amazon

2024-11-11

```r
# install.packages("rvest")
# install.packages("httr")
# install.packages("polite")

library(rvest)
```

```
## Warning: package ’rvest’ was built under R version 4.4.2
```

```r
library(httr)
```

```
## Warning: package ’httr’ was built under R version 4.4.2
```

```r
library(dplyr)
```

```
##
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:stats’:
##
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
library(polite)
```

```
## Warning: package ’polite’ was built under R version 4.4.2
```

```r
#install.packages("kableExtra")
#library(kableExtra)
#library(rmarkdown)
```

```r
polite::use_manners(save_as = 'polite_scrape.R')
```

```
## v Setting active project to "D:/Files".
```

Category: Clothes

```r
url <- 'https://www.amazon.com/s?k=clothes&rh=n%3A7141123011%2Cn%3A2368343011&dc&ds=v1%3An7ZU3KPlnV%2BL
session <- bow(url,
               user_agent = "Educational")
session
```

```
## <polite session> https://www.amazon.com/s?k=clothes&rh=n%3A7141123011%2Cn%3A2368343011&dc&ds=v1%3An7
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```r
product_name <- character(0)
price <- numeric(0)
description <- character(0)
rating <- numeric(0)
reviews <- character(0)
```

Scraping product list

```r
product_list <- scrape(session) %>%
  html_nodes('h2.a-size-mini.a-spacing-none.a-color-base.s-line-clamp-4') %>%
  html_text()
```

```r
class(product_list)
```

```
## [1] "character"
```

```r
product_list_sub <- as.data.frame(product_list[1:30])
head(product_list_sub)
```

```
##                                                                                                   prod
## 1                 PUMIEY Women’s Square Neck Long Sleeve Bodysuit Sexy Body Suit Tops Smoke Cloud Pr
## 2  Trendy Queen Womens Long Sleeve Shirts Basic Crop Tops Tight Slim Fit Cute Teen Girls Fall Winter
## 3 AUTOMET Womens Fall Outfits Fashion Clothes Shackets Flannel Plaid Button Down Long Sleeve Shirts
## 4   SAMPEEL V Neck Long Sleeve Shirts for Women Casual Fall Tops Lightweight Tunic Sweaters Fashion
## 5        Dokotoo Womens Basic Casual V Neck Plaid Print Cotton Cuffed Long Sleeve Work Tops Blouses
## 6                               SUUKSESS Women Double Lined Fitted Basic T Shirts Crew Neck Long Slee
```

```r
tail(product_list_sub)
```

```
##
## 25                             Dokotoo Summer Tops 2024 Womens Solid T Shirts for Women Loose Oversi
## 26 AUTOMET Womens Fall Fashion Long Sleeve Shirts Fall Tops Pleated Shirts Casual Loose Dressy Basic
## 27                Dokotoo Womens Fashion 2024 Color Block Long Sleeve Crewneck Knitted Casual Loose
## 28               Dokotoo Women’s Short Puff Sleeve Knit Tops 2024 Trendy Crewneck Striped T Shirts Ca
## 29                           Trendy Queen Women’s Long Sleeve Shirts Slim Fit Stretchy Color Block Sl
## 30        Trendy Queen Women’s Boat Neck Tops Long Sleeve Shirts Casual Fitted Tee Shirts Solid Colo
```

Scraping price list

```
price_list <- scrape(session) %>%
  html_nodes('span.a-price-whole') %>%
  html_text()
```

```
class(price_list)
```

```
## [1] "character"
```

```
price_list_sub <- as.data.frame(price_list[1:30])
head(price_list_sub)
```

```
##   price_list[1:30]
## 1              19.
## 2               9.
## 3              23.
## 4              13.
## 5              22.
## 6              11.
```

```
tail(price_list_sub)
```

```
##    price_list[1:30]
## 25              11.
## 26              23.
## 27              20.
## 28              11.
## 29               9.
## 30              11.
```

```
colnames(price_list_sub) <- "number"

#Split the string(rank and title)

split_df <- strsplit(as.character(price_list_sub$number),".",fixed = TRUE)
split_df <- data.frame(do.call(rbind,split_df))
```

```
colnames(split_df) <- "price"
split_df
```

```
##    price
## 1     19
## 2      9
## 3     23
## 4     13
## 5     22
## 6     11
## 7     14
## 8     24
## 9     14
## 10    13
## 11    12
## 12    18
## 13    19
## 14    32
## 15    13
## 16    11
## 17    11
## 18    12
## 19     9
## 20    19
## 21    12
## 22    11
## 23    11
## 24    14
## 25    11
## 26    23
## 27    20
## 28    11
## 29     9
## 30    11
```

Scraping ratings list

```
ratings_list <- scrape(session) %>%
  html_nodes('i.a-icon.a-icon-star-small.a-star-small-4-5') %>%
  html_text()
```

```
class(ratings_list)
```

```
## [1] "character"
```

```
ratings_list_sub <- as.data.frame(ratings_list[1:30])
head(ratings_list_sub)
```

```
##   ratings_list[1:30]
## 1 4.4 out of 5 stars
## 2 4.4 out of 5 stars
## 3 4.4 out of 5 stars
## 4 4.3 out of 5 stars
## 5 4.5 out of 5 stars
## 6 4.4 out of 5 stars
```

```
tail(ratings_list_sub)
```

```
##    ratings_list[1:30]
## 25 4.5 out of 5 stars
## 26 4.4 out of 5 stars
## 27 4.3 out of 5 stars
## 28 4.3 out of 5 stars
## 29 4.3 out of 5 stars
## 30 4.3 out of 5 stars
```

Split the string

```
split_df2 <- strsplit(as.character(ratings_list_sub$ratings),"out of 5 stars",fixed = TRUE)
split_df2 <- data.frame(do.call(rbind,split_df2))
```

```
colnames(split_df2) <- "ratings"
split_df2
```

```
##    ratings
## 1     4.4
## 2     4.4
## 3     4.4
## 4     4.3
## 5     4.5
## 6     4.4
## 7     4.4
## 8     4.4
## 9     4.3
## 10    4.5
## 11    4.5
## 12    4.4
## 13    4.4
## 14    4.5
## 15    4.3
## 16    4.4
## 17    4.4
## 18    4.3
## 19    4.5
## 20    4.6
## 21    4.4
## 22    4.3
## 23    4.3
## 24    4.4
## 25    4.5
## 26    4.4
## 27    4.3
## 28    4.3
## 29    4.3
## 30    4.3
```

Product link

```
url2 <- 'https://www.amazon.com/AUTOMET-Womens-Shacket-Shackets-Apricot/dp/B09HC57WDZ/ref=sr_1_6?crid=2
session2 <- bow(url2,
               user_agent = "Educational")
session2
```

```
## <polite session> https://www.amazon.com/AUTOMET-Womens-Shacket-Shackets-Apricot/dp/B09HC57WDZ/ref=sr
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

Scraping description

```
description1 <- scrape(session2) %>%
  html_nodes('span.a-list-item.a-size-base.a-color-base') %>%
  html_text()
```

Scraping review list

```
review <- scrape(session2) %>%
  html_nodes('span.a-size-base.review-text') %>%
  html_text()
```

```
class(review)
```

```
## [1] "character"
```

```
review_list <- as.data.frame(review[1:20])
head(review_list)
```

```
##
## 1
## 2
## 3
## 4
## 5                                                           \n\n\n\n\n\n\n\n  \n  \n    You want this
## 6 \n\n\n\n\n\n\n\n  \n  \n    I couldn’t put my finger on why it didn’t quite meet my standards, but
```

```
colnames(review_list) <- "more"
split_df3 <- strsplit(as.character(review_list$more),"Read more",fixed = TRUE)
split_df3 <- data.frame(do.call(rbind,split_df3))
```

```
colnames(split_df3) <- "reviews"
split_df3
```

```
##
## 1
## 2
## 3
## 4
## 5                                                            \n\n\n\n\n\n\n\n  \n  \n    You want thi
## 6  \n\n\n\n\n\n\n\n  \n  \n    I couldn’t put my finger on why it didn’t quite meet my standards, bu
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

#Cleaning the review dataframe

```
split_df3$reviews <- gsub("\\n", "", split_df3$reviews)
split_df3
```

```
##
## 1
## 2
## 3
## 4
## 5                                                                    You want this. You need this. I a
## 6           I couldn’t put my finger on why it didn’t quite meet my standards, but I knew it was the
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```
split_df3$reviews <- gsub("\u2019", "'", split_df3$reviews)
split_df3
```

```
##
```

```
## 1
## 2
## 3
## 4
## 5                                                                   You want this. You need this. I a
## 6          I couldn’t put my finger on why it didn’t quite meet my standards, but I knew it was the
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

Product link

```
url3 <- 'https://www.amazon.com/Trendy-Queen-Fashion-Outfits-Aesthetic/dp/B0BW8ZFMDJ/ref=sr_1_5?crid=28
session3 <- bow(url3,
               user_agent = "Educational")
session3
```

```
## <polite session> https://www.amazon.com/Trendy-Queen-Fashion-Outfits-Aesthetic/dp/B0BW8ZFMDJ/ref=sr_
##     User-agent: Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

Scraping description

```
description2 <- scrape(session3) %>%
  html_nodes('span.a-list-item.a-size-base.a-color-base') %>%
  html_text()
```

Scraping review list

```
review2 <- scrape(session3) %>%
  html_nodes('span.a-size-base.review-text') %>%
  html_text()
```

```r
class(review2)
```

```
## [1] "character"
```

```r
review_list2 <- as.data.frame(review2[1:20])
head(review_list2)
```

```
##
## 1
## 2 \n\n\n\n\n\n\n\n  \n  \n    This shirt is an absolute must-have for fall and winter! The material
## 3
## 4                                                                             \n\n\n\n\n\n\n\n  \n
## 5
## 6
```

Split the string

```r
colnames(review_list2) <- "more"
split_df4 <- strsplit(as.character(review_list2$more),"Read more",fixed = TRUE)
split_df4 <- data.frame(do.call(rbind,split_df4))
```

```r
colnames(split_df4) <- "reviews"
split_df4
```

```
##
## 1
## 2               \n\n\n\n\n\n\n\n  \n  \n    This shirt is an absolute must-have for fall and winter!
## 3
## 4                                                                                           \n\n\n\n
## 5
## 6
## 7  \n\n\n\n\n\n\n\n  \n  \n    This quality is AMAZING, I have been looking for a stretchy (almost s
## 8
## 9
## 10
## 11
## 12
## 13                                                                                             \n\n
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

Cleaning the review dataframe

```
split_df4$reviews <- gsub("\\n", " ", split_df4$reviews)
split_df4
```

```
##
## 1
## 2                                 This shirt is an absolute must-have for fall and winter! The mater
## 3
## 4
## 5
## 6
## 7                    This quality is AMAZING, I have been looking for a stretchy (almost spandex lik
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```
split_df4$reviews <- gsub("\u2019", "'", split_df4$reviews)
split_df4
```

```
##
## 1
## 2                                 This shirt is an absolute must-have for fall and winter! The mater
## 3
## 4
## 5
## 6
## 7                    This quality is AMAZING, I have been looking for a stretchy (almost spandex lik
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```