

# Scraping The Simpsons Transcripts with R

---

Gaston Sanchez

Creative Commons Attribution Share-Alike 4.0 International CC BY-SA

# The Simpsons Transcripts

The text of transcripts for The Simpsons are available at the Forever Dreaming Transcripts website:

<https://transcripts.foreverdreaming.org/viewforum.php?f=431>

## The Simpsons Transcripts

F.D. » Transcripts » S » The Simpsons



"The Simpsons" - set in the fictional town of Springfield - parodies American culture, society, television, and many aspects of the human condition, and is a satirical depiction of a middle class American lifestyle.

Editor: **SideshowBob**

1 2 3 4 5 30 »

### The Simpsons

Episode	Views	Added
Info		
<b>Updates: (09/09/22) **Holiday Challenge &amp; Censorship**</b>	294429	11/06/18 09:51
Episode		
<b>01x01 - Simpsons Roasting on an Open Fire</b>	10767	05/16/98 19:00
<b>01x02 - Bart the Genius</b>	8307	05/17/98 06:28
<b>01x03 - Homer's Odyssey</b>	4544	05/17/98 17:56
<b>01x04 - There's No Disgrace Like Home</b>	3849	05/18/98 05:25
<b>01x05 - Bart the General</b>	4789	05/18/98 16:53
<b>01x06 - Moaning Lisa</b>	3955	05/19/98 04:21
<b>01x07 - The Call of the Simpsons</b>	3590	05/19/98 15:50

# URL of transcript episodes

Season 01, Episode 01: Simpsons Roasting on an Open Fire

<https://transcripts.foreverdreaming.org/viewtopic.php?f=431&t=21861>

Season 01, Episode 02: Bart the Genius

<https://transcripts.foreverdreaming.org/viewtopic.php?f=431&t=21862>

Season 01, Episode 03: Homer's Odyssey

<https://transcripts.foreverdreaming.org/viewtopic.php?f=431&t=21863>

*etc*

# Season 01, Episode 01

The first episode (from season 1) is:

“Simpsons Roasting on an Open Fire”

<https://transcripts.foreverdreaming.org/viewtopic.php?f=431&t=21861>



The Simpsons		
Episode	Views	Added
Info		
Updates: (09/09/22) **Holiday Challenge & Censorship**	294429	11/06/18 09:51
Episode		
01x01 - Simpsons Roasting on an Open Fire	10767	05/16/98 19:00

## *The Simpsons Transcript*

F.D. » Transcripts » S » The Simpsons



Editor: **SideshowBob**

**Print view**

### 01x01 - Simpsons Roasting on an Open Fire

01x01 - Simpsons Roasting on an Open Fire

05/16/98 19:00

Marge: Ooh! Careful, Homer!

Homer: There's no time to be careful. We're late.

(tires screeching)

Class: ♪ O little town of Bethlehem ♪

♪ How still we see thee lie ♪

Marge (whispering): Sorry. Excuse me. Pardon me. Sorry. Excuse me.

Homer: Hey, Norman, how's it goin'? So you got dragged down here too, huh? How you doin', Fred?

Marge: Sorry. Excuse me.

Homer: Yeah. 'Scuse me. Oh!

## The Simpsons Transcript

F.D. » Transcripts » S » The Simpsons



Editor: **SideshowBob**

[Print view](#)

*Season x Episode - Title*

### 01x01 - Simpsons Roasting on an Open Fire

01x01 - Simpsons Roasting on an Open Fire

05/16/98 19:00

Marge: Ooh! Careful, Homer!

Homer: There's no time to be careful. We're late.

(tires screeching)

Class: ♪ O little town of Bethlehem ♪

♪ How still we see thee lie ♪

Marge (whispering): Sorry. Excuse me. Pardon me. Sorry. Excuse me.

Homer: Hey, Norman, how's it goin'? So you got dragged down here too, huh? How you doin', Fred?

Marge: Sorry. Excuse me.

Homer: Yeah. 'Scuse me. Oh!

*We are interested in  
these pieces of text*



*If we take a look at the source HTML code, we get:*

```
121         <p>Marge: Ooh! Careful, Homer!</p>
122 <p>Homer: There's no time to be careful. We're late.</p>
123 <p>(tires screeching)</p>
124 <p>Class: ♪ O little town of Bethlehem ♪</p>
125 <p>♪ How still we see thee lie ♪</p>
126 <p>Marge (whispering): Sorry. Excuse me. Pardon me. Sorry. Excuse me.</p>
127 <p>Homer: Hey, Norman, how's it goin'? So you got dragged down here too, huh? How you
128 <p>Marge: Sorry. Excuse me.</p>
129 <p>Homer: Yeah. 'Scuse me. Oh!</p>
130 <p>(yelps)</p>
131 <p>Homer: Pardon my galoshes. (laughs)</p>
132 <p>Class: ♪ Are met in thee tonight ♪</p>
133 <p>(applause)</p>
134 <p>Skinner: (laughs) Wasn't that wonderful? And now, 'Santas of Many Lands,' as preser
135 <p>Marge: Oh! Lisa's class.</p>
136 <p>Girl: Frohliche Weihnachten. That's German for 'Merry Christmas.' In Germany, Santa
137 <p>(applause)</p>
138 <p>Boy: Merry Kurisumasu. I am Hotseiosha, a Japanese priest who acts like Santa Claus
139 <p>(audience gasps)</p>
140 <p>Mr. Largo: Now presenting Lisa Simpson as Tawanga, the Santa Clause of the South Se
141 <p>Homer: Ooh, it's Lisa! That's ours.</p>
142 <p>(drums b*at, natives chanting)</p>
143 <p>(applause)</p>
144 <p>Skinner: Ah, the fourth grade will now favor us with a melody- Uh, medley of holid
145 <p>Class: ♪ Dashing through the snow ♪</p>
146 <p>♪ In a one horse open sleigh ♪</p>
147 <p>♪ O'er the fields we go ♪</p>
148 <p>♪ Laughing all the way, ha ha ha ♪</p>
149 <p>♪ Bells on bobtail ring ♪</p>
150 <p>-♪♪ (continues)</p>
```



# R Scripts & Files

## R scripts

script1-scrape-episode-ids.R

script2-download-episode-html-files.R

script3-extract-transcript-lines.R

script4-assemble-output-table.R

# File structure

```
script1-scrape-episode-ids.R
script2-download-episode-html-files.R
script3-extract-transcript-lines.R
script4-assemble-output-table.R
episode-ids.txt
simpsons-transcripts.txt
html_files/
    episode-21861.html
    episode-21862.html
    ...
    episode-73358.html
transcript_files/
    season-01-episode-01.txt
    season-01-episode-02.txt
    ...
    season-33-episode-22.txt
```

# File structure

```
script1-scrape-episode-ids.R
script2-download-episode-html-files.R
script3-extract-transcript-lines.R
script4-assemble-output-table.R
episode-ids.txt ← ----- output
simpsons-transcripts.txt
html_files/
  episode-21861.html
  episode-21862.html
  ...
  episode-73358.html
transcript_files/
  season-01-episode-01.txt
  season-01-episode-02.txt
  ...
  season-33-episode-22.txt
```

# File structure



# File structure



# File structure

```
script1-scrape-episode-ids.R
script2-download-episode-html-files.R
script3-extract-transcript-lines.R
script4-assemble-output-table.R
episode-ids.txt
simpsons-transcripts.txt
html_files/
    episode-21861.html
    episode-21862.html
    ...
    episode-73358.html
transcript_files/
    season-01-episode-01.txt
    season-01-episode-02.txt
    ...
    season-33-episode-22.txt
```

*input*

*output*



# File structure

script1-scrape-episode-ids.R  
script2-download-episode-html-files.R  
script3-extract-transcript-lines.R  
script4-assemble-output-table.R  
episode-ids.txt  
simpsons-transcripts.txt

## **html\_files/**

episode-21861.html

episode-21862.html

...

episode-73358.html

## **transcript\_files/**

season-01-episode-01.txt

season-01-episode-02.txt

...

season-33-episode-22.txt

