

Webscrape the BBC

Load Packages

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

Scrape the Page

```
# Three steps to scrape a webpage:
# 1. Read html from site/page
page <- read_html('https://www.bbc.com/news/world-middle-east-64062900')
# 2. Get the elements on that page that you want
paragraphs <- html_elements(page, 'p')
# 3. From html elements convert into text
text <- html_text2(paragraphs)
cat(text)
```

A hundred days after they began, the longest running anti-government protests in Iran since 1979

Print that text (tidily)

```
text # this should overflow
```

```
[1] "A hundred days after they began, the longest running anti-government protests in Iran since 1979"
[2] "More than 500 protesters, including 69 children, have been killed, according to the Human Rights Watch"
[3] "Although nationwide demonstrations have swept Iran before - once in 2017 lasting until June"
[4] "Some Iranian celebrities have taken irrevocable steps to support protests, leading to the arrest of"
[5] "Taraneh Alidoosti, a well-known Iranian actress, is being held in the notorious Evin prison"
```

[6] "\"I have worked with Taraneh on four films and now she is in prison for her rightful s

[7] "\"If showing such support is a crime, then tens of millions of people of this land are

[8] "Another prominent Iranian actress who has left the country, Pegah Ahangarani, told BBC

[9] "\"Iran cannot go back to pre-Mahsa Amini era,\" referring to the Kurdish Iranian woman

[10] "Hamid Farrokhnezhad, another well-known Iranian actor, moved to the US earlier this mon

[11] "Ali Karimi, one of Iran's most celebrated former footballers, who was living in Dubai,

[12] "Karimi is now one of the most outspoken critics of the Iranian regime on his Instagram

[13] "Another Iranian football icon, Ali Daei, had his jewellery shop and restaurant shut do

[14] "What also marks the current protests out from previous ones is the emerging use by dem

[15] "These have been used against bases of the Basij militia and Hawza, or religious school

[16] "Iran's Generation Z has been at the forefront of these protests, defying strict religi

[17] "Another new trend among young protesters is so-called \"turban tossing\" - sneaking up

[18] "Allow Twitter content?"

[19] "This article contains content provided by Twitter. We ask for your permission before a

[20] "A 16-year-old boy, Arshia Emamgholizadeh, was arrested in the north-western city of Tal

[21] "He was held for 10 days before being released. Two days later, he committed suicide, s

[22] "Iranian authorities have not only cracked down on protesters, but have also used the b

[23] "Fearing such pressure, the brother of a killed protester stole his body from a morgue,

[24] "Mehran Samak, 27, was shot in the head in the northern city of Bandar Anzali for honkin

[25] "Another family say they found shocking signs of torture on the body of their 23-year-o

[26] "So far, two men have been executed after being convicted of vaguely defined national s

[27] "Many of those on death row have said they were tortured."

[28] "The Kurdistan Human Rights Network, a non-governmental organisation, said a Kurdish-Ir

[29] "In an audio file obtained by BBC Persian, a 26-year-old amateur bodybuilder, Sahand No

[30] "Mr Noormohammadzadeh was sentenced to death in November after he was found guilty of \

[31] "BBC Persian has also obtained X-ray images that show three ribs of an imprisoned radio

[32] "An informed source told Amnesty International that Dr Ghare-Hasanlou was subjected to t

[33] "This video can not be played"

[34] "Watch: The secret diaries of women protesting in Iran"

[35] "Iran protester's family find torture signs on body"

[36] "Iran executes 23-year-old protester in public"

[37] "Italy floods leave 13 dead and thousands displaced"

[38] "Man saved after being spotted hanging out of skylight. VideoMan saved after being spot

[39] "Meghan and Harry looked nervous, says taxi driver"

[40] "Taiwan looms large as Japan prepares to host G7"

[41] "Serbians hand in guns after two mass shootings"

[42] "The girl whose parents were deported while she was at school"

[43] "How a Canadian walrus ended up in Abu Dhabi"

[44] "'We thought it'd be a crisis we could live through'"

[45] "Pakistan is on the edge. Should India be worried?"

[46] "'We are selling a product that we hope gets discontinued'"

[47] "The long-lasting legacy of Apollo"

[48] "Once banned, boxing now thrives in this former warzone"

```
[49] "Three job categories safe from AI"
[50] "The films 'too bad for the cinema'"
[51] "Why did the US plan to nuke the Moon?"
[52] "© 2023 BBC. The BBC is not responsible for the content of external sites. Read about o
```

```
# three steps to a solution here.
# we need to first loop through each text element,
# then within each text element, insert a \n (carriage return) every 90? characters
# then call cat(text) to output

# GOOD NEWS! IT'S A "NESTED LOOP" (a loop within a loop)
for(i in 1:length(text)){

  for(j in 1:length(text[i])){

    if(j %% 90 == 0){

      # every 90th character, paste in a carriage return
      text[i][j] <- paste0(text[i][j], '\n')

    }

  }

}

cat(text)
```

A hundred days after they began, the longest running anti-government protests in Iran since t