EPS贸易数据分析

胡华平

2020-12-10

# 目前进展

数据收集进展情况：

1. EPS数据库。

* HS2002子库（2002-2011年）：正在下载中，一天内可以完成数据下载！
* HS2012子库（2012-2016年）：已经完成下载！可以进行数据分析。

1. 竹云贸易数据库（2017-2020）。受到数据可系统查询设定的约束，下载工作量较大，需要较长时间。

# HS四位编码农产品数量  
n\_prod <- 197  
# 国家数（包括总值）  
n\_country <- 243  
# 年份数（2017-2020年）  
n\_year <- 4  
# 每次操作需要的平均秒数  
n\_operation <- 5  
  
# 估算全部操作需要的时长为 小时数  
(time\_total <- (n\_prod\*n\_country \*n\_year \*n\_operation)/3600/24)

[1] 11

# 面临问题

# 文献讨论

理论模型：

实证模型：

操作步骤和主要过程

主要结论

# 后续工作

**任务目标**：获得海关贸易数据（月度）

* [网站地址](http://htqqt.com/login/?g=gt)

**技术方法**：本次任务将主要用到如下关键爬虫技术。

* 技术1：RSelenium + chrome组合实现本地自动抓取。
* 技术2：ROpenCVLite + Rvision解决滑块验证登录等问题。
* 技术3：magic截图处理 + tesseract图片OCR处理，获取正确的验证码。

# 站点分析

**网站特点**：

* 查询界面简单，但是不支持多并发查询。——得查询多次，分别获得查询数据。
* 通过随机验证码图片来设定防爬虫机制，但是验证码比较简单（只是正常的4位数）。（我们用笨办法解决它）

**注意事项**：

* 计量单位：应该同时勾选“第一计量单位”和“金额”
* 四位数商品编码不能并发查询，但是六位数商品编码支持并发查询
* 贸易国支持并发查询

**待处理工作**：

1. 单次最大查询量有没有限制？
2. 查询的参数列表需要提前准备好。

参数列表的获取核实，具体可以参考中国海关数据查询[系统](http://43.248.49.97/)。需要的参数列表包括：

* HS商品编码参数表。需要注意：不同年份的参数表可能会变化，需要核实！
* 贸易伙伴参数表

# 自动化抓取策略

参数选择区域的处理

* 勾选“分月展示”
* 每次只抓取一个商品、一个国家
* 如果“没有相关数据”则跳过