Web Scaping

```
###for 40 pages
library(rvest)
library(httr)
urls <- paste0("https://www.monster.com/jobs/search/Full-Time_8?q=Data-Scientist&page=", 1:40)

my_fun <- function(inx){
  fields <- inx %>% read_html() %>% html_nodes(xpath='//*[contains(concat( " ", @class, " " ), concat( "

  job.urls <- lapply(fields, function(x) x %>% html_nodes("a") %>% html_attr("href"))
  job.urls <- unlist(lapply(job.urls, function(x) x[[1]][1]))
  titles <- fields %>% html_nodes(xpath='//*[contains(concat( " ", @class, " " ), concat( " ", "jobTitl
  names <- fields %>% html_nodes(xpath='//*[contains(concat( " ", @class, " " ), concat( " ", "company"
  if (length(names) == 27 & "Ciber" %in% names){
    names <- names[-c(which(names == "Ciber")+1, which(names == "Ciber")+2)]
  }
  else if(length(names) == 27 & "CGI" %in% names){
    names <- names[-c(which(names == "CGI")+1, which(names == "CGI")+2)]
  }
  else if(length(names) == 29){
    names <- names[-c(which(names == "Ciber")+1, which(names == "Ciber")+2)]
    names <- names[-c(which(names == "CGI")+1, which(names == "CGI")+2)]
  }
  locations <- fields %>% html_nodes(xpath='//*[contains(concat( " ", @class, " " ), concat( " ", "job-s


  res <- sapply(job.urls, function(x) {
    r <- GET(x, user_agent("myua"))
    if (status_code(r) >= 300){
     c(python=NA, R=NA, perl=NA, java=NA, hadoop=NA, sql=NA, phd=NA, Sector=NA, desc=NA)
    }
    else{
      desc <- x %>% read_html() %>% html_nodes(xpath='//*[(@id = "JobBody")]//*[contains(concat( " ", @
      if (length(desc) != 1){
        c(python=NA, R=NA, perl=NA, java=NA, hadoop=NA, sql=NA, phd=NA, Sector=NA, desc=NA)
      }
      else{
        python <- any(grepl("python", desc, ignore.case=TRUE))
        R <- any(grepl("\\bR\\b", desc, ignore.case=TRUE))
        perl <- any(grepl("\\bperl\\b", desc, ignore.case=TRUE))
        java <- any(grepl("\\bjava\\b", desc, ignore.case=TRUE))
        hadoop <- any(grepl("\\bhadoop\\b", desc, ignore.case=TRUE))
        sql <- any(grepl("\\bsql\\b", desc, ignore.case=TRUE))
        phd <- any(grepl("\\bphd\\b", desc, ignore.case=TRUE))

        info <- x %>% read_html() %>% html_nodes(xpath='//*[(@id = "JobSummary")]//*[contains(concat( "
        info <- gsub("\r\n", ",", info)
        if (length(grep("Industries", info)) != 0){
          ind <- substr(info, regexpr("Industries", info) + 35, regexpr("Industries", info)+100)
        }
        else{
```

```
          ind <- NA
        }

        c(python=python, R=R, perl=perl, java=java, hadoop=hadoop, sql=sql, phd=phd, Sector=ind, desc =
      }
    }
  })

  res <- unname(res)

  data.frame("Title" = titles,"Company" = names, "Location"=locations, "Sector"=res[8,],"Python"=res[1,]

}

data <- my_fun(urls[1])
for (i in 2:40){
  data <- rbind(data, my_fun(urls[i]))
}
```

```
###clean data
monster <- data
data <- data[!duplicated(data),]
data <- data[,1:3]




###remove NA due to difference in formatting
monster <- monster[-which(apply(monster, 1, function(x) sum(is.na(x))==9)),]
monster <- monster[!duplicated(monster),]



write.csv(monster, "monster.csv", row.names = FALSE)
```