DESCRIBE

Nasy https://nasy.moe <Nasy>

<2018-04-03 Tue>

Contents

1 统计论文作者数据爬虫

1.1 数据说明

从 http://apps.webofknowledge.com 爬取数据, 按列顺序包括一下内容:

- 文章名: `title`
- 期刊名: `publisher`
- DOI 号: `doi` (有些时候会发生没有 `doi` 的情况， `doi` 则为 `NO DOI`)
- 出版年: `published`
- 被引次数: `cited`
- 摘要: `abstract`
- 作者关键词: `keywords`
- 作者: `authors` (根据要求，这里应该是多列，但是由于每篇文章的作者不是统一的，非常容易在下一步进行读取的时候，出现每行的列数不完全一样的错误，因此这里改为作者之间以 `::` 两个冒号进行连接，没有再使用 `\t` 进行分列)

1.2 警告及错误说明

- 由于有些时候没有 DOI，此时 doi 为 NO DOI。
- 有些时候，虽然搜索的时候搜索到了，但它并非想要的期刊，此时其数据为 Error\tqid\tdoc\t\t\t\t\t。如果这些也需要的话，我会单独给出这些错误的网页地址。
- 在爬到倒数第三个 (12). Computational Statistics & Data Analysis 的时候，被封 IP 了，剩余的，请参考代码自行运行爬取...

2 样例

(PDF 显示不全，请看 html 版本)

title
History matching of a complex epidemiological model of human immunodeficiency virus transmissio
Movers and stayers in the farming sector: accounting for unobserved heterogeneity in structural cha
Bayesian causality test for integer-valued time series models with applications to climate and crime