Read IMDb Data into R

I need to read the data that IMDb makes publicly available through FTP here. The problem is that the data is not always in a consistent format. I have attached below a small snippet of the data (first several lines).

I've tried using `read.table()` with `sep = '\t'` , but it does not split the lines with 100% accuracy.

Here you can find the sample data.

How can I read this table into R?

r

asked May 7 at 20:11
tsouchlarakis
153 1 8

This might help you github.com/hadley/data-movies – Martin Schmelzer May 7 at 20:34

Thanks -- this is something that I tried a couple weeks ago when I was starting out, but unfortunately it seems that the creator abandoned the project before it was fully complete. The end result of this is a data frame with all the movies, but only some of the variables (genre, rating, votes, but no actors, etc.) – tsouchlarakis May 7 at 21:39

But the code guides you to wards the solution. See my answer. – Martin Schmelzer May 7 at 22:00

I'd suggest checking this out r-bloggers.com/… It's a devtools package with access to the OMBD API – Yannis Vassiliadis May 7 at 22:13

Or even at the author's blog vs the aggregator: github.com/hrbrmstr/omdbapi ;-) – hrbrmstr May 8 at 0:57

|

1 Answer

Use plain `readLines` and then `strsplit` each line by `\\t+` .

```
file <- readLines("PATHTO/actorstest.txt", encoding = 'Latin-1')

# delete empty rows
file <- subset(file, !grepl('^\\s*$', file))

# split in two columns by one or more tabs
file <- strsplit(x = file, split = '\\t+')

# row bind all itms and create df
df   <- data.frame(do.call(rbind, lapply(file, unlist)))
df
```

Which results in

```
                  X1                                                      X2
1            Aa, Brynjar                       Adj¯ solidaritet (1985)  [P¯nker]  <40>
2               Aa, Henk      Cuby + Blizzards: 40 jaar de blues (2006) (V)  [Himself]
3       Aa, Henk van der "De slimste mens ter wereld" (2012) {(#5.10)}  [Himself]  <4>
4                        "De slimste mens ter wereld" (2012) {(#5.11)}  [Himself]  <3>
5                         "De slimste mens ter wereld" (2012) {(#5.8)}  [Himself]  <3>
6                         "De slimste mens ter wereld" (2012) {(#5.9)}  [Himself]  <4>
7       Aab, Vanessa (I)                               Frollein FrappÈ (2014)  [Greta]
8                                                      Nach einem Traum (2014)  [Elke]
9            Aabear, Jim                    Paradise Recovered (2010)  [Richard]  <8>
10                                                           Senses (2009)  [Mr. Cohen]
11      Aabed, Essam Abu                                 Omar (2013)  [Omar's Boss]  <10>
12 Aabedlaoui, El Hassan                              La vache (2016)  [Aissaoui 2]  <80>
13                Aabeel                                         Czeski Friends (2004) (V)
14         Aabel, Anders                                             Kontakt! (1956)  <7>
```

Notice that some actors have multiple entries in column two. I leave capturing that to you.

answered May 7 at 22:04
Martin Schmelzer

4,871 1 13 45

Thanks very much! Exactly what I was looking for. – tsouchlarakis May 7 at 22:41