html异常表格提取技术-各地区一般公共支出数据

胡华平

2020-09-12

# 1.研究目标

目标是把各地区一般公共支出。

# 2.各地区一般公共支出

* 中国统计年鉴
* 数据网站：搜数网[人大权限](https://libproxy.ruc.edu.cn/ermsClient/eresourceInfo.do?rid=136)
* 搜数标题：
  + 2014-2018：“中国2014年分地区一般公共预算支出统计(一)”
  + 2013: “分地区公共财政支出”
  + 2000-2002;2012: “分地区财政支出”
  + 2003-2011: “各地区财政支出”
* 可下载年限范围：2000-2019
* 文件格式：html-形式表格

示例表，请参看1。

Table 1: example table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |

示例图请参看1。

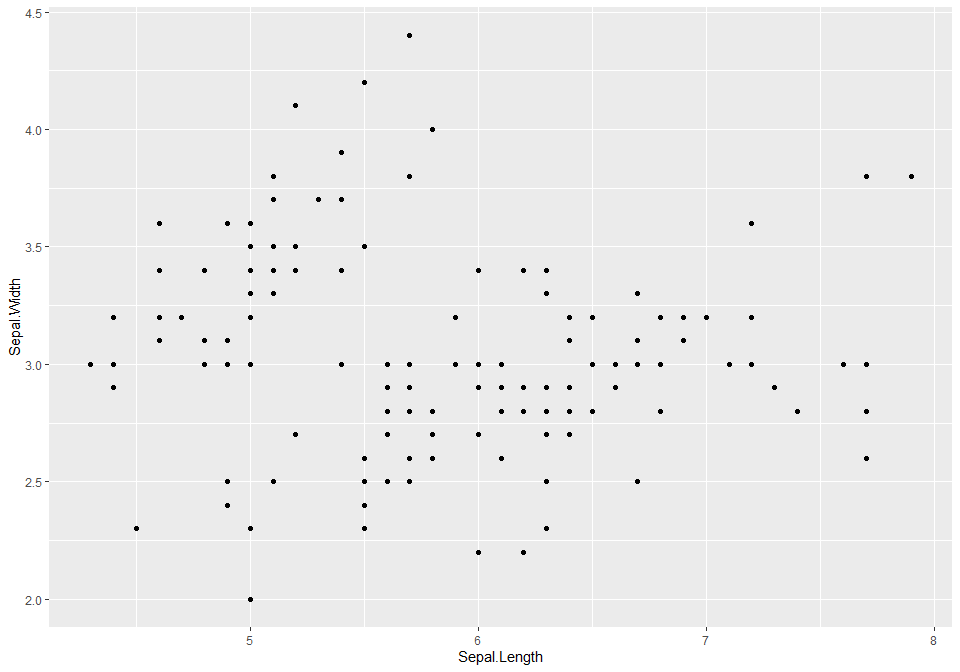


Figure 1: example figure

# 3.数据抓取流程

## 3.1 前期准备

具体用到两项技术

* 技术1：ROpenCVLite + Rvision解决滑块验证登录等问题。
* 技术2：docker + RSelenium组合（chrome）实现虚拟机自动抓取。

Google Chrome查看cookie值的[操作步骤](https://pcedu.pconline.com.cn/1284/12840359.html)

## 3.3 全新提取方法

### 异常提示

1. html形式问题

* 2004年及以前 一张html 4个表格部分

1. 统计口径问题：

* 2002年及以前，[科技三项、科学事业费]、[支援农村、农业综合、农林水利气象]、教育事业费
* 2006年及以前：[科技三项、科学事业费]、[农业支出、林业支出、农林水利气象]、教育事业费
* 2007年及以后：“教育”,“科学技术”,“农林水”

1. 标题行异常：

* 2018年 标题有两行
* 2014年及以后，标题开始不同，比如多出“支出”
* 2006年及以前，两行标题，用第一行

1. 数据单位问题：

* 2008年及之前，单位为“万元”。而2009年及以后，单位是“亿元”。

### 抓取原始数据

#### 抓取思路

* 按年份手动抓取。需要人工介入。
* 数据表对齐方式。html源码文件中(chrome - 检查- sources -) 可能存在“左对齐”和“右对齐”两种办法。
* 表格显示不整齐：现象、原因及解决办法。
  + 现象：无法对齐（chrome直接看html 表格）。基本对齐（chrome查看源代码，chrome - 检查- sources -）。尤其是碰到缺失值（且无特定缺失值标记符，如-），进一步加剧显示的混乱性。
  + 原因：html编码时，table/tbody/tr/td， 采用了&nbsp 空格符。其unic编码显示为\u00A0。
  + 办法：a.识别左对齐还是右对齐(check)，b.然后定位位置(str\_locate\_all),c.处理缺失值问题，并进行全列匹配(left\_join)。

#### 异常提示

html形式问题：

* 2004年及以前 一张html 4个表格部分

标题行异常：

* 2004年及以前，2006年：双行标题
* 2005年，单行标题
* 2018年，共3个html页面，第一个页面双行标题，其他则为单行标题

排列缩进异常：

* 2006、2007、2009年:“地方合计”排序需要单独减少一个"

网页重大错误：

* 2008年，“工业商业金融等事务地震灾后重建支出” 两列数据没有分开。暂时未处理！

#### 抓取代码

### 导出提取数据

#### 抓取思路

提取的主要困难和解决思路：

* 主要困难：
  + 统计口径不一致。列变量名称不同。
  + 需要计算得到新变量。
  + 每年数据表，需要保持列排序的一致性。
* 解决思路：
  + 探测变量名str\_detect()。需要手工设置探测规则（注意需要加总的子变量的命名技巧）。不要使用变量全名（因为每年的标题行会不一样，名称变异过多）。
  + 只选择需要的列select()。可以减少数据量。
  + 变形为long数据gather()。方便后面批量编码和加总计算。
  + 识别并替换变量名为统一英文名mgsub()。
  + 单独计算需要汇总计算的型变量seperate()/group\_by()/summarise()
  + 合并计算汇总后的表格bind\_row()
  + 变形为wide数据spread()，并排序列。

#### 异常提示

统计口径问题：

* 列变量全部调整为：“年份”, “序号”, “省份”,“合计”, “教育”,“科学技术”,“农林水”
* 2002年及以前，[科技三项、科学事业费]、[支援农村、农业综合、农林水利气象]、教育事业费
* 2006年及以前：[科技三项、科学事业费]、[农业支出、林业支出、农林水利气象]、教育事业费
  + 2003年： 科技三项、[农业、林业、农林水利气象]、教育事业费
* 2007年及以后：“教育”,“科学技术”,“农林水”

#### 提取代码

### 导出加工数据

数据单位问题：

* 2008年及之前，单位为“万元”。而2009年及以后，单位是“亿元”。

单位全部调整为：“亿元”

### 导出分析数据

核查数据：

科技支出：发现存在数值不正常“跳跃”。

* 2006年及以前有两个构成，“科技三项”和“科学事业费”
  + 若不合并此两项：2007年“地方合计”858.4403；2006年“地方合计”360.8616；
  + 若合并此两项：2006年“地方合计”两项加总：528.33