中德中心 | 2019-11-19 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191119_5438737.html |
| 3 | 1 | 光学中心学生党支部开展环卫工人送温暖活动 | 研究生部 | 2019-11-19 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191119_5438731.html |
| 4 | 2 | 九三学社长春光机所委员会获九三中央表彰 | 党委办公室 | 2019-11-19 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191119_5438647.html |
| 5 | 3 | 航测部学生党支部开展义务扫雪活动 | 研究生部 | 2019-11-18 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191118_5437982.html |
| 6 | 4 | 长春光机所举行2019年研究生应急疏散及灭火演练 | 党委办公室 | 2019-11-18 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191118_5437891.html |
| 7 | 5 | Light中心党支部开展“献衣物，送爱心”活动 | Light中心 | 2019-11-16 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191116_5437393.html |
| 8 | 6 | 国家光栅工程中心学术论坛成功举办 | 王玮 | 2019-11-13 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191113_5430808.html |
| 9 | 7 | 研究生学生党支部举办致敬大街先生活动 | 研究生部 | 2019-11-11 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191112_5430226.html |
| 10 | 8 | 中国科学院长春软件测评中心顺利通过国家认可委员会实验室能力验证 | 质检中心 | 2019-11-11 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191111_5428504.html |
| 11 | 9 | 研究生部举办人文系列讲座——当前国际政治格局分析 | 研究生部 | 2019-11-07 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191107_5424679.html |
| 12 | 10 | 应用光学国家重点实验室第45期应光论坛成功举办 | 应光室 | 2019-11-07 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191107_5423362.html |
| 13 | 11 | 航测三部举办非线性控制专题学术交流活动 | 航测三部 | 2019-11-04 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191104_5420323.html |
| 14 | 12 | 长春光机所参加全国科学院联盟理事会第八次全体会议 | 长光集团 | 2019-11-02 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191102_5419562.html |
| 15 | 13 | 吉林省光学学会第八次会员代表大会暨2019年学术年会胜利召开 | 应光室 | 2019-11-02 | http://www.ciomp.cas.cn/xwdt/zhxw/201911/t20191102_5419561.html |
| 16 | 14 | 长春光机所举办机械设计精品培训班 | 人力资源处 | 2019-10-31 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191031_5416146.html |
| 17 | 15 | 2018级硕士生召开开题考核经验分享会 | 研究生部 | 2019-10-30 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191030_5414749.html |
| 18 | 16 | 长春光机所管理系统年轻干部第三理论学习小组正式成立 | 国际合作处 | 2019-10-29 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191029_5413726.html |
| 19 | 17 | 长春光机所举办“国家重大科研仪器研制项目验收与档案管理”培训 | 基础处 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413320.html |
| 20 | 18 | 长春光机所举办“国际专利申请”专题培训 | 发光室 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413200.html |
| 21 | 19 | 航测一部年轻干部理论学习小组成立 | 航测一部 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413199.html |
| 22 | 20 | 光栅中心党支部上党课深入学习“八个明确”和“十四个坚持” | 光栅中心 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413197.html |
| 23 | 21 | 长春光机所“光明前行”年轻干部联合理论学习小组成立 | 发光室 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413193.html |
| 24 | 22 | 保密处党支部开展党课学习和主题党日活动 | 保密管理处 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5413185.html |
| 25 | 23 | 应光室党支部召开年轻干部理论学习小组成立大会 | 应光室 | 2019-10-28 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191028_5412923.html |
| 26 | 24 | Light中心召开年轻干部理论学习小组成立大会 | Light中心 | 2019-10-26 | http://www.ciomp.cas.cn/xwdt/zhxw/201910/t20191026_5412790.html |

爬取新闻的新闻内容如图：


# 3. 数据分析和结果

## 3.1 jieba分词

中文文本需要通过分词获得单个的词语，Jieba 是一种基于 python 的常用的中文分词工具，可对一段文字中的词汇进行分词。jieba库提供三种分词模式，即精确模式、全模式和搜索引擎模式。精确模式：把文本精确的切分开，不存在冗余单词；全模式：把文本中所有可能的词语都扫描出来，有冗

余；搜索引擎模式：在精确模式基础上，对长词再次切分。本文使用精确模式，用jieba库自带函数jieba.cut(s)实现。

首先对文本进行处理，去掉一些分割字符

```
20      content = content.replace('\\', '').replace(' ', '').replace('\r\n', '').replace('[', '').replace(']', '').replace(
21          '"', '').replace("'", '')
```

然后通过jieba.cut进行分词，存到wordlist

```
26      word_list = ' '.join(jieba.cut(mydic[date]))
```

## 3.2 wordcloud制作词云

wordcloud是优秀的词云展示第三方库，以词语为基本单位，通过图形可视化的方式，更加直观和艺术的展示文本。

从wordlist读取数据，设置词云图片的参数，并将结果按日期进行存储

```
27      wordcloud = WordCloud(font_path='simhei.ttf',
28                            width=800,
29                            height=600,
30                            background_color='white').generate(word_list)
31      # plt.imshow(wordcloud)
32      # plt.show()
33      wordcloud.to_file(path.join(r'D:\PycharmProjects\ciomp_spider\ciomp_spider\cloudword', date + ".png"))
```

词云如下图所示：

图 3.1 词云（2019.11.19）

## 3.3 曲线图

首先使用pandas读取表中部门和日期信息，再使用unique函数查询处理后的部门的个数，一共有57个部门，并写入新的表格中。如下图所示：

```
df = pd.read_excel('E:/1111/news_index.xlsx', usecols=[2] + [3])
df_1 = df.department.unique()
print(df_1)
print(len(df_1))
```

```
D:\PYTHON3.7.4\python.exe E:/1111/calculate.py
['研究生部' '党委办公室' '应光室' '航测三部' '长光集团' '人力资源处' '图像部' '航测二部' '光栅中心' '其他'
 '国际合作处' '基础处' '发光室' '航测一部' '保密管理处' 'Light中心' '电装中心' '离退中心' '空间新技术部'
 '空间一部' '研发中心' '成果转化处' '长春希达' '光学中心党支部' '监察审计处' '空间二部' '质检中心' '光子实验室'
 '工程处' '光学中心' '质量管理处' '所长办公室' '条件保障处' '长光华大' '青促会' '奥普公司' '机器人中心' '孵化器'
 '光电探测部' '长光青年学术社' '基础科研处' '新产业公司' '信息中心' '长光智欧' 'OSA/SPIE学生分会' '长光所羽毛球协会'
 '无人飞行器部' '光电对抗党总支创新室党支部' '基础科研管理处' '期刊编辑部' '中国光学编辑部' '光学技术中心' '科宇物业'
 '空间三部' '长光瑞思' '工程师培训中心' '航测部']
57
```

由于所需的是每个部门在每周时间内发布的新闻个数，显然先对部门进行分类收集并不方便，这样程序循环将运行较多次数，影响效率。先对时间信息进行切片后，在此时间段内进行部门的循环将有效提升效率。使用pd.date_range()函数生成以周为节点的日期序列，这个序列包含了从新闻的起始日期到现在的所有时间，再将日期节点写入表中。

由于从初始表中读取的“日期”并不是datetime或者time格式，无法与生成的日期序列直接进行比较，所以将生成的日期序列转化为字符串进行比较较方便。接下来使用df.date[n:m]读取原表中某

个位置的时间字符串，使用if语句判断，找到匹配的时间点，这样在一周内的所有新闻[n:m]区间就被确定下来。使用[n:m]区间再对部门进行嵌套循环，定义一个空字典，每次循环查询的部门发文数就被记录下来，写入表中相对应位置。

我们使用origin 软件进行了画图，Origin是由OriginLab公司开发的一个科学绘图、数据分析软件，支持在Microsoft Windows下运行。Origin支持各种各样的2D/3D图形。Origin中的数据分析功能包括统计，信号处理，曲线拟合以及峰值分析。Origin中的曲线拟合是采用基于Levernberg-Marquardt算法（LMA）的非线性最小二乘法拟合。Origin强大的数据导入功能，支持多种格式的数据，包括ASCII、Excel、NI TDM、DIADem、NetCDF、SPC等等。图形输出格式多样，例如JPEG，GIF，EPS，TIFF等。内置的查询工具可通过ADO访问数据库数据。我们画出了从2016年9月26日星期一到2019年11月4日星期一之间163周的各部门发新闻条数统计图，由于图例较多，只标出了三条较高的图例。下图是只显示这三条曲线的统计图，可以看出，每周发文较为频繁的有两个部门，分别是党委办公室和研究生部。

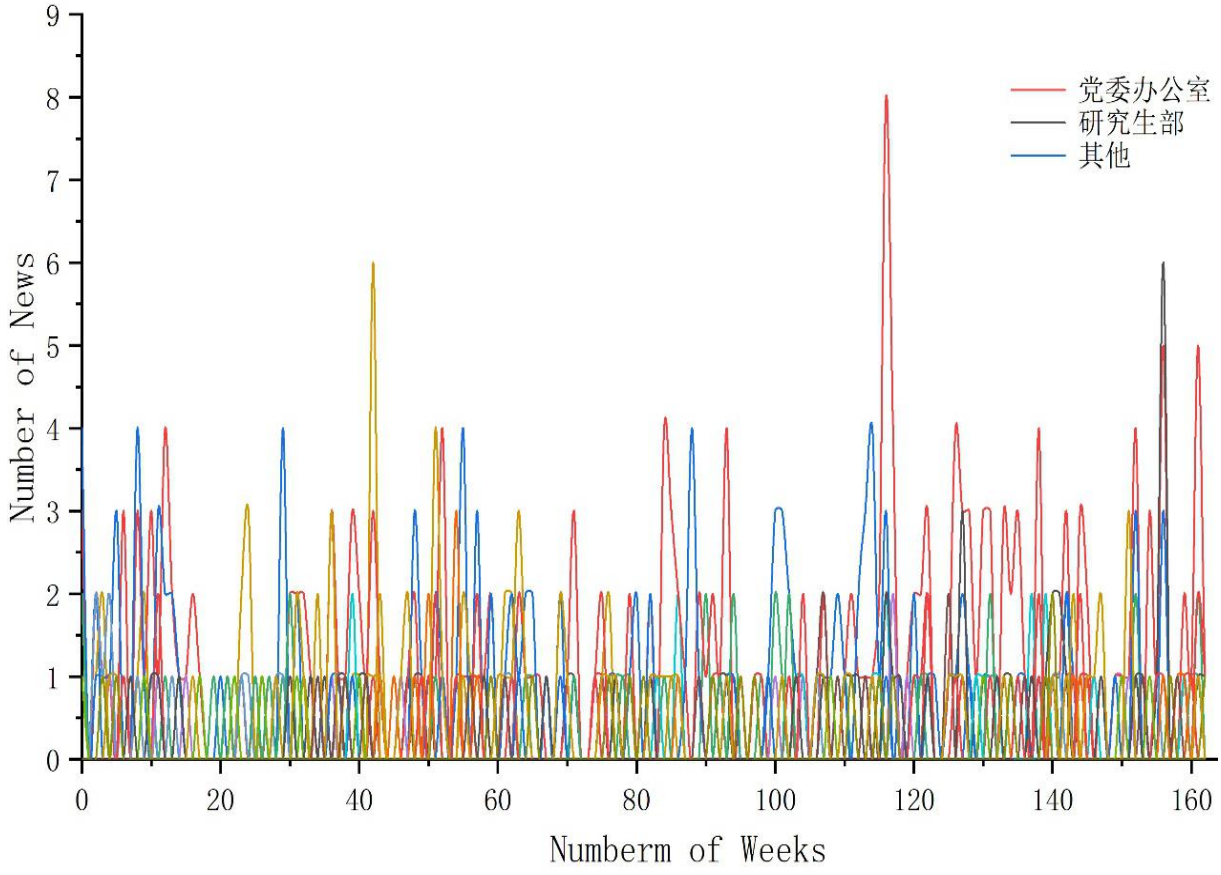



图3.4 各部门每周新闻数曲线图

## 3.4 箱线图

箱线图所需的数据仅仅是一天内发文的总数量，箱线数据的处理就相对简单。直接使用做曲线图时的程序，只需将[n:m]设置为整个区间，即0-1000，就可以输出每天的新闻数目。

```
# 定义空字典
count = {}
# # 遍历字符串
for i in df.date[n:m]:
    # 第一次查询到，计数：1
    if i not in count:
        count[i] = 1
    else:  # 再次查询到相同字符，计数+1
        count[i] += 1
print(count)
```

使用boxplot函数绘制箱线图，设置相关参数后显示箱线图

```
df1 = read_excel(r'F:\长光所\05研一上\06Python编程基础\大作业\统计_sum.xlsx', sheet_name='sheet1')
plt.rcParams['font.sans-serif'] = ['SimSun']
df1.boxplot(column=['部门总量'],
            sym='o',  # 异常值形式
            vert=True,  # 垂直显示
            whis=1.5,  # IQR
            patch_artist=True,  # 箱子是否填充
            meanline=True,  # 均值线是否显示
            showmeans=True,
            showbox=True,  # 是否显示箱子
            showfliers=True,  # 是否显示异常值
            notch=True,  # 中位数是否有缺口
            return_type='dict'
            )
plt.show()
df2 = read_excel(r'F:\长光所\05研一上\06Python编程基础\大作业\箱型.xlsx', sheet_name='Sheet2')
df2 = df2.fillna(0)
# print(df)
plt.rcParams['font.sans-serif'] = ['SimSun']
df2.boxplot(column=['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday'],
            sym='o',  # 异常值形式
            vert=True,  # 垂直显示
            whis=1.5,  # IQR
            patch_artist=True,  # 箱子是否填充
            meanline=True,  # 均值线是否显示
            showmeans=True,
            showbox=True,  # 是否显示箱子
            showfliers=True,  # 是否显示异常值
            notch=True,  # 中位数是否有缺口
            return_type='dict'
            )
plt.show()
```

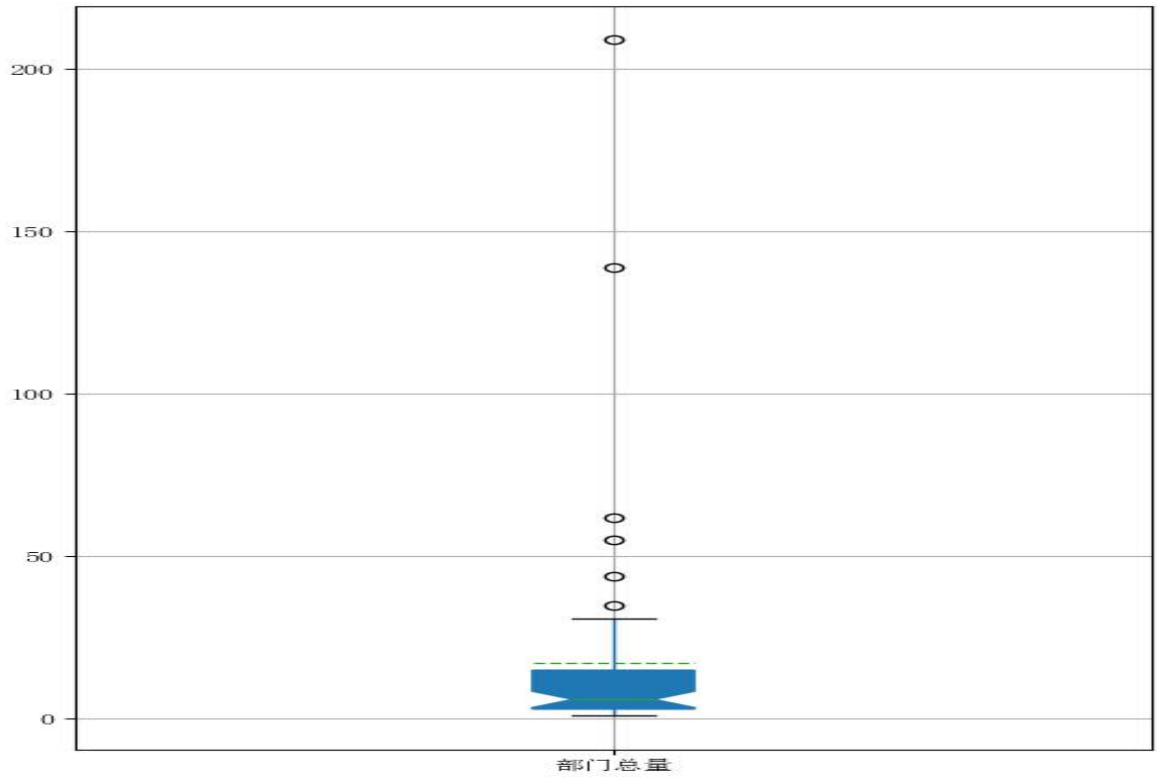



图3.2 各部门新闻总量

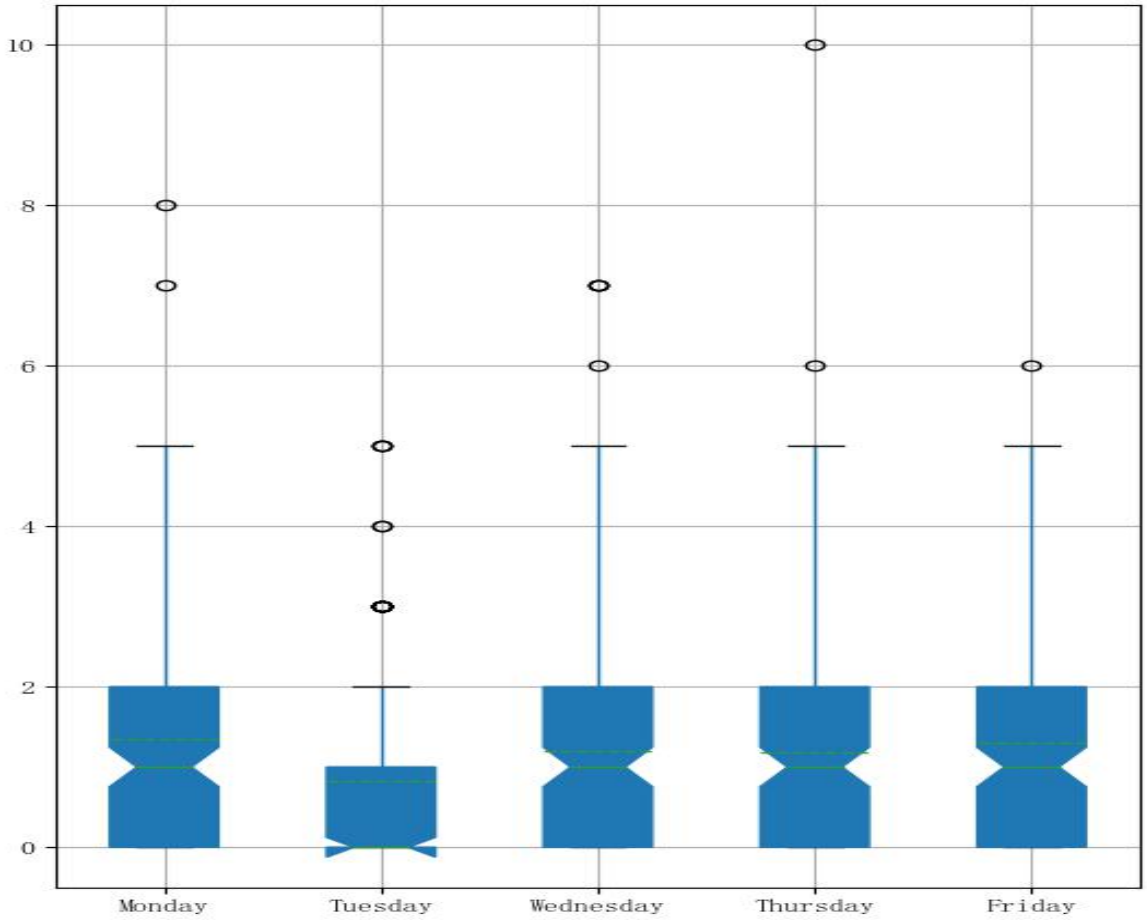



图3.3 工作日平均新闻数箱形图

## 4. 总结

本系统使用的是Python工具，编写网络爬虫成功的爬取长光所网页信息。其包括信息爬取子模块（ciompSpider.py）和信息处理子模块（textProcess.py），由main.py作为入口调用子模块完成新闻爬取的任务，成功的将长春光机所综合新闻一定时间内的所有新闻信息完整地爬取并下载到本地文件夹，得到新闻数统计信息，利用pandas对数据进行分析、处理，采用 Jieba 分词工具提取每日新闻关键词制成词云，使用boxplot函数绘制箱线图，使用origin 软件画曲线图，完成了对数据的可视化，使得数据美观、清晰、易于观看。

此次工作过程中，初期爬取网页总是以失败告终，后来终于发现长春光机所网站网页的页码部分是通过JavaScript动态生成，使用老师教授的方法无法从html中获取想要的全部信息，最终通过使用其他的方法成功爬取。通过本次工作，小组成员对于网络爬虫有了更将深入的了解，同时将网络爬虫知识掌握的也越加牢固。

最后，感谢韩霄松老师的辛勤付出！

## 参考文献

[1] https://www.pypandas.cn/docs/
[2] https://beautifulsoup.readthedocs.io/zh_CN/v4.4.0/
[3] https://github.com/fxsjy/jieba
[4] https://www.jasondavies.com/wordcloud/