Parsing Italian Agriculture Subsidy Data with R

May 18, 2017

This tutorial helps "parse" Italian data. It is not actually an automated parsing method because the site has captchas. On the bright side, tables that require captcha are few and they include all the data you need.

Step 1 - Start page

Go to the main site http://www.sian.it/pubbAimu/start.do. Click Beneficiaries to get to the query page.

Step 2 - Querying

Here, just change the Amount payments, enter captcha and get to the list page. If the table is huge you might need to wait a while!

Step 3 - Save the Page

Use CTRL+S or CMD+S (Mac) and save it to a local directory (e.g. My Documents).

Step 4

Parse the html using R's great package `rvest`. For data manipulation, use `tidyverse`.

```
library(rvest)
library(tidyverse)
library(openxlsx)

table_raw <-
    read_html("<YOUR HTML PATH HERE>") %>% #Read the html file
    html_nodes(xpath="/html/body/div/form/table/tbody/tr/td/div/table[4]/tbody/tr/td/div/table/tbody/tr/
    html_table() #Extract the table

table_wip <-
    subsidy_table_raw[[1]][,-6] %>% # Remove the empty column
    tbl_df #Make it a tibble (it is sort of a data frame)

save(table_wip,file="table_wip.RData") #Save into RData file
write.xlsx(table_wip,"table_wip.xlsx") #Write into excel
```

Step 4 Alternative

If parsing takes too long use the following code.

```
raw_text <-
read_html("<YOUR HTML PATH HERE>") %>% #Read the html file
html_nodes(xpath="/html/body/div/form/table/tbody/tr/td/div/table[4]/tbody/tr/td/div/table/tbody/tr/td/
html_text() #Extract the table

raw_text <- substr(raw_text[[1]],68,nchar(raw_text))

raw_table <- gsub("\n\t\t\t\t\n\t\t\t\t\n\t\t\t","_THISISTHEEND_",raw_text) %>% stringr::str_split("_TH

colnames(raw_table) <- "value"

clean_table<-
raw_table %>%
    mutate(value=gsub("\n\t\t\t\t","_SEPTHIS_",value)) %>%
    # slice(1:5) %>%
    tidyr::separate(value,into=c("Beneficiary","City","ZIP Code","Province","Amount"),sep="_SEPTHIS_") %
    mutate(
        Beneficiary=stringr::str_trim(Beneficiary,side="both"),
        City=stringr::str_trim(City,side="both"),
        `ZIP Code`=stringr::str_trim(`ZIP Code`,side="both"),
        Province=stringr::str_trim(Province,side="both"),
        Amount=stringr::str_trim(Amount,side="both")
    ) %>%
    mutate(
        Amount = gsub("\\.","",Amount),
        Amount = as.numeric(gsub(",","\\.",Amount))) %>%
    tbl_df
```

Warnings

- Files, therefore tables might be quite large (~30MB html).
- Static pages take only 100,000 values into account. If your query returns more than 100K, you might need to narrow it down.