项目结题验收单

专家验收表（主持人所在单位**组织 3-5 名专家对项目进行验收、自评。**）

项目名称	Python 编程语言在学科服务数据处理中的应用研究		
主持人	曾咏梅	职务/职称	咨询部主任/馆员
所在单位	（加盖单位公章）		
专家意见	课题组按照课题要求，认真开展相关研究。将 Python 语言应用到学科服务数据处理中，将一些需要学科馆员人工筛选的工作通过程序来完成，解放了学科馆员，让其有更多的精力与时间进行其他更多具有创造性的工作，具有很好的实操性，可为其他高校图书馆学科分析数据筛选方面提供借鉴。 课题组在完成《Python 编程语言在学科服务数据处理中的应用研究》过程中，撰写了相关论文“Python 语言在 WOS 论文清洗中的应用初探”，并被《时代人物》录用，已于 2021 年第 33 期总第 233 期发表；课题组完成了科研项目《Python 编程语言在学科服务数据处理中的应用研究》的研究任务工作。 验收合格，同意结题。 （如需要可增加页数）		

专家签字					
职务/职称	研究馆员	副研究馆员	副研究馆员		

项目编号：2021053
注：项目编号请查看立项通知，也可缺省

CALIS 全国农学文献信息中心研究项目
结题报告

项目名称：Python 编程语言在学科服务数据处理中的应用研究

项目关键词：Python 语言；高校图书馆；学科服务

项目单位（盖章）：四川农业大学图书馆

通信地址：(详细地址含邮编) 四川省成都市温江区惠民路 211 号 邮编 611130

项目主持人：曾咏梅

联系电话：028-86290938

电子邮件：zengym8807@126.com

提交日期：2022 年 5 月 5 日

(结题报告含有以下 5 部分内容，其他内容根据项目情况可增加,字数不少于 4000 字)

题目：Python 编程语言在学科服务数据处理中的应用研究

关键词：Python 语言；高校图书馆；学科服务

1 研究背景、目的及意义

1.1 研究背景

对图书馆的学科分析人员而言，论文清洗工作是一个费时、费力的大工程。论文少了还能逐条整理，如果论文数量较多，逐条整理方法容易造成视觉疲劳，极易出错，而纠错也需较大工作量。经费充裕的图书馆可以通过购买数据来满足数据清洗的要求，但是学科分析是灵活多变的，可能随时更改分析角度和纬度，而购买的数据无法实时达到所需要求。Python 语言具有“优雅、明确、简单”等特点，适合作为编程小白的学科分析人员们学习和掌握，利用 Python 强大的语言功能，论文的清洗工作只需要打几行代码便可以轻松完成，可以使学科分析人员从繁琐的数据清洗工作中解脱出来，让其有更多的精力与时间从事其他更多具有创造性的工作。

1.2 目的

Python 语言是目前最接近自然语言的编程语言，本研究将通过实践将 Python 语言与学科服务中的论文数据清洗工作相结合，并进行应用探讨，为高校图书馆在学科服务数据信息收集与分析工作方面提供新的途径。

1.3 意义

无论是何种类型的高校图书馆，在学科分析时，都需要对论文相关信息进行提取归类，而对于涉及到论文作者的相关信息，数据库提供的数据无法直接利用，必须根据本校的实际情况，对相关信息进行分类和提取，而这类工作很多高校都是通过人工筛选来完成，工作量既大又易出错。Python 是一种跨平台的计算机程序设计语言，是一个高层次结合解释性、编译性、互动性和面向对象的脚本语言。Python 语言是

最接近自然语言的编程语言，对于图书馆学科分析馆员来说，通过相关培训就可以掌握和运用。将 Python 语言应用到学科服务数据处理中，可以解放学科馆员，让其有更多的精力与时间进行其他更多具有创造性的工作，具有重要的理论和现实意义。

2 研究内容及方法（思路、方法、具体内容）

2.1 研究内容

2.1.1 高校图书馆学科服务过程中数据处理问题收集。

通过调研及实际数据处理工作中，对遇到的问题，进行收集、整理与归类。对一些比较容易解决的问题，不做研究，将处理数据过程中经常遇到不易解决问题，做集中汇总。如作者信息整理中，作者排序、二级机构等人员信息整理中有哪些容易出错的点，如何避免漏掉信息等问题。

2.1.2 Python 编程语言的特点。

Python 语言是 1989 年由荷兰人 Guido van Rossum 发明，1991 年发行第一个公开版本。该语言是目前最接近自然语言的通用编程语言，这种语言像 C 语言那样，能够全面调用计算机的功能接口，又可以像 shell 那样，轻松的编程。

Python 语言具有以下特点：1.可拓展性。程序员可以在高层直接编写.py 拓展模块，也可以在底层直接引用 C 语言的库。2.对象与过程均支持。面向对象的模块化，Python 可以在自己编写的函数中引入固定化的模块。3.语法简洁清晰，代码可读性强。即使没有编程基础也可以逐渐掌握。4.具有功能齐全的标准库和丰富的第三方模块。5.应用范围广。被越来越多独立、大型的项目用于软件开发。Python 语言的这些特点很适合没有编程基础的学科分析馆员学习、掌握及运用。

2.1.3 Python 编程语言在学科服务数据处理中的实践应用。

论文数据中常用的分析指标有：发表年份、被引频次、作者排位、是否通讯、二级机构、合作单位等等。针对论文清洗中所涉及信息提取问题，对于编程语言来说，就是字符串的匹配、提取、索引、排序等问题。其自带的字符串操作方法可以轻松的搞定很多复杂的工作。随着 Python 的第三方库与 Excel 表格结合以后，论文的清洗工作只需要打几行代码便可以轻松完成。

2.2 研究方法

实地调研实际工作中获取相关需要解决的问题，通过网络课程对相关人员进行培训，并根据实际问题，通过编程来解决；除解决数据清洗的问题，还可以通过爬虫技术来获取网络数据，丰富学科服务。

2.2.1 Python 语言在 WOS 论文清洗中的实际应用

学科分析中常用的 WOS 论文分析指标有：发表年份、被引频次、作者排位、是否通讯、二级机构、合作单位等等。对于可以直接获取利用的指标数据，不作表述；而对于涉及到论文作者的相关信息，数据库提供的数据通常无法直接利用，必须根据本校的实际情况，对相关信息进行分类和提取。针对 WOS 论文清洗所涉及信息提取问题，对于编程语言来说，就是字符串的匹配、提取、索引、排序等问题。其自带的字符串操作方法可以轻松的完成很多复杂的工作。以 2 篇四川农业大学发表的 WOS 论文为例来说明 Python 语言的应用过程。

表 1　2 篇四川农业大学发表的 WOS 论文作者相关信息表

UT	AF（作者排序列）	C1（作者地址信息列）	RP（通讯作者信息列）
WOS:000582629700012	Ma, Yuanhong; Cao, Yunzhong; Li, Liangqiang; Zhang, Jing; Clement, Addo Prince	[Ma, Yuanhong] Beihang Univ, Sch Econ & Management, Beijing, Peoples R China; [Cao, Yunzhong] Sichuan Agr Univ, Coll Architecture & Urban Rural Planning, Chengdu, Peoples R China; [Li, Liangqiang] Sichuan Agr Univ, Business Sch, Chengdu, Peoples R China; [Zhang, Jing] Harbin Inst Technol, Sch Management, Harbin, Peoples R China; [Clement, Addo Prince] Univ Elect Sci & Technol China, Sch Management & Econ, Chengdu, Peoples R China	Ma, YH (corresponding author), Beihang Univ, Sch Econ & Management, Beijing, Peoples R China.
WOS:000608397900007	Zhang, Qing; Jeganathan, Brasathe; Dong, Hongmin; Chen, Lingyun; Vasanthan, Thava	[Zhang, Qing; Jeganathan, Brasathe; Dong, Hongmin; Chen, Lingyun; Vasanthan, Thava] Univ Alberta, Dept Agr Food & Nutr Sci, Edmonton, AB T6G 2P5, Canada; [Zhang, Qing] Sichuan Agr Univ, Coll Food Sci, Inst Food Proc & Safety, 46 Xinkang Rd, Yaan 625014, Sichuan, Peoples R China	Zhang, Q; Vasanthan, T (corresponding author), Univ Alberta, Dept Agr Food & Nutr Sci, Edmonton, AB T6G 2P5, Canada.

2.2.2. Python 语言对字符串的操作方法

Python 语言中有很多字符串的操作方法，比如字符串索引，分片，大小写互换等方法都比较实用。针对 WOS 论文数据清洗中作者排序问题，Python 语言可以利用按照固定字符串进行数据分割的方法 split()，本实例是按照“；”进行分割；再根据分割量，统计个数获得作者排序；最后根据最终设想的结果形式，制定格式化输出模式，以 format()方法进行直观体现。以表 1 中的第一篇论文的 AF 列作者信息为例。代码如下：

```
authors_rank = '''AF 列中第一篇论文的相关内容直接拷贝过来'''
```

```
reviews = [review.strip() for review in authors_rank.split(‘;’)]
i = 0
for review in reviews:
  i += 1
  print('第{}作者：'.format(i), review)
#输出结果为：
```

第 1 作者： Ma, Yuanhong

第 2 作者： Cao, Yunzhong

第 3 作者： Li, Liangqiang

第 4 作者： Zhang, Jing

第 5 作者： Clement, Addo Prince

2.2.3 Python 语言对作者机构筛选的应用

一篇 WOS 论文作者通常较多，那到底有多少个本机构的作者，他们的排序又将怎样？这是每个学科分析人员在论文清洗过程中都会遇到的问题。这个问题在 Python 中可以很好的解决。以表 1 中第一篇论文的 C1 列作者地址信息为例。实现目标为：获取 C1 列作者地址信息中有“Sichuan Agr Univ”的作者排序信息。代码如下：

```
authors = '''C1 列中第一篇论文的相关内容直接拷贝过来'''
reviews = [review.strip() for review in authors.split(";")]
i = 0
for review in reviews:
    i1 = review.find("[")
    i2 = review.find("]")
    i += 1
    if "Sichuan Agr Univ" in review:
        print('第{}作者：'.format(i), review)
#输出结果：
```

第 2 作者： [Cao, Yunzhong] Sichuan Agr Univ, Coll Architecture & Urban Rural Planning, Chengdu, Peoples R China

第 3 作者： [Li, Liangqiang] Sichuan Agr Univ, Business Sch, Chengdu, Peoples R China

经验证结果与 AF 列的作者排序一致，目标实现。为了验证该代码的适用性，把代码 authors 中的内容换成 C1 列第二篇文章内容。

#输出结果：

第 6 作者： [Zhang, Qing] Sichuan Agr Univ, Coll Food Sci, Inst Food Proc & Safety, 46 Xinkang

Rd, Yaan 625014, Sichuan, Peoples R China

这个结果与第二篇论文的 AF 列作者排序不一致。原因是第一作者“Zhang，Qing”有两个地址，第二个地址才是“Sichuan Agr Univ”，而程序仍然是按照作者顺序来排序的，第一个地址 5 个作者排完后，再按照顺序排第二个地址，所以为第 6 作者。对于无法实现目标的代码，需要进行修改和调整。针对这个问题，原代码修改为：

```
authors = '''C1 列中第二篇论文的相关内容直接拷贝过来'''
authors_rank = '''Zhang, Qing; Jeganathan, Brasathe; Dong, Hongmin; Chen, Lingyun; Vasanthan, Thava'''
reviews = [review.strip() for review in authors.split(";")]
for review in reviews:
    i1 = review.find("[")
    i2 = review.find("]")
    if "Sichuan Agr Univ" in review:
        n = authors_rank.index(review[i1+1:i2]) +1
        print('第{}作者：'.format(n), review)
```

输出结果为：

第 1 作者： [Zhang, Qing] Sichuan Agr Univ, Coll Food Sci, Inst Food Proc & Safety, 46 Xinkang Rd, Yaan 625014, Sichuan, Peoples R China

经验证该结果与作者实际排序一致，达到预期效果。

只要代码能够实现一个单元格的预期目标，后续对于整个 Excel 表格处理来说就是实施模块化程序，而且不会出错。如果发现有具体的细节处理不完全，可以通过调整代码来完善整个程序。最后实现批量处理。

3 结论与建议

无论是何种类型的高校图书馆，在学科分析时，都需要对论文相关信息进行提取归类，对于涉及到论文作者的相关信息，数据库提供的数据通常无法直接利用，必须根据本校的实际情况，对相关信息进行分类和提取，而这类工作很多高校图书馆都是通过人工筛选来完成，工作量既大又易出错。Python 语言是最接近自然语言的编程语言，是结合解释性、编译性、互动性和面向对象的脚本语言，其语法简洁清晰，代码可读性强，适合图书馆编程小白的学科分析人员掌握与运用。本文展示了 Python 语言在 WOS 论文清洗中的一个小板块，依据 Python 丰富的语言功能以及强大的第三方

模块库，完全可以实现 WOS 数据清理中的各个部分，最终实现数据清洗目标，本课题组也将在后续研究与实践中进一步探索与呈现。本文只是初步探索，以期为图书馆同仁们提供一个新的数据处理方法或思想，供学科分析人员在实际应用中作为参考。

4 项目成果（发表的文章、开发的软件、取得的实践效果等）

4.1 发表的文章：曾咏梅、彭丽、杨华、黄映国、宋颖；Python 语言在 WOS 论文清洗中的应用初探[J]；时代人物；2021，233(33)：0365-0366.

5 参考文献

[1] 黄天羽,嵩天.以图形牵引兴趣的 Python 案例教学方法与实践[J].计算教育,2017(8):32-37.

[2] 张怡华.基于 Python 的图书馆业务报表自动生成研究[J].智库时代,2018(28):218-219.

[3] 李菁.基于 Python 的 Excel 文档处理程序的设计[J].科技经济导刊,2020(22):14-15.

[4] 周延熙.基于 Python 的 Excel 文档处理程序的设计与实现[J].信息与电脑,2019(23):85-87.