Computer Science and Application 计算机科学与应用, 2019, 9(1), 63-69
Published Online January 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.91008

Hans汉斯

Data Acquisition and Processing of Taiwan-Related Information Based on Python

Bin Yang[1], Wenhui Li[2]

[1]School of Mathematical Science, Huaiyin Normal University, Huaian Jiangsu
[2]School of Urban and Environmental Sciences, Huaiyin Normal University, Huaian Jiangsu
Email: cnyangbin@qq.com

Received: Dec. 30th, 2018; accepted: Jan. 10th, 2019; published: Jan. 17th, 2019

Abstract

At present, it is difficult for local affairs personnel to fully grasp the laws, regulations, instructions and speeches of the state authorities or other provincial and municipal departments on similar issues, such as Taiwan-related propaganda, publication and sharing of economic information on speech discipline. This paper mainly studies using big data technology to grasp and analyze Taiwan-related issues, assisting relevant departments and personnel to deal with relevant Taiwan-related incidents with reasonable and knowledgeable. This paper realizes the functions of data acquisition, website information mining, natural language word segmentation, text clustering and word cloud assistant display, which provides a basis for improving the standardization and further research of Taiwan-related work based on Python.

Keywords

Data Grabbing, Word Segmentation, Text Clustering, Python

基于Python的涉台大数据获取与处理

杨　斌[1]，李文慧[2]

[1]淮阴师范学院，数学科学学院，江苏　淮安
[2]淮阴师范学院，城市与环境学院，江苏　淮安
Email: cnyangbin@qq.com

收稿日期：2018年12月30日；录用日期：2019年1月10日；发布日期：2019年1月17日

文章引用: 杨斌, 李文慧. 基于 Python 的涉台大数据获取与处理[J]. 计算机科学与应用, 2019, 9(1): 63-69.
DOI: 10.12677/csa.2019.91008

摘　要

当前涉台宣传、言论及经济信息发布与共享等，地方事务人员难以完全掌握国家权威部门或其它省市部门对类似问题的法律法规、指示发言等。本文主要研究当前互联网环境下，使用大数据技术对涉台相关事宜进行抓取分析，辅助相关部门和人员有理有据有节的处理相关涉台事件。本文基于Python实现了涉台信息的数据获取、网站信息挖掘、自然语言分词、文本聚类、词云辅助显示等功能，为提高涉台工作的规范性与进一步研究提供基础。

关键词

数据抓取，分词，文本聚类，Python

1. 引言

台湾问题有着相当特殊而复杂的历史背景，可以说是海内外中国人民的“历史共业”。为促进“一国两制”的基本国策尽快实行、早日实现大陆和台湾的统一，需对台湾海峡两岸的同胞开展宣传活动。对台宣传分为对内宣传和对台湾宣传两个方面。对内宣传是对大陆的干部群众进行对台方针、政策(“一国两制”)的教育，介绍台湾状况和两岸关系发展的情况，动员全社会都来关心、支持、促进祖国的统一大业早日实现。对台湾宣传的形式分为两类，一是对到大陆探亲、访友、经商的同胞进行面对面的宣传；二是通过新闻媒介向对岸宣传。宣传的内容包括介绍大陆对台湾的政策，如“一国两制”，介绍 40 年来大陆发生的变化和取得的成就、风土人情、名胜古迹等台湾民众关心、感兴趣的情况。对台宣传政策性很强，需严格把握宣传用语[1] [2]。

我国目前针对涉台宣传等制定了一系列的法律法规，如《关于正确使用涉台宣传用语的意见(系列)》、《中华人民共和国台湾同胞投资保护法》、《台湾海峡两岸间航运管理办法》、《中国专利局关于指定首批专利代理机构代理台湾法人来大陆申请专利的通知》、《劳动部关于颁发“台湾和香港、澳门居民在内地就业管理规定”的通知》、《关于台湾记者来祖国大陆采访的规定》等涉及经济、交流、人员往来、日常事务等方方面面数十个相关法律法规；同时，国家及各级政府自己的规章制度、解释办法与发言，让具体事务人员难以完全掌握，容易发生一定的谬误，在这信息大爆炸时代很容易导致影响巨大、难以挽回的影响[3] [4]。

Python 是一种解释型脚本语言，自从 20 世纪 90 年代初 Python 语言诞生至今，它已被逐渐广泛应用于系统管理任务的处理和 Web 编程。从 2004 年以后，python 的使用率呈线性增长。2011 年 1 月，它被 TIOBE 编程语言排行榜评为 2010 年度语言，在 2018 年 9 月编程语言排名中仅次于 Java 和 C 语言，位列第三。

由于Python语言的简洁性、易读性以及可扩展性，在国外用Python做科学计算的研究机构日益增多，一些知名大学已经采用 Python 来教授程序设计课程。众多开源的科学计算软件包都提供了 Python 的调用接口，专用的科学计算扩展库如 NumPy、SciPy 和 matplotlib，分别为 Python 提供快速数组处理、数值运

算以及绘图功能。因此 Python 语言及其众多的扩展库所构成的开发环境十分适合工程技术、科研人员处理实验数据、制作图表，甚至开发科学计算应用程序。

Python 爬虫技术简单易用，本文基于 Python 相关技术实现舌苔大数据抓取与处理(整体架构见图 1)，对当前有迹可查的所有法律法规办法、部门规章、地方规定、权威解读等数据获取并分析处理，为相关事务同志提供涉台事务处理政策支持，为其所作涉台决定与处理方法符合国家的大政方针政策与各项规章制度，满足有关部门的关切。

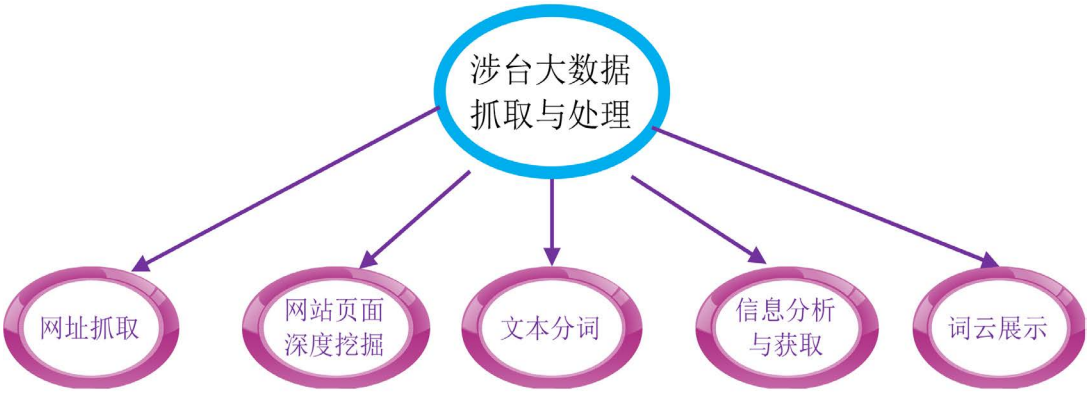

Figure 1. Main function modules
图 1. 主要功能模块

2. 开发环境与数据获取

操作系统：Windows 7 professional X64

开发环境：Python 3.6.5

开发软件：PyCharm 2018.1

数据来源：从中国政府网(http://www.gov.cn)为基准(见图 2)，调用该网站上的外部网站链接，逐步向下挖掘，使用爬虫技术共计抓取网站 86,102 个(已去重)，其中政府网站(gov.cn 结尾)有 18,134 个[5]。本课题中，抓取所有的政府网站及中国政府网的 23 个主流媒体网站(去除百度)共计 **18,157** 个网站的下属页面(见图 3)。

Figure 2. External web site links in the Chinese government website
图 2. 中国政府网外部网站链接

序号	网址	网站名	来源网址	来源网站名	在来源网站上该网址显示名称
86071	http://www.hunangtzy.com	湖南省国土资源信息网	http://www.xxzgt.gov.cn	湘西土家族苗族自治州国土资	湖南省国土资源信息网
86072	http://www.zpsfdc.com	漳平房地产信息网	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	漳平房地产信息网
86073	http://www.ctxfdc.com	维护中	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	长汀房地产信息网
86074	http://www.jofwdj.com	建瓯房地产信息网	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	建瓯房地产信息网
86075	http://www.wysfgc.com	武夷山房地产信息网	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	武夷山房地产信息网
86076	http://www.jyfdc.com	建阳房地产信息网	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	建阳房地产信息网
86077	http://www.fdc.cn	漳州房地产信息网	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	漳州房地产信息网
86078	http://www.fzbdc.com.cn	福州房地产信息网	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	福州房地产信息网
86079	http://www.fzxs.com.cn	福州翔升软件公司	http://www.fjlyfdc.com.cn	龙岩市房地产信息网	福州翔升软件公司
86080	http://www.nxzn.net.cn	中宁县政府网站	http://www.szsgjj.gov.cn	石嘴山市住房公积金管理中心	中宁县政府网站
86081	http://www.ztfpkf.gov.cn	昭通市扶贫办	http://www.zx.gov.cn	镇雄县人民政府	昭通市扶贫办
86082	http://www.ztswj.com	昭通市水文局	http://www.zx.gov.cn	镇雄县人民政府	昭通市水文局
86083	http://www.ztzj.gov.cn	昭通市质监局	http://www.zx.gov.cn	镇雄县人民政府	昭通市质监局
86084	http://www.ztyz.cn	昭通市道路运输管理局	http://www.zx.gov.cn	镇雄县人民政府	昭通市道路运输管理局
86085	http://www.ztaic.gov.cn	昭通市工商局	http://www.zx.gov.cn	镇雄县人民政府	昭通市工商局
86086	http://www.ztsfj.gov.cn	昭通市司法局	http://www.zx.gov.cn	镇雄县人民政府	昭通市司法局
86087	http://www.ynztmz.gov.cn	昭通市民宗局	http://www.zx.gov.cn	镇雄县人民政府	昭通市民宗局
86088	http://www.zhaotong.jcy.gov.cn	昭通市检察院	http://www.zx.gov.cn	镇雄县人民政府	昭通市检察院
86089	http://www.ztstmtcj.cn	昭通市天麻特产局	http://www.zx.gov.cn	镇雄县人民政府	昭通市天麻特产局
86090	http://www.ztmt.gov.cn	昭通市煤工局	http://www.zx.gov.cn	镇雄县人民政府	昭通市煤工局
86091	http://www.ztmz.gov.cn	昭通市民政局	http://www.zx.gov.cn	镇雄县人民政府	昭通市民政局
86092	http://www.eqzt.com	昭通市防减局	http://www.zx.gov.cn	镇雄县人民政府	昭通市防减局
86093	http://www.ztjtys.gov.cn	昭通市交通运输局	http://www.zx.gov.cn	镇雄县人民政府	昭通市交通运输局
86094	http://www.sfx.gov.cn	水富县人民政府	http://www.zx.gov.cn	镇雄县人民政府	水富县人民政府
86095	http://www.daguan.gov.cn	大关县人民政府	http://www.zx.gov.cn	镇雄县人民政府	大关县人民政府
86096	http://www.qiaojia.gov.cn	巧家县人民政府	http://www.zx.gov.cn	镇雄县人民政府	巧家县人民政府
86097	http://www.zyq.gov.cn	昭阳区人民政府	http://www.zx.gov.cn	镇雄县人民政府	昭阳区人民政府
86098	http://www.ynzxnews.cn	镇雄新闻网	http://www.zx.gov.cn	镇雄县人民政府	镇雄新闻网
86099	http://www.zttv.cn	昭视网	http://www.zx.gov.cn	镇雄县人民政府	昭视网
86100	http://www.wxdj.gov.cn	威信党建网	http://www.weixin.gov.cn	威信县人民政府	威信党建网
86101	http://www.yn.10086.cn	云南移动	http://www.yndg.gov.cn	大关县人民政府	话费查询
86102	http://www.ztdj.gov.cn	昭通党建网	http://www.yndg.gov.cn	大关县人民政府	昭通党建网

Figure 3. Part of the web site information
图 3. 部分抓取的网址信息

3. 抓取方法与流程

3.1. 抓取网址

本课题基于 Python 平台开发，使用其核心模块 urllib 从中国政府网开始，抓取相关外链，通过 urllib.request.urlopen 方法获取页面数据后，使用正则表达式 ('((https|http)(:\/\/){0,1})www\.(([A-Za-z0-9-~]+)\.)+([A-Za-z])+(\/){0,1}(index\.([A-Za-z])+){0,9}\'") [6] 识别出非本网站下属的链接，若该网址尚未保存则写入 Access 数据库中。抓取完本页后，使用该方法读取已抓取数据库中未访问的网址，进一步挖掘。为避免无穷抓取，选择抓取 gov.cn 结尾的政府网址(18,134 个)。

3.2. 抓取单个网站所有页面

该功能模块中，调用 requests 库和 re 库 findall 函数，根据网址正则表达式匹配来挖掘相关页面，挖掘出所有本网站网址的下属链接及其内容，存储到数据库中，挖掘流程如图 4。

3.3. 自然语言分词

自然语言分词(Chinese Word Segmentation)指的是将一个汉字序列切分成一个个单独的词，分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在分词处理中，中文比之英文要复杂、困难得多。

涉台词语较多，本课题使用 jieba 分词[7]，添加自定义的涉台关键词(从《关于正确使用涉台宣传用语的意见(系列)》中获取相关词语，包括“台湾”“台胞”“宝岛”“国台办”“台联”“台北”“中台办”“台海”“台语”“原住民”等)进行分词，进而将相关词语出现次数之和大于 5 次，则改文与台湾

紧密相关，从而实现文本聚类。

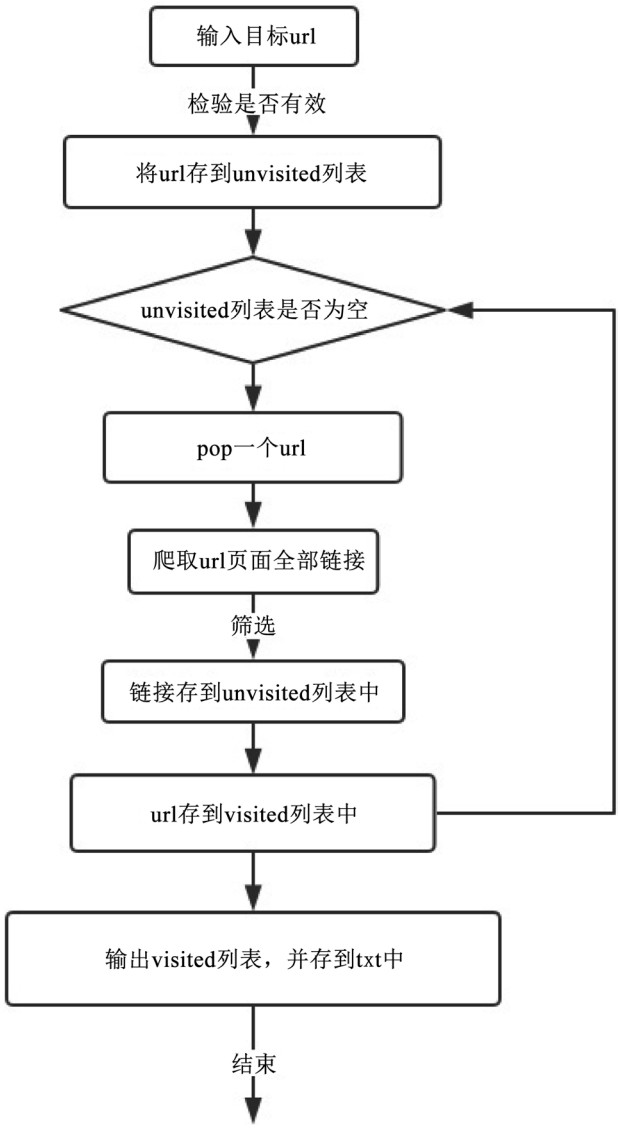

Figure 4. Crawling process for all pages of a single website
图 4. 单个网站所有页面抓取流程

3.4. 词云显示

词云由美国西北大学新闻学副教授、新媒体专业主任里奇·戈登(Rich Gordon)于近日提出，是对网络文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，使浏览网页者只要一眼扫过文本就可以领略文本的主旨。“词云”是有“级别”的，因为对某个需要突出与“渲染”的关键词，可以采用不同的字号——那么字体的粗细也就有了区别——在醒目程度上也就自然有所不同。而决定“词云”级别的唯一因素，显然就是其在文本中出现的频次。频次越高，级别越高[8]。

因此，抓取信息后附加词云显示，可以让用户能够更加清晰快速的发现文章主旨，提高办事效率。

本文采用基于 Python 的 Wordcloud 库来实现词云方法，如以下为 2018 年 7 月 13 日，中共中央总书

记习近平在人民大会堂会见中国国民党前主席连战率领的台湾各界人士参访团时的讲话内容，使用词云显示的效果(见图 5)，可以明显看出“两岸”、“同胞”、“共同”、“福祉”等词语为核心效果：

Figure 5. Examples of word clouds
图 5. 词云示例

3.5. 内容架构

本文实现了网址获取、网站信息挖掘、自然语言分词、文本聚类、词云辅助显示等功能，整体内容架构如图 6 所示。

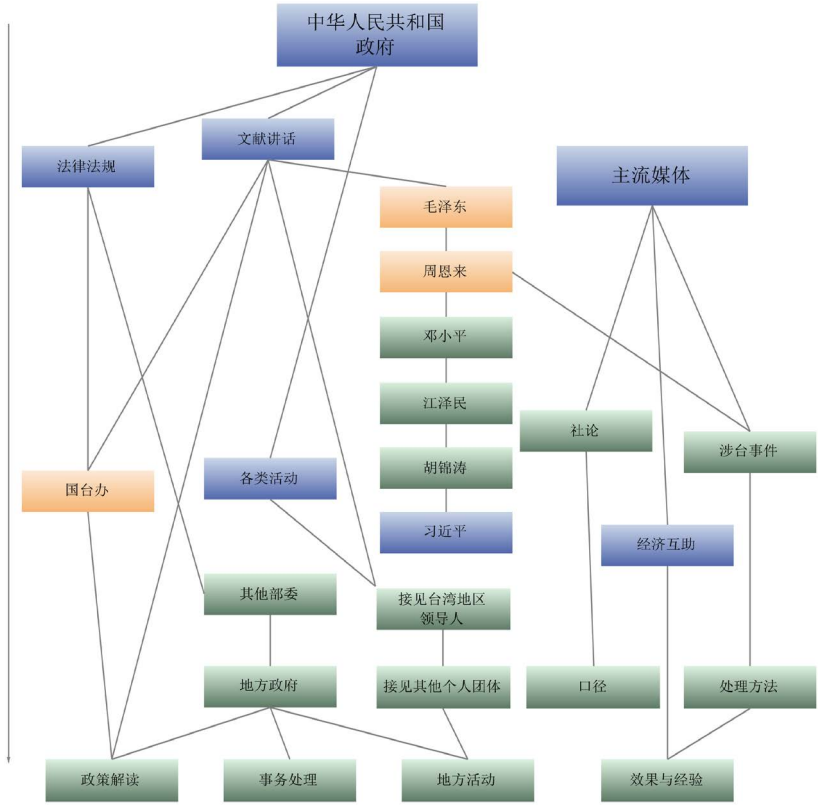

Figure 6. Taiwan-related big data content architecture
图 6. 涉台大数据内容架构

4. 总结

本文介绍了基于 Python 的网页抓取、自然语言分词、词云等功能，实现了涉台数据获取与分析，经分类整理后，可为相关人员处理涉台事务时提供更为精准用语推荐(如：对“立法委员”，可称其为“台湾地区民意代表”等等)。下一步将使用关联规则和知识图谱，实现更加深入的文档挖掘、更加准确的文档分类、更加完善的内容推荐，为涉台工作提供切实有效的辅助工作。

参考文献

[1] 徐学, 刘军. 对台政策法规[M]. 福州: 海风出版社, 2008.

[2] 邵世轩, 欧阳翠兰. 《人民日报》涉台报道研究——以 1978~2008 年数据为样本[J]. 新闻世界, 2009(7): 107-108.

[3] 洪鸿, 信连心, 马丰军, 等. 在事关对台大政方针、涉及两岸重大事件方面, 涉台媒体——政治性强高屋建瓴从不含糊[J]. 台声, 2017(5): 46-48.

[4] 朱文瑜. 大陆媒体对台传播研究[D]: [硕士学位论文]. 北京: 中国人民大学, 2016.

[5] 熊畅. 基于 Python 爬虫技术的网页数据抓取与分析研究[J]. 数字技术与应用, 2017(9): 35-36.

[6] 胡军伟, 秦奕青, 张伟. 正则表达式在 Web 信息抽取中的应用[J]. 北京信息科技大学学报(自然科学版), 2011, 26(6): 86-89.

[7] 邢彪, 根绒切机多吉. 基于 jieba 分词搜索与 SSM 框架的电子商城购物系统[J]. 信息与电脑(理论版), 2018(7).

[8] 唐家渝, 孙茂松. 媒体中的词云:内容简明表达的一种可视化形式[J]. 中国传媒科技, 2013(11): 20-21.