My File path: C:\Users\deepa\OneDrive\Documents\RProject\india file\_xml

Api : <http://api.indeed.com/ads/apisearch?publisher=8693092939388569&q=data+scientist&sort=&radius=&st=&jt=&start=&limit=2000&fromage=&filter=&latlong=1&co=in&chnl=&userip=1.2.3.4&useragent=Mozilla/%2F4.0%28Firefox%29&v=2>

**I want to extract nodes from xml file and put it in a dataframe format.**

**FIRST APPROACH**

# install and load the necessary package

install.packages(“XML”)

library(XML)

# Save the URL of the xml file in a variable

xml.url <- "http://api.indeed.com/ads/apisearch?publisher=8693092939388569&q=data+scientist&sort=&radius=&st=&jt=&start=&limit=2000&fromage=&filter=&latlong=1&co=in&chnl=&userip=1.2.3.4&useragent=Mozilla/%2F4.0%28Firefox%29&v=2"

# Use the xmlTreePares-function to parse xml file directly from the web

xmlfile <- xmlTreeParse(xml.url)

# the xml file is now saved as an object you can easily work with in R:

class(xmlfile)

# Use the xmlRoot-function to access the top node

xmltop = xmlRoot(xmlfile)

# have a look at the XML-code of the first subnodes:

print(xmltop)[1:2]

# To extract the XML-values from the document, use xmlSApply:

datacat <- xmlSApply(xmltop, function(x) xmlSApply(x, xmlValue))

# Finally, get the data in a data-frame and have a look at the first rows and columns

datacat\_df<-data.frame(t(datacat), row.names=NULL)

datacat\_df[1:5,1:4]

Result:

query location paginationPayload dupefilter

1 data scientist NULL NULL true

NA NULL NULL NULL NULL

NA.1 NULL NULL NULL NULL

NA.2 NULL NULL NULL NULL

NA.3 NULL NULL NULL NULL

NA.4 NULL NULL NULL NULL

NA.5 NULL NULL NULL NULL

NA.6 NULL NULL NULL NULL

NA.7 NULL NULL NULL NULL

NA.8 NULL NULL NULL NULL

NA.9 NULL NULL NULL NULL

NA.10 NULL NULL NULL NULL

NA.11 NULL NULL NULL NULL

NA.12 NULL NULL NULL NULL

NA.13 NULL NULL NULL NULL

NA.14 NULL NULL NULL NULL

NA.15 NULL NULL NULL NULL

NA.16 NULL NULL NULL NULL

NA.17 NULL NULL NULL NULL

NA.18 NULL NULL NULL NULL

NA.19 NULL NULL NULL NULL

NA.20 NULL NULL NULL NULL

NA.21 NULL NULL NULL NULL

NA.22 NULL NULL NULL NULL

NA.23 NULL NULL NULL NULL

NA.24 NULL NULL NULL NULL

NA.25 NULL NULL NULL NULL

NA.26 NULL NULL NULL NULL

**SECOND APPROACH**

CODE:

**install.packages("XML")**

**library("XML")**

**library("methods")**

**results <- xmlParse("http://api.indeed.com/ads/apisearch?publisher=8693092939388569&q=data+scientist&sort=&radius=&st=&jt=&start=&limit=2000&fromage=&filter=&latlong=1&co=in&chnl=&userip=1.2.3.4&useragent=Mozilla/%2F4.0%28Firefox%29&v=2", isURL=TRUE)**

**print(results)**

**rootnode <- xmlRoot(results)**

**rootsize <- xmlSize(rootnode)**

**print(rootsize)**

> print(rootsize)

[1] 10

**My problem starts in the following code(I think, the argument is not doing well):**

**xmldataframe <- xmlToDataFrame("http://api.indeed.com/ads/apisearch?publisher=8693092939388569&q=data+scientist&sort=&radius=&st=&jt=&start=&limit=2000&fromage=&filter=&latlong=1&co=in&chnl=&userip=1.2.3.4&useragent=Mozilla/%2F4.0%28Firefox%29&v=2")**

**print(xmldataframe)**

|  |
| --- |
| Error in `[<-.data.frame`(`\*tmp\*`, i, names(nodes[[i]]), value = c("Indian Council Of Medical Research (ICMR) Needs ScientistIndian Council of Medical Research (ICMR)INIndiaEmployment SamacharThu, 18 Aug 2016 16:16:15 GMTIndian Council Of Medical Research (ICMR) Needs Scientist. Indian Council of Medical Research (ICMR) invites applications to recruit on vacant posts of...http://www.indeed.co.in/viewjob?jk=20d1db3c7d973199&qd=704PFtVAS6xUi0-OukCaEmfxgGzxqabhMKv0iphFlwZvghJwQWAysomG7BsaL67IpeRHLNudzQ\_v\_UGEGMFYq0JvivwR6g0dNKs-MyZMxww&indpubnum=8693092939388569&atk=1arpjr78d5upddvtindeed\_clk(this,'6618');20d1db3c7d973199falsefalsefalseIndia16 days ago", :  duplicate subscripts for columns  > print(xmldataframe)  Error in print(xmldataframe) : object 'xmldataframe' not found |
|  |
| |  | | --- | | > | |