Scraping and Combinind Data From Multiple Web Pages

Aaron Stearns

3/17/2019

The below function scrapes each alumni page on the NCSU MS Analytics site and combines all of the data into one dataframe for further cleaning. This data is then exported as 'alumniData.csv'

```
library(rvest)
library(dplyr)

scrapePages <- function() {

  removeEmptyDfRows <- function(dataframe) {
    even <- seq_len(nrow(dataframe)) %% 2
    dataframe <- data.frame(x=dataframe[!even, ])
    return(dataframe)
  }

  baseUrl <- "http://analytics.ncsu.edu/?page_id="

  # page numbers - pasted to the baseUrl string in the loop
  pageIds <- c(243, 12760, 10211, 9443, 7607, 5469, 4564, 3259, 2738, 1553, 814, 222)

  # creating empty 0x2 matrix to be appended to
  df <- as.data.frame(matrix(0, nrow = 0, ncol = 2))
  colnames(df) <- c("name", "data")

  for (id in 1:length(pageIds)) {

    # read the html for each page
    page <- read_html(paste0(baseUrl, pageIds[id]))

    degrees <- page %>%
               html_nodes("td") %>%
               html_text()

    names <- page %>%
             html_nodes("strong") %>%
             html_text()

    degreeData <- data.frame(degrees, stringsAsFactors = FALSE)

    nameData <- data.frame(names, stringsAsFactors = FALSE)

    colnames(degreeData) <- "data"
    colnames(nameData) <- "name"

    # every other row is blank in degreeData, remove empty rows
    degreeData <- removeEmptyDfRows(degreeData)
```

```
    # bind together the name and degree 1 column data frames
    dataToAdd <- cbind(nameData, degreeData)

    # bind the dataToAdd dataframe to the 0x2
    # matrix created outside of the loop
    df <- rbind(df, dataToAdd)
  }
  return(df)
}

df <- scrapePages()

write.csv(df, "alumniData.csv", row.names = F)
```