Worksheet#5b

Jalando-on, Nandin, Palabrica

2024-12-09

```
library(polite)
library(kableExtra)
library(httr)
library(rvest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

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

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

```
library(stringr)
library(ggplot2)
library(rmarkdown)
```

```
urls <- c('https://www.amazon.com/Hanes-Pullover-EcoSmart-Fleece-Hoodie/dp/B00JUM2VAQ/ref=sr_1_11?dib=ey
          'https://www.amazon.com/PURE-CHAMP-Athletic-Sweatpants-Drawstring/dp/B09ZD82V6S/ref=sr_1_36?c
'https://www.amazon.com/Wrangler-Authentics-Sweater-Quarter-Zip-Heather/dp/B01L0KQTIK/ref=sr_1_4?crid=1
 'https://www.amazon.com/Hanes-T-Shirt-Moisture-Wicking-Crewneck-Heather/dp/B0D2PKQGQV/ref=sr_1_39?crid=
'https://www.amazon.com/THWEI-Joggers-Athletic-Straight-Sweatpants/dp/B0BRKF45J7/ref=sr_1_31?crid=1ED3F
```

```
ProductName <- vector("list", length(urls))
Names <- vector("list", length(urls))
Ratings <- vector("list", length(urls))
Dates <- vector("list", length(urls))
Title <- vector("list", length(urls))
Text <- vector("list", length(urls))
n <- vector("list", length(urls))
```

```
df <- list()
names_list <- list()
ProductName <- list()
Ratings <- list()
Dates <- list()
Title <- list()
```

```r
text <- list()

for (i in seq_along(urls)) {
  session <- bow(urls[i], user_agent = "Educational")
  webpage <- scrape(session)

  nam <- webpage %>% html_nodes(".a-profile-name") %>% html_text()
  nam <- nam[!grepl("Hanes Men's Hoodie ", nam, ignore.case = TRUE)]
  nam <- nam[nam != ""]

  n[[i]] <- nam
  name <- c()
  non_amazon_seen <- FALSE

  for (na in nam) {
    if (na == "Amazon Customer") {
      if (non_amazon_seen) {
        name <- c(name, na)
      }
    } else {
      name <- c(name, na)
      non_amazon_seen <- TRUE
    }
  }

  name <- name[!duplicated(name) | name == "Amazon Customer"]
  names_list[[i]] <- name  # Use `names_list` instead of `names`

  ProductName[[i]] <- webpage %>%
    html_nodes('.a-size-large.product-title-word-break') %>%
    html_text()
  ProductName[[i]] <- rep(ProductName[[i]], length.out = length(names_list[[i]]))

  rate <- webpage %>% html_nodes(".a-icon-alt") %>% html_text()
  rati <- rate[!grepl("Previous page|Next page|Previous set of slides|Next set of slides", rate, ignore
  rat <- gsub(" out of 5 stars", "", rati)
  rats <- rat
  if (length(rats) > length(name)) {
    rats <- tail(rats, length(name))
  } else if (length(rats) < length(name)) {
    rats <- c(rats, rep(NA, length(name) - length(rats)))
  }

  Ratings[[i]] <- rats

  dat <- webpage %>% html_nodes(".a-size-base.a-color-secondary.review-date") %>% html_text()
  date <- gsub("Reviewed.*on ", "", dat)
  Dates[[i]] <- date

  titl <- webpage %>% html_nodes(".a-size-base.review-title.a-color-base.review-title-content.a-text-bo
  tit <- gsub("Reviewed.*on ", "", titl)
  ti <- gsub(".*stars\\s*", "", tit)
  t <- gsub("\\s+", " ", ti)
```

```
  Title[[i]] <- t

  tex <- webpage %>% html_nodes(".a-expander-content.reviewText.review-text-content.a-expander-partial-
  te <- gsub("\\n", " ", tex)
  t <- gsub("\\s+", " ", te)
  text[[i]] <- trimws(t)
}
```

```
cate <- c("Category 1: MEN'S CLOTHING", "Category 2", "Category 3", "Category 4", "Category 5")
category <- vector("list", length(cate))
for (i in seq_along(cate)) {
    category[[i]] <- cate[i]
}
```

```
for (i in seq_along(cate)) {
    category[[i]] <- rep(category[[i]], length.out = length(names_list[[i]]))  # Rename names to names_
}
```

```
names_list <- list(rep("Name 1", 3), rep("Name 2", 4), rep("Name 3", 5), rep("Name 4", 6), rep("Name 5"

productnumbe <- c("Product 1", "Product 2", "Product 3", "Product 4", "Product 5", "Product 6", "Produc
productnumber <- vector("list", length(productnumbe))
for (i in seq_along(productnumbe)) {
    productnumber[[i]] <- productnumbe[i]
}

urls <- c("url1", "url2", "url3", "url4", "url5")

for (i in seq_along(urls)) {
    if (i <= length(names_list)) {
        productnumber[[i]] <- rep(productnumber[[i]], length.out = length(names_list[[i]]))
    }
}
```

```
names_list <- list()


names_list <- list(c("User1", "User2", "User3"))

min_length <- min(length(category[[1]]), length(productnumber[[1]]), length(ProductName[[1]]),
                  length(names_list[[1]]), length(Ratings[[1]]), length(Dates[[1]]),
                  length(Title[[1]]), length(text[[1]]))

category[[1]] <- category[[1]][1:min_length]
productnumber[[1]] <- productnumber[[1]][1:min_length]
ProductName[[1]] <- ProductName[[1]][1:min_length]
names_list[[1]] <- names_list[[1]][1:min_length]
Ratings[[1]] <- Ratings[[1]][1:min_length]
Dates[[1]] <- Dates[[1]][1:min_length]
Title[[1]] <- Title[[1]][1:min_length]
text[[1]] <- text[[1]][1:min_length]

cloth1 <- data.frame(
  Category = category[[1]],
  Product_number = productnumber[[1]],
```

```
  Name_of_Product = ProductName[[1]],
  Username = names_list[[1]],
  Rating = Ratings[[1]],
  Date = Dates[[1]],
  Title_of_Review = Title[[1]],
  Text_of_Review = text[[1]],
  stringsAsFactors = FALSE
)

head(cloth1, 50)
```

```
##                     Category Product_number
## 1 Category 1: MEN'S CLOTHING      Product 1
## 2 Category 1: MEN'S CLOTHING      Product 1
## 3 Category 1: MEN'S CLOTHING      Product 1
##                                                                                Name_of_Product
## 1         Hanes Men's Hoodie, EcoSmart Fleece Hoodie, Hooded Sweatshirt for Men
## 2         Hanes Men's Hoodie, EcoSmart Fleece Hoodie, Hooded Sweatshirt for Men
## 3         Hanes Men's Hoodie, EcoSmart Fleece Hoodie, Hooded Sweatshirt for Men
##   Username Rating             Date
## 1    User1    5.0 December 10, 2024
## 2    User2    5.0  November 6, 2024
## 3    User3    4.0  October 15, 2024
##                                        Title_of_Review
## 1                                    Super comfortable!
## 2 Great Quality and Value - Super Warm and Comfortable!
## 3               Comfy, affordable, and could be better
##
## 1
## 2 The Hanes Men’s EcoSmart Fleece Hoodie has exceeded my expectations for an affordable, everyday sw
## 3
```