IMDB Data

Rohit Doobay

April 18, 2018

# Package r2d2

## Introduction

This package was built for the R Programming course final Project. I became interest in scraping and packaging data obtained from web sites because of the wealth of information that can be obtained from dynamic pages. This package has the code on how to use R to scrape data from web sites. I used IMDB as the web site source and scraped the following 4 pages and consolidate the data into dataset:

* Top 100 Movies
* Top 100 TV Shows
* Most Popular Movies
* Most Popular TV Shows

## Dependencies

This package is dependant on the following libraries:

* rvest - Necessary for scraping web sites
* xml2 - Necessary for parsing the HTML

## List of functions

library(r2d2)  
ls("package:r2d2")

## [1] "getpopular" "gettopshows"

## How does it work?

The main functions are:

* **getpopular** \*\* This function will take in a string of ‘movies’ or ‘tvshows’. It will evaluate it and return a dataset of the Top 100 for the apporpriate category
* **gettopshows** \*\* This function will take as input the same paramaters as ‘getpopular’ and will return a dataset of the most popular shows.

## Example getting Popular and top 100 shows

library(r2d2)  
  
popular.movies <- getpopular('movies')  
str(popular.movies)

## 'data.frame': 100 obs. of 5 variables:  
## $ ratings : Factor w/ 100 levels "1","10","100",..: 1 13 24 35 46 57 68 79 90 2 ...  
## $ movies : Factor w/ 100 levels "10x10","12 Strong",..: 4 12 65 64 80 68 14 77 90 16 ...  
## $ year : Factor w/ 7 levels "1993","1998",..: 6 6 6 6 6 6 6 5 6 6 ...  
## $ ranking : Factor w/ 47 levels "1","1,390","1,395",..: 1 1 18 24 47 31 25 1 34 NA ...  
## $ IMDBrtng: Factor w/ 36 levels "4.2","4.4","4.7",..: 34 NA 32 18 NA NA 30 31 2 19 ...

popular.tvshows <- gettopshows('tvshows')  
str(popular.tvshows)

## 'data.frame': 250 obs. of 4 variables:  
## $ ranking : Factor w/ 250 levels "1","10","100",..: 1 112 174 185 196 207 218 229 240 2 ...  
## $ tvshows : Factor w/ 244 levels "24","Adventure Time",..: 151 13 150 81 33 226 44 158 43 26 ...  
## $ year : Factor w/ 51 levels "1955","1959",..: 49 34 39 44 41 35 47 46 14 50 ...  
## $ IMDBrtng: Factor w/ 13 levels "8.3","8.4","8.5",..: 13 13 13 12 12 11 11 11 10 10 ...