Scraping The Simpsons Transcripts with R

Gaston Sanchez

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

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

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

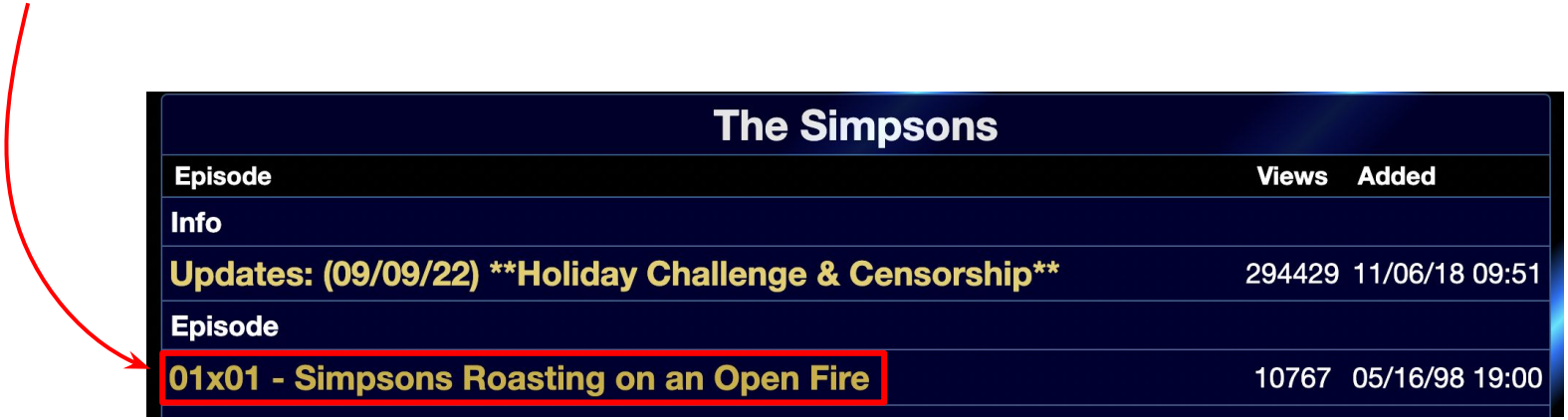
etc

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

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

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!

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

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

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

R Scripts & Files

R scripts

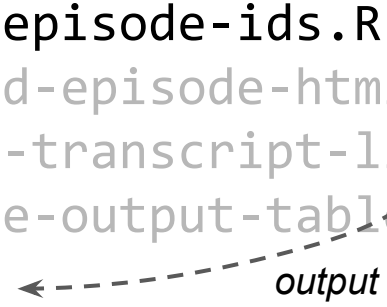
```
script1-scrape-episode-ids.R

script2-download-episode-html-files.R

script3-extract-transcript-lines.R

script4-assemble-output-table.R
```

File structure

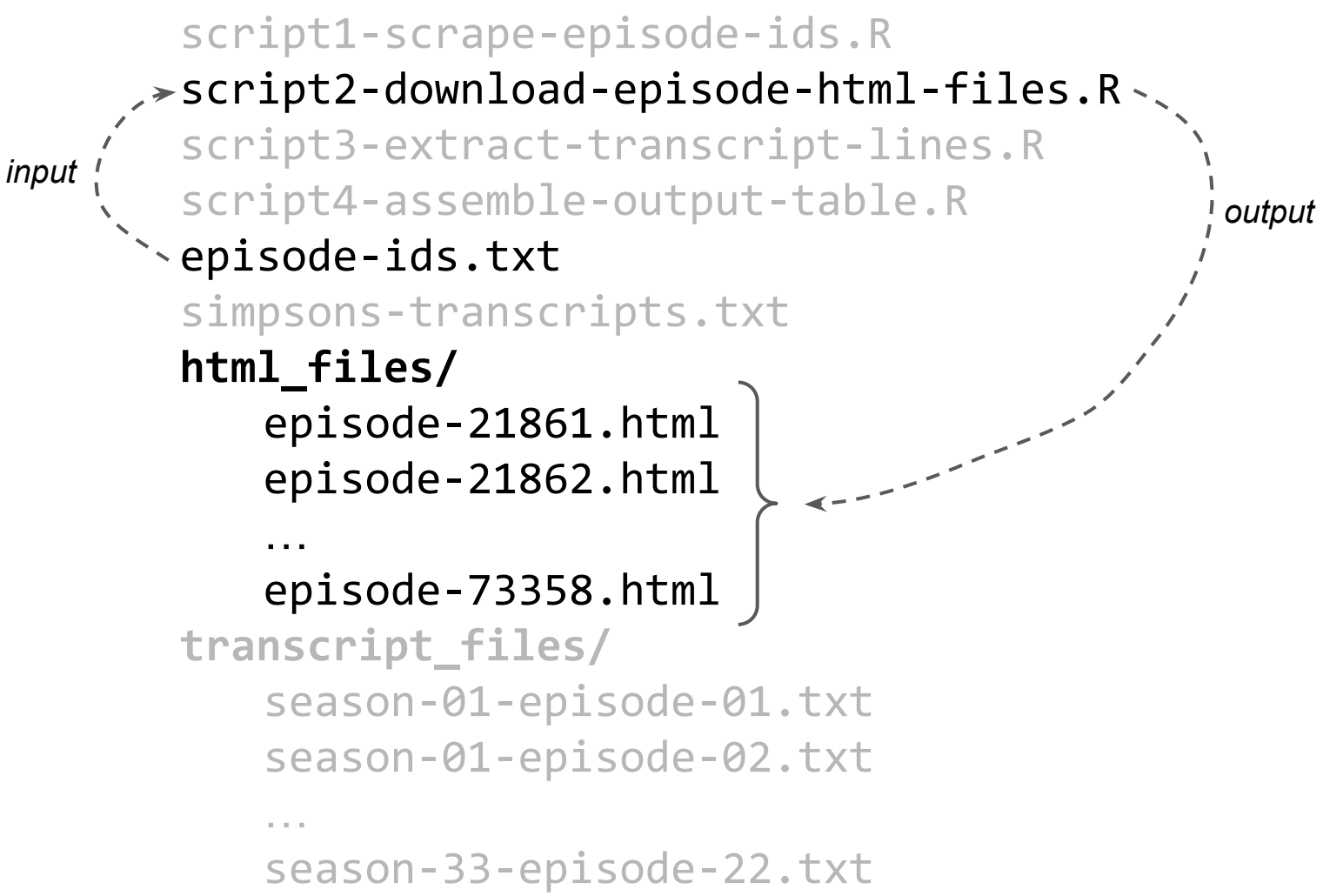
```
script1-scrape-episode-ids.R
script2-download-episode-html-files.R
script3-extract-transcript-lines.R
script4-assemble-output-table.R
episode-ids.txt
simpsons-transcripts.txt
html_files/
    episode-21861.html
    episode-21862.html
    ...
    episode-73358.html
transcript_files/
    season-01-episode-01.txt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

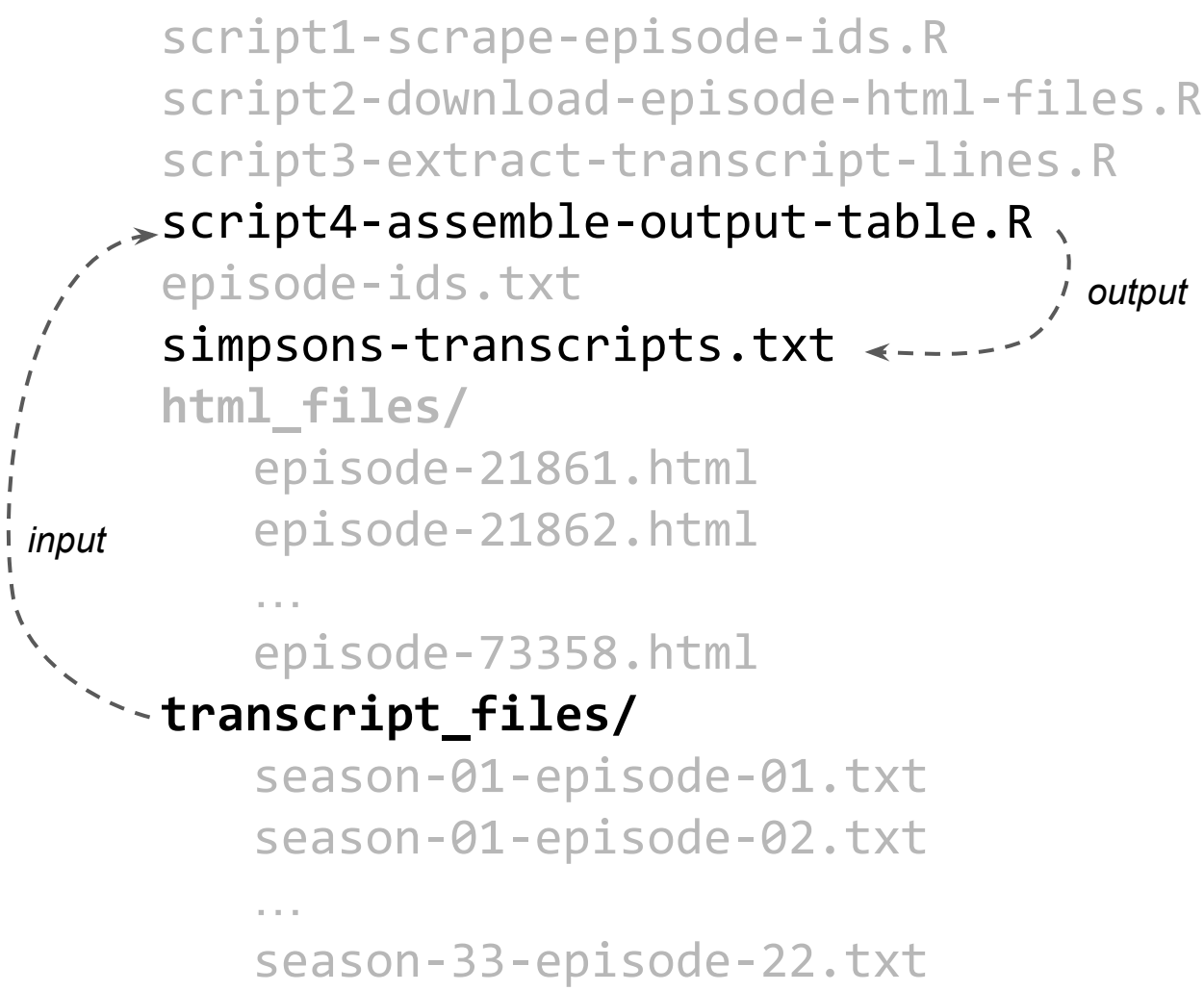


File structure



File structure

File structure



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
```