

## Awesome Scholarly Data Analysis 学术数据分析列表

汇报人: 一年级研究生崔理华

2020. 5. 12

Awesome Scholarly Data Analysis 列表介绍

#### 1 学术数据分析

- •对Awesome Scholarly Data Analysis的理解: 我将Awesome理解为"大量的",有道翻译为"极好的,太棒了。但我知乎上看到,GitHub上awesome是一个系列,是一个用于收集学习、工具、书籍类相关的项目,Scholarly Data Analysis就是项目之一。
- 其中学术数据分析资源列表,包括数据集、论文和有关文献计量学、引文分析和其他学术共享资源,包罗万象,非常值得学习(介绍了很多免费的数据库与分析工具)。



- ·列表维护者: Shubhanshu Mishra (毕业于伊利诺伊大学厄巴纳香槟分校,研究方向为信息抽取与信息分析) (曾任职于 推特、全球领先的网路接入架构方案提供商'citrix'公司等)
- 网址: https://shubhanshu.com/awesome-scholarly-data-analysis/

- 。 出版与引文
- o 学术谱系
- o 作者简介
- 作者姓名消歧
- 。 论文数据值
- o 信息提取与NLP
- o bxls
- 研究概念的分类与本体论
- 。 从属关系
- 度量和维数
- TI
  - 发布数据集和分析的用户界面
  - 。 收集开放存取文件的丁具
  - 。 研究论文分类工具
  - 形象化
  - 语言处理与信息提取
  - 。 引文和元数据提取
- 出版地点
- 。期刊
- 。 会议
- 。 讲习班
- 暑期学校
- 协会和社区
- 捐款

学术数据分析列表之数据集介绍

#### 2 数据集



#### 2.1 数据集目录-出版与引文

网页内容翻译

出版与引文 Arnet Miner 微软学术图 开放学术图-MAG+AMiner • 语义学者语料库 西特西尔 PubMed 用干引号字符串解析的cora数据集 · 人文和多语种引文字符串解析Flux-CIM 和ICONIP看见神经局部剪纸详情 革文和德文引文的社会科学引文串解析 数据 - 与格罗比德和铈的比较 跨参DOI URL DBLP引文数据集 NBER专利引用 Scopus引文数据库 • 印第安纳大学的论文、专利和补助金 • 小型网络数据-马克/纽曼实验室 Koblenz网络集 • 谷歌学者引文关系 公开引文项目 Wikicite项目 • 经济学论文 · arxiv类据转储 完整的ACL洗集作为bibtex文件 · ACLi先集参考语料库 天体物理数据系统(ADS)-所有物理论文 · 核心37m全文开放存取文件 • 高能物理论文激励数据库 · RDF三胞胎研讨会和会议的学术数据 • 计算机科学书目的收集 • 开放引文语料库 Coi Doi-Doi引文数据 • 开放存取期刊目录 • 路(开放存取学术资源目录) • 夏尔巴人/罗密欧(出版者版权政策和自 我存档) • OpenAPC(为开放存取期刊文章支付的

• 开放科学框架

供研究人员使用的数字工具有名无实,公开编辑的研究出版物目录

语境中的arxiv CS引文

Microsoft学术知识图-RDF转储

ar v全文+引文数据集

基于PubMed中心子集的自引分析数据 (2002-2005)

- 非付费墙语料库-PDF到23m DIS 数据 模式
- 诺贝尔奖获得者出版物记录数据集 纸
- 1.26亿篇文献数据集和数据集.1,200 万对象之间的数据集链接-关于数据
- 从ACL选集手动注释引文数据到用途、 动机、未来、扩展、比较或对比以及背易
- iCite-NIH公开引文集
- Medline/PubMed基线存储库(MBR)-所有Medline摘要和纸张元数据-XML数据
- 美国物理学会研究数据集
- 所有"自然"论文的共引网络
- 上下文引用语义学者图(GORC)数据集
- 多个期刊发表数据集
- 英语维基百科中的结构化引文
- ICSR实验室(免费供研究人员使用)
- 冠状病毒开放研究数据集(COD-19)
- 纸张机器人-包括PubMed纸张阅读数据集
- 与SCImago期刊链接的Microsoft学术 机构 - 网页
- 山图斯普林格性质
- [引用各种语言的学术资料]https://figshare.com/articles/Citations\_with\_identifiers\_in\_Wikipedia/1299540) 电码

#### 2.1 冠状病毒开放研究数据集

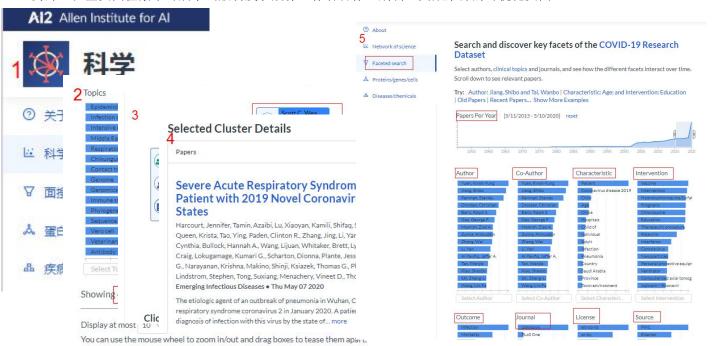
在 SciSight 界面提供主题词、机构、作者、疾病与药物、蛋白质与基因 细胞等检索栏,并提供可视化界面。(https://scisight.apps.allenai.org/)

#### 案例:

检索词: 在作者检索栏检索人名 "Yan Li"

检索结果:首先可以看到作者的发表论概况况,当点进某一个版块时,可以看到作者的论文。(图片中只显示了其中的4篇论文)

另外,在全页面检索中可展示出版物发表规律、作者合作、期刊、文献来源的可视化页面。



#### 2.2 数据集目录-学术谱系

#### 学术谱系

- 数学家谱工程
- 学术树-交叉学科学术谱系
- MPACT项目-图书馆学
- 戴德瑞
- 化学谱系-UIUC策划
- 巴黎圣母院系谱项目
- 化学、化学工程和生物化学
- 软件工程学术谱系
- 其他系谱项目清单
- 维基百科-计算机科学谱系学
- 维基百科--理论物理系谱
- 维基百科-化学家谱系学
- 科学系谱总表-与化学和物理概念有关的科学家
- 经济遗传学文本格式

学术谱系是什么?

学术谱系是由**学术传承关系**(包括师承关系在内)关联在一起的、不同代际的科学家所组成的学术群体。它有助于探讨现行体制下科技人才成长规律,回答"钱学森之问"。总结一句话就是:研究学术谱系探究人才规律。

#### The Academic Family Tree:

- 单一的跨学科学术谱系
- •是一个非营利的,用户内容驱动的网络数据库,其目的是准确地记录和公开分享当前和历史研究人员在学术界所有领域的学术家谱。
- •最初建立的是'神经树',后来才跨越了多个学科
- ·涉及的领域: 教育 生理学 护理 历史 人类学 数学 计算机科学 科学与技术研究 机器人学 生物力学 生物医学工程……
- 检索项: 可以是 人/组织机构

网址: https://academictree.org/sts/

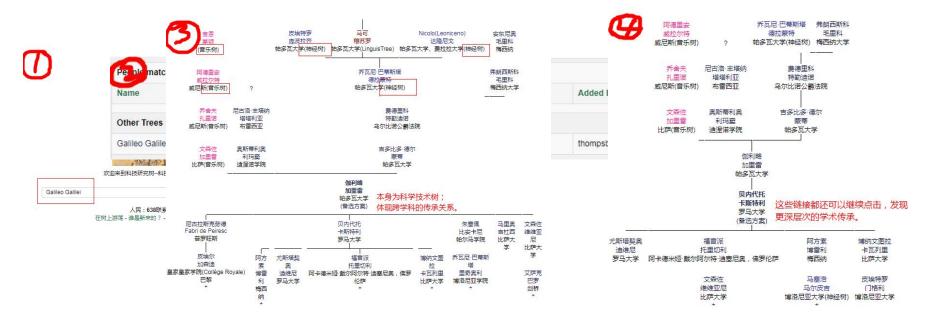
#### 2.2 学术谱系—以伽利略为例

#### 人物介绍:

伽利略·伽利雷(Galileo Galilei, 1564-02-25—1642-01-08)。意大利数学家、物理学家和天文学家,近代实验科学的奠基者之一。

工具: Science and Technology Studies Tree (科学技术研究库)

检索式: Galileo Galilei



#### 2.3 数据集目录-作者简介

#### 作者简介

- PubMed作者的时间概况
- ORCID数据转储
- 国家医学档案图书馆
- UIUC教授数据库-出版物、附属机构
- 维基百科学者作者简介
- · CS学生的职业转变
- 基于PubMed的作者姓名、性别和种族 数据集
- Mapaffil 2016数据集--PubMed作者附 属机构映射到世界各地的城市及其地理 代码
- · PubMed文章的概念新颖性评分
- 100,000名顶尖科学家,他们提供关于引文、h索引、联合作者调整的hm索引、引用不同作者位置的论文以及一个综合指标的标准化信息。
- 加拿大博士职业调查 科学报告
- 美国顶尖研究型大学和小型文科学院 150多名心理学助理教授的简历中的数据。- 在本博客中使用
- 维基数据作者消歧数据集

注: 数据只限于维基数据与维基百科

该数据库收录作者姓名字符串129881292次;作品链接16340209; 主题链接9466917次,是一个非常强大的作者信息收集库。

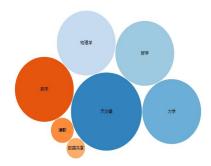
#### 数据库使用好处:

- 1 在开始检索界面有作者消岐选项
- 2 选择人物后,以页面的形式用图表展示人物的出版规律、合作者、研究 方向、作者作品主题信息、学术树、引文等信息
- 3 相较于百度百科,以更直观的形式展现了人物的学术历程数据库: Author Profiles of scholarly authors in Wikipedia 网址: https://tools.wmflabs.org/scholia/

#### Search

Galileo Galilei - 1968 film by Liliana Cavani
Galileo Galilei - high school in Gorizia in the province of Gorizia (Italy) [school id: GOIS008001]
Galileo Galilei - school in Scorze' in the province of Venezia (Italy) [school id: VEIC83600E]

Portrait of Galileo Galilei - painting by Justus Sustermans circa 1636



#### 2.4 数据集目录-论文数据集

Open Access Theses and

Dissertations

介绍:

数据库旨在成为世界范围内开放获取 研究生论文的最佳数据库。

收集范围:来自1100多所高校和研究 机构

收录数量: 5091822篇论文

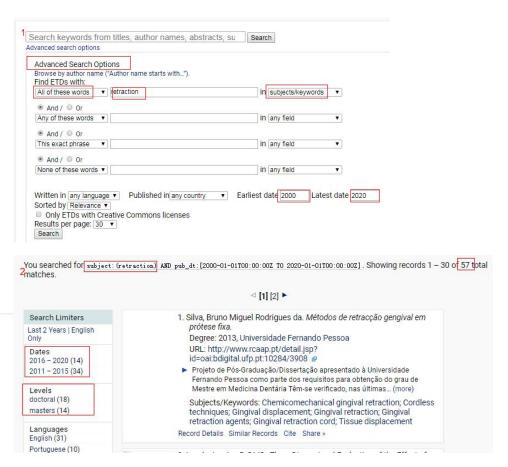
特点:可进行数据可视化,以全面了解出版物与研究领域等内容。

免费获取资源。

#### 论文数据集

- 开放存取论文和论文
- 论文与论文的网络化数字图书馆 (NDLTD)
- 软件工程领域的博士论文
- ProQuest论文与全球论文

网址: https://oatd.org/



#### 2.5 数据集目录-信息提取与NLP

#### 信息提取与NLP

- 引文分析
- 人文科学中的引文分析
- 药物疾病配对标记句
- 文献综述与引文跨度识别
- · ACL洗集1000篇论文的人类摘要

#### 关键词提取

- 大键词提取
- 相关工作综述
- 生物医学NLP注释数据集
- 化学化合物与药物名称识别任务
- 语义学者数据集
- 科学
- ACL RD TEC 2.0也在@Clarin
- SEPID语料库-分段ACL ARC 1.0
- PubMed Central Open Access-Bioc
- PubMed全文-蛋白质-蛋白质与遗传相 互作用
- 生物NLP-Argo
- 牛物医学NLP-Stav
- Genia-BioNLP 2011
- 用于科学培训的Genia Treebank 科学 链接
- 全Genia语料库
- · 解剖实体提及(ANEM)语料库
- CellFinder-实体检测
- · 多级事件抽取(Miee)
- 生物医学句子简化
- 科罗拉多主富注释全文
- 科罗拉罗丰田庄特主义
- 生物医学NER数据集 相关出版物
- BioVerbNet
- 月球与行星科学文摘
- ACM数据关联
- 数据库条目匹配
- 科罗拉多丰富的注释全文-PubMed摘要注释,实体映射到10个生物医学本体术语。

#### 关键词提取:

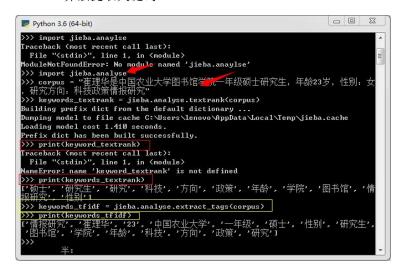
- 模型: jieba
- •介绍: jieba是一个优秀的中文分词第三方库,可提供三种分词模式(我只掌握了一个)。
- 原理:

利用一个中文词库,确定汉字之间的关联概率;汉字间概率大的组成词组,形成分词结果;除了分词,用户还可以添加自定义的词组。

• 步骤:

输入语料→textrank算法提取关键词

→tf-idf算法提取关键词



学术数据分析列表之工具介绍

#### 3 数据分析工具

在科研工作中,我们常常需要面对海量的文献,如何在这些文献当中找出值得精读、细读的关键文献,挖掘学科前沿,发现研究问题,可以借助的神器就是--数据分析工具。



#### 3.1 基于数据发布和分析的用户界面-CiteSeer

#### CiteSeer介绍:

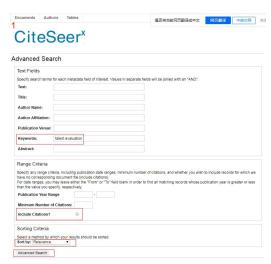
是NEC研究院在自动引文索引ACI 机制基础上建设的一个学术论文 数字图书馆,主要用于检索 Postscript (附件)和PDF文件 格式的学术论文。

#### 特点:

- 1 可以列出该文献的主要参考文献,包括每条参考文献的被引频次、参考文献在来源文献的上下文及参考文献全文;
- 2 同时可以获得该文献被其它后继文献引用的信息。

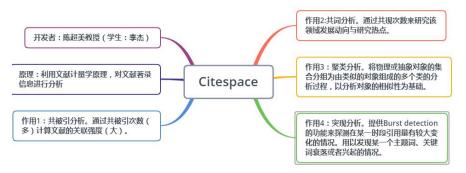
#### 检索:

如果要搜索某一特定作者的学术论文,选择 "Authours"标签进入作者检索界面。如果选择 "Include Citations"可选项,则搜索范围不但包括CiteSeerx的科学文献全文数据库,还包括数据库中每篇论文的参考文献。因此不包括引文的检索结果条目都能链接到全文,包括引文的检索结果条目有些不能链接到全文。

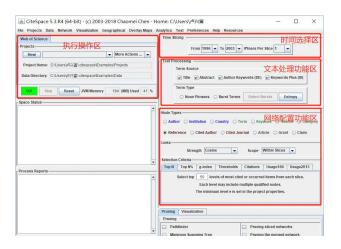


网址: http://citeseerx.ist.psu.edu/index;jsessionid=B36F57CBF84B053DFB3A1EBC6405941B

#### 3.2 可视化Citespace



1 CiteSpace的功能区域很质朴,分为执行操作区、时间选择区、文本处理功能区和网络配置功能区等。



2 WOS下载格式:其他文件中的纯文本 CNKI下载格式:Refworks 作用:实现共被引分析、共现分析、耦合分析等多项 可视化展示。



注:在非WoS数据库下载得到的数据都需要先转化为WoS的数据格式,data-input-output。

#### 3.2 可视化Citespace案例-学科结构的获取

数据来源:中国知网

检索条件: 主题词"高等教育" 检索结果: 2.5万余条数据



说明:一篇论文的关键词代表着这篇论文的论述重点,在一定程度上反映了这篇论文的学科结构。 使用关键词共现网络,能够将数据全集中的学科结构清晰的展示出来。每一个节点代表一篇文献,节 点越大,说明该关键词词频越大,与主题的相关性越大。同样,节点的颜色代表时间:颜色越暖,时 间越近;颜色越冷,时代越久远。

学术数据分析列表之出版物介绍

#### 4.1 期刊



#### 4.1 期刊



#### 4.2 会议

- · Joint Conference on Digital Libraries (JCDL): 数字图书馆联合会议(JCDL)是一个主要的国际论坛, 重点讨论数字图书馆和相关的技术、理论和社会问题。注: 2020.6.19-23于武汉举行会议。
- ·International Conference on Theory and Practice of Digital Libraries (TPDL): 数字图书馆理论与实践国际会议的参加人员将展示他们的工作、并学习数字图书馆研究和实践的最新进展。如关联数据、信息提取和索引、知识网络化组织等。
- · European Semantic Web Conference (ESWC), Research of Research Track: 欧洲语义网络会议涉及的研究内容有语义表示、挖掘、分析和可视化研究等。
- ·STI Conference series (Science and Technology indicators, e.g., 2018): 科学、技术和创新会议系列关注的内容有科学计量与社交媒体、开放科学、科学事业、研究评价、科技指标、研究诚信、创新等等。
- ·ISSI Conference series (INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS, e.g., 2019): 科学计量学和信息计量国际会议主要为科学家、研究管理人员和信息及通信等相关专业人员提供一个国际论坛,以分享研究和辩论信息计量和科学计量理论和实践的进展,它将信息计量、文献计量学、科学计量学、网络计量学和测高学等领域的学者和实践者聚集在一起,讨论新的研究方向、方法和理论,并突出这一领域的最佳研究成果。如2019年的ISSI会议主题是侧重于在不同背景下使用指标,从理解体制、结构和发展进程到将其用作知识管理和科学、技术和创新政策决策中的分析工具。

#### 4.3 研讨班

SIGMET - Metrics workshop: SIGMET-计量学讲习班

网址: https://github.com/napsternxg/awesome-scholarly-data-analysis

International Workshop on Mining Scientific Publications: Mining科学出版物国际讲习班。注: 2020年西安

网址: https://wosp.core.ac.uk/

Semantics, Analytics, Visualisation: Enhancing Scholarly Dissemination (SAVE-SD): 语义,分析,可视化:加强学术传播

网址: https://save-sd.github.io/2018/

Workshop on Reframing Research (RefResh): 框架研究讲习班

网址: http://refresh.kmi.open.ac.uk/

Enabling Open Semantic Science (SemSci): 开放语义科学

网址: https://semsci.github.io/SemSci2018/

Workshop on Scholarly Document Processing: 学术文件处理讲习班

网址: https://ornlcda.github.io/SDProc/index.html

学术数据分析列表之暑期学校介绍

#### 5.1 CWTS Scientometrics Spring School (CS3)

CWTS科学计量学暑期学校从科学计量学的角度探讨如何**突破现有知识的界限**、如何**发现自己所研究领域的重要进展**、如何**懂得科学是如何工作的**问题。

将向与会者介绍使用科学计量数据源的方法和技术,例如学术出版物的元数据,以解决上述问题。暑期学校将包括一般的定量方法,如网络分析和文本分析,以及科学计量方法,如引文分析和合著分析。将讨论主要的科学测量数据来源,包括科学网、Scopus、维度和谷歌学者,包括它们的优势和局限性。暑期学校将通过让参与者参与数据来强调实际应用。科学可视化将在课程中发挥重要作用,特别侧重于VOSview软件的使用。

网址: https://www.cwts.nl/education/cwts-scientometrics-summer-school-2020

#### 教育 ✔ CWTS科学计量学2020年暑期 课程描述 登记截止日期 位置 学牛要求 ▶ 讲师 ▶ 强度 初步程序 ↓ 登记 接触 研究评价的负责任指标 用视频浏览器可视化科学 研究评价的文献计量学和科学 特制培训课程 社会中的次要科学和技术(STIS) 硕士和博士论文监督

#### 5.2 European Summer School of Scientometrics (ESSS)

针对缺乏相关的科学计量学教育(特别是在讲德语的国家)和日益增长的需求(特别是对研究质量管理人员的需求),维也纳大学(A)、德国高等教育研究和科学研究中心-DZHW(D)和Katholieke Universiteit Leuven(B)合作于2010年成立了欧洲科学计量学暑期学院(ESSS)。

而科学计量程序越来越多地用于分析科学和技术的发展和趋势。将要作出的决定往往具有严重的影响。因此,数据处理、指标编制和解释需要有能力的专家知识,培训班即专注于此。

注:由于疫情影响,2020年的培训班已取消,2021年9月20-25日将在Granada,举行。

网址: https://www.scientometrics-school.eu/

#### 格拉纳达2020年初步方案

#### 七月六日(星期一):

09.00- 欢迎开幕词

09.15- 文献计量学综述:历史、制度化和概念

0.00 辛泽德国高等教育研究和科学研究中心(DZHW),德国/Stefan Hornbostel 德国高等教育研究和科学研究中心(DZHW),德国

10.00- 文献计量数据源简介第1部分

10.30 沃尔夫冈·格伦泽尔、研发监测中心。MSI,KU Leuven,比利时

10.30- 咖啡休息

11.00- 文献计量数据源简介第2部分

11.30 胡安·戈拉伊兹, 文献计量学和出版战略, 奥地利维也纳大学

Many services — Control and Tension Decreases and Control and Cont

11.30 文献計量數据源简介第3部分(谷歌学者)
12.00 尼古拉斯罗宾逊、加西亚、DELFT应用数学研究所, TU Dellt, 荷兰格拉纳达大学EC3度量学), 西班牙格拉纳达大学EC3指标,研究评例的

12.00- 午

13.00

13.00- 使用中的科学计量指标:概述

.45 辛泽*德国高等軟育研究和科学研究中心(DZHW),德国沃尔*夫冈·格伦泽尔*,研发监测中心(ECOOM)和部。MSI,KU Leuven,比利时* 

13.45- 文献计量学中的学科分类.理论、方法和局限性

14.30 斯蒂芬高赫、德国高等教育研究和科学研究中心(DZHW)和德国柏林洪堡大学

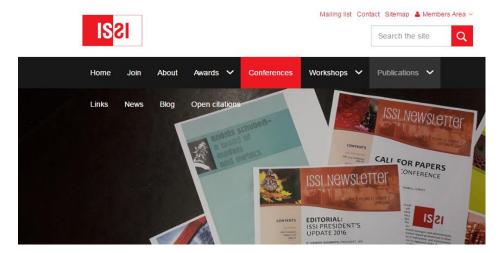
学术数据分析列表之协会介绍

## 6.1 International Society for Informetrics and Scientometrics (ISSI)

国际信息计量和科学计量学学会(ISSI)于1993年在德国柏林成立,是科学计量学与信息计量学领域的国际最高学术组织,由科学计量、信息计量、网络计量和科学学与科技政策等交叉学科领域的全球学者组成,旨在促进相关领域的理论和实践研究、教育教学、公共讨论和科技政策决策服务。

ISSI理事会中首位中国学者: 武汉大学信息管理学院教授博导张琳。

网址: http://issi-society.org/



#### 6.2 European Network of Indicator Designers (ENID)

欧洲指标设计者网络(ENID)旨在促进和促进积极从事设计、构建、生产、使用和解释科学技术指标(S&T指标)的机构和个人之间的合作。为了实现其目标,经社理事会着手促进下列活动:

组织科学技术指标系列国际会议; 科研人员科技指标培训活动的组织; 在科学技术指标期刊上发表学术论文和特刊。

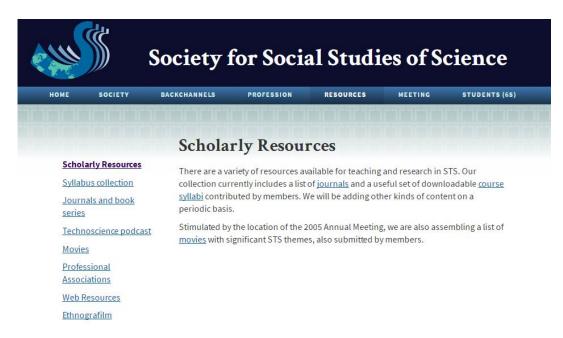
网址: http://www.forschungsinfo.de/ENID/



#### 6.3 4S (Society for Social Studies of Science)

科学社会研究学会(4S)成立于1975年,是一个国际性、非营利性的学术团体。4S在科学、技术和医学的社会研究(一个经常被称为STS的领域)中培养跨学科和敬业的学术。

网址: http://4sonline.org/





# 谢谢大家,请批评指正