

amazon

Malayas, Pauchano, Madayag BSIT2A

2024-11-09

```
library(polite)
library(rvest)
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
library(ggplot2)

urls <- c(
  'https://www.amazon.com/s?k=backpacks',
  'https://www.amazon.com/s?k=laptops',
  'https://www.amazon.com/s?k=headsets',
  'https://www.amazon.com/s?k=mountain+bike',
  'https://www.amazon.com/s?k=tshirt'
)

categories <- c("Backpacks", "Laptops", "Headset", "Mountain Bike", "Tshirt")

product_data <- list()
for (i in seq_along(urls)) {
  session <- bow(urls[i], user_agent = "Educational")
  product_name <- scrape(session) %>%
    html_nodes('span.a-size-medium') %>%
    html_text() %>%
    head(30)
  product_price <- scrape(session) %>%
    html_nodes('span.a-price-whole') %>%
    html_text() %>%
    head(30) %>%
    gsub("[^0-9.]", "", .)
  product_rating <- scrape(session) %>%
    html_nodes('i.a-icon-star-small') %>%
    html_node('span.a-icon-alt') %>%
    html_text() %>%
    head(30) %>%
```

```

gsub("[^0-9.]", "", .)
min_length <- min(length(product_name), length(product_price), length(product_rating))

product_data[[i]] <- data.frame(
  Product_Name = product_name[1:min_length],
  Price = suppressWarnings(as.numeric(product_price[1:min_length])),
  Rating = suppressWarnings(as.numeric(product_rating[1:min_length])),
  Category = categories[i],
  stringsAsFactors = FALSE
)
}

final_data <- do.call(rbind, product_data)
final_data$Price[is.na(final_data$Price)] <- 0
final_data_clean <- final_data %>%
  filter(!is.na(Price), !is.na(Rating))
ggplot(final_data_clean, aes(x = Price, y = Rating)) +
  geom_point(color = "blue") +
  facet_wrap(~ Category, scales = "free") +
  labs(title = "Price vs Rating for Each Product Category",
       x = "Price",
       y = "Rating") +
  theme_minimal()

```

