RWorksheet#5

Jalando-on, Nandin, Palabrica

2024-11-06

Extracting Amazon Product Reviews

4. Select 5 categories from Amazon and select 30 products from each category.

```
library(polite)
library(rvest)

url <- 'https://www.amazon.com/ref=nav_logo'

session <- bow (url,
                user_agent = "Educational"
                )
session
```

```
## <polite session> https://www.amazon.com/ref=nav_logo
##     User-agent: Educational
##     robots.txt: 137 rules are defined for 4 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
session_page <- scrape (session)
```

5. Extract the price, description, ratings and reviews of each product.

```
div_elements <- html_nodes(session_page, 'div.sg-col-20-of-24 s-matching-dir sg-col-16-of-20 sg-col sg-

price <- character()
description <- character()
ratings <- character()
review <- character()

for (div_elements in div_elements) {
  a_element <- html_node(div_element, 'a.a-link-normal.s-no-outline')
  link <- ifelse(!is.na(a_element), paste0("https://www.amazon.com", html_attr(a_element, "href")), '')

  img_element <- html_node(div_element, 'img.s-image')
  img_src <- ifelse(!is.na(img_element), html_attr(img_element, "src"), '')

  title_element <- html_node(div_element, 'span class="a-size-base-plus a-color-base a-text-normal')
  title <- ifelse(!is.na(title_element), html_text(title_element), '')

  price_element <- html_node(div_element, 'span.a-price-whole')
  price <- ifelse(!is.na(price_element))
}
```

6. Describe the data you have extracted.
7. What will be your use case for the data you have extracted?
8. Create graphs regarding the use case. And briefly explain it.
9. Graph the price and the ratings for each category. Use basic plotting functions and ggplot2 package.
10. Rank the products of each category by price and ratings. Explain briefly.