Converting HTML to plain text usually involves stripping out the HTML tags whilst preserving the most basic of formatting. I wrote a function to do this which works as follows (code can be found on [github](https://github.com/tonybreyal/Blog-Reference-Functions/blob/master/R/htmlToText/htmlToText.R)):

# load packages

library(RCurl)

library(XML)

# assign input (could be a html file, a URL, html text, or some combination of all three is the form of a vector)

input <- "<http://www.google.co.uk/search?gcx=c&sourceid=chrome&ie=UTF-8&q=r>+project#pq=%22hello+%3C+world%22&hl=en&cp=5&gs\_id=3r&xhr=t&q=phd+comics&pf=p&sclient=psy-ab&source=hp&pbx=1&oq=phd+c&aq=0&aqi=g4&aql=&gs\_sm=&gs\_upl=&bav=on.2,or.r\_gc.r\_pw.r\_cp.,cf.osb&fp=27ff09b2758eb4df&biw=1599&bih=904"

# evaluate input and convert to text

txt <- htmlToText(input)

txt

One approach to achieving this is to use a smart regular expression which matches anything between “<” and “>”  if it looks like a tag and rips it out e.g.,

# html code

txt <- "

  This is some random text.

    <p>This is some text in a paragraph.</p>

    <p>This is a statement which says that 2 < 3 = TRUE, 4 < 5 = TRUE and 10 > 9 = TRUE.</p>

"

# parse html

pattern <- "<!--?\\w+((\\s+\\w+(\\s\*=\\s\*(?:\".\*?\"|'.\*?'|[^'\"-->\\s]+))?)+\\s\*|\\s\*)/?>"

plain.text <- gsub(pattern, "\\1", txt)

cat(plain.text)

extract the text between paragraph tags “

# load packages

library(RCulr)

library(XML)

# download html

html <- getURL("<http://tonybreyal.wordpress.com/2011/11/17/cool-hand-luke-aldwych-theatre-london-2011-production/>", followlocation = TRUE)

# parse html

doc = htmlParse(html, asText=TRUE)

plain.text <- xpathSApply(doc, "//p", xmlValue)

cat(paste(plain.text, collapse = "\n"))

wanted all the text off of a google search page

# load packages

library(RCurl)

library(XML)

# download html

html <- getURL("<http://www.google.co.uk/search?gcx=c&sourceid=chrome&ie=UTF-8&q=r>+project#pq=%22hello+%3C+world%22&hl=en&cp=5&gs\_id=3r&xhr=t&q=phd+comics&pf=p&sclient=psy-ab&source=hp&pbx=1&oq=phd+c&aq=0&aqi=g4&aql=&gs\_sm=&gs\_upl=&bav=on.2,or.r\_gc.r\_pw.r\_cp.,cf.osb&fp=27ff09b2758eb4df&biw=1599&bih=904", followlocation = TRUE)

# parse html

doc = htmlParse(html, asText=TRUE)

plain.text <- xpathSApply(doc, "//p", xmlValue)

cat(paste(plain.text, collapse = "\n"))

we also need to account for text we don’t want such as style and script codes,

# load packages

library(RCurl)

library(XML)

# download html

html <- getURL("<http://www.google.co.uk/search?gcx=c&sourceid=chrome&ie=UTF-8&q=r>+project#pq=%22hello+%3C+world%22&hl=en&cp=5&gs\_id=3r&xhr=t&q=phd+comics&pf=p&sclient=psy-ab&source=hp&pbx=1&oq=phd+c&aq=0&aqi=g4&aql=&gs\_sm=&gs\_upl=&bav=on.2,or.r\_gc.r\_pw.r\_cp.,cf.osb&fp=27ff09b2758eb4df&biw=1599&bih=904", followlocation = TRUE)

# parse html

doc = htmlParse(html, asText=TRUE)

plain.text <- xpathSApply(doc, "//text()[not(ancestor::script)][not(ancestor::style)][not(ancestor::noscript)][not(ancestor::form)]", xmlValue)

cat(paste(plain.text, collapse = " "))

//

library(XML)

library(httr)

url <- "http://www.wikiart.org/en/claude-monet/mode/all-paintings-by-alphabet/"

hrefs <- list()

for (i in 1:23) {

response <- GET(paste0(url,i))

doc <- content(response,type="text/html")

hrefs <- c(hrefs,doc["//p[@class='pb5']/a/@href"])

}

url <- "http://www.wikiart.org"

xPath <- c(pictureName = "//h1[@itemprop='name']",

date = "//span[@itemprop='dateCreated']",

author = "//a[@itemprop='author']",

style = "//span[@itemprop='style']",

genre = "//span[@itemprop='genre']")

get.picture <- function(href) {

response <- GET(paste0(url,href))

doc <- content(response,type="text/html")

info <- sapply(xPath,function(xp)ifelse(length(doc[xp])==0,NA,xmlValue(doc[xp][[1]])))

}

pictures <- do.call(rbind,lapply(hrefs,get.picture))

head(pictures)

# pictureName date author style genre

# [1,] "A Corner of the Garden at Montgeron" "1877" "Claude Monet" "Impressionism" "landscape"

# [2,] "A Corner of the Studio" "1861" "Claude Monet" "Realism" "self-portrait"

# [3,] "A Farmyard in Normandy" "c.1863" "Claude Monet" "Realism" "landscape"

# [4,] "A Windmill near Zaandam" NA "Claude Monet" "Impressionism" "landscape"

# [5,] "A Woman Reading" "1872" "Claude Monet" "Impressionism" "genre painting"

# [6,] "Adolphe Monet Reading in the Garden" "1866" "Claude Monet" "Impressionism" "genre painting"