727-HW3

Yuchen Ding

2024-10-08

Web Scraping

```
url <- read_html("https://en.wikipedia.org/wiki/Grand_Boulevard,_Chicago")
tables <- html_table(url, fill = TRUE)
str(tables)
```

```
## List of 7
##  $ : tibble [26 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ Grand Boulevard: chr [1:26] "Community area" "Community Area 38 - Grand Boulevard" "The Harold
##   ..$ Grand Boulevard: chr [1:26] "Community area" "Community Area 38 - Grand Boulevard" "The Harold
##  $ : tibble [11 x 4] (S3: tbl_df/tbl/data.frame)
##   ..$ Census
##   ..$ Pop.
##   ..$ .mw-parser-output .sr-only{border:0;clip:rect(0,0,0,0);clip-path:polygon(0px 0px,0px 0px,0px 0p
##   ..$ %±
##  $ : tibble [6 x 17] (S3: tbl_df/tbl/data.frame)
##   ..$ .mw-parser-output .navbar{display:inline;font-size:88%;font-weight:normal}.mw-parser-output .na
##   ..$ .mw-parser-output .navbar{display:inline;font-size:88%;font-weight:normal}.mw-parser-output .na
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##   ..$
##  $ : tibble [5 x 3] (S3: tbl_df/tbl/data.frame)
##   ..$ X1: chr [1:5] "Armour Square, Chicago" "" "Fuller Park, Chicago" "" ...
##   ..$ X2: chr [1:5] "Douglas, Chicago" "" "Grand Boulevard, Chicago" "" ...
##   ..$ X3: chr [1:5] "Oakland, Chicago" "" "Kenwood, Chicago" "" ...
##  $ : tibble [9 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ vteCommunity areas in Chicago: chr [1:9] "Far North" "Northwest" "North" "Central" ...
##   ..$ vteCommunity areas in Chicago: chr [1:9] "Rogers Park\nWest Ridge\nUptown\nLincoln Square\nEdi
##  $ : tibble [2 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ vteNeighborhoods in Chicago: chr [1:2] "Recognized by the city" "Other districts and areas rec
##   ..$ vteNeighborhoods in Chicago: chr [1:2] "Albany Park\nAndersonville\nArcher Heights\nAshburn\nA
```

```
##  $ : tibble [2 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ vte Chicago: chr [1:2] "Architecture\nBeaches\nClimate\nColleges and universities\nCommunity ar
##   ..$ vte Chicago: chr [1:2] "Architecture\nBeaches\nClimate\nColleges and universities\nCommunity ar
```

```
pop <- tables[[2]][c("Census","Pop.")]
pop <- pop[-11, ]
pop$area <- "Grand_Boulevard"
```

Expanding to More Pages

```
adjacent <- tables[[4]][c("X1","X2","X3")]
adjacent <- adjacent[-c(2,4), ]
adjacent
```

```
## # A tibble: 3 x 3
##   X1                     X2                       X3
##   <chr>                  <chr>                    <chr>
## 1 Armour Square, Chicago Douglas, Chicago         Oakland, Chicago
## 2 Fuller Park, Chicago   Grand Boulevard, Chicago Kenwood, Chicago
## 3 New City, Chicago      Washington Park, Chicago Hyde Park, Chicago
```

```
east_of_grand <- c(adjacent$X3)
print(east_of_grand)
```

```
## [1] "Oakland, Chicago"   "Kenwood, Chicago"   "Hyde Park, Chicago"
```

```
east_of_grand <- gsub(" ", "_", east_of_grand)
east_of_grand
```

```
## [1] "Oakland,_Chicago"   "Kenwood,_Chicago"   "Hyde_Park,_Chicago"
```

```
pops <- pop

url0 <- "https://en.wikipedia.org/wiki/"

for (i in east_of_grand){
  urls <- paste0(url0, i)
  pages <- read_html(urls)
  tables <- html_table(pages, fill = TRUE)
  pop_tables <- tables[[2]][c("Census","Pop.")]
  pop_tables <- pop_tables[-nrow(pop_tables), ]

  pop_tables <- pop_tables %>%
    mutate(area = i)

  pops <- bind_rows(pops, pop_tables)
}

pops
```

```
## # A tibble: 42 x 3
##    Census Pop.    area
##    <chr>  <chr>   <chr>
##  1 1930   87,005  Grand_Boulevard
##  2 1940   103,256 Grand_Boulevard
##  3 1950   114,557 Grand_Boulevard
```

```
##  4 1960   80,036  Grand_Boulevard
##  5 1970   80,166  Grand_Boulevard
##  6 1980   53,741  Grand_Boulevard
##  7 1990   35,897  Grand_Boulevard
##  8 2000   28,006  Grand_Boulevard
##  9 2010   21,929  Grand_Boulevard
## 10 2020   24,589  Grand_Boulevard
## # i 32 more rows
```

Scraping and Analyzing Text Data

```
library(rvest)
library(dplyr)
library(tidytext)

locations <- c("Armour Square, Chicago",    "Douglas, Chicago", "Oakland, Chicago", "Fuller Park, Chicag
locations <- gsub(" ", "_", locations)
location_data <- tibble(Location = character(), Description = character())

for (i in locations){
  urls <- paste0(url0, i)
  pages <- read_html(urls)
  description <- pages %>%
    html_nodes("p") %>%
    html_text() %>%
    paste(collapse = " ")

  location_data <- location_data %>%
    add_row(Location = i, Description = description)
}

print(location_data)
```

```
## # A tibble: 9 x 2
##   Location                 Description
##   <chr>                    <chr>
## 1 Armour_Square,_Chicago   "\n Armour Square is a Chicago neighborhood on the c~
## 2 Douglas,_Chicago         "\n Douglas, on the South Side of Chicago, Illinois,~
## 3 Oakland,_Chicago         "Oakland, located on the South Side of Chicago, Illi~
## 4 Fuller_Park,_Chicago     "Fuller Park is the 37th of Chicago's 77 community a~
## 5 Grand_Boulevard,_Chicago "\n Grand Boulevard on the South Side of Chicago, Il~
## 6 Kenwood,_Chicago         "\n Kenwood, one of Chicago's 77 community areas, is~
## 7 New_City,_Chicago        "\n New City is one of Chicago's 77 official communi~
## 8 Washington_Park,_Chicago "Washington Park, Chicago may refer to:\n"
## 9 Hyde_Park,_Chicago       "\n Hyde Park is a neighborhood on the South Side of~
```

```
location_words <- location_data %>%
  unnest_tokens(word, Description)

data(stop_words)
location_words <- location_words %>%
  anti_join(stop_words, by = "word")

location_words %>%
```

```
  count(Location, word, sort = TRUE)
```

```
## # A tibble: 2,992 x 3
##    Location              word            n
##    <chr>                 <chr>       <int>
##  1 Hyde_Park,_Chicago    park           74
##  2 Hyde_Park,_Chicago    hyde           69
##  3 Hyde_Park,_Chicago    chicago        34
##  4 Fuller_Park,_Chicago  park           26
##  5 Fuller_Park,_Chicago  fuller         25
##  6 Oakland,_Chicago      oakland        25
##  7 Kenwood,_Chicago      kenwood        24
##  8 Hyde_Park,_Chicago    street         22
##  9 Douglas,_Chicago      bronzeville    21
## 10 Fuller_Park,_Chicago  2              21
## # i 2,982 more rows
```

```
counts <- location_words %>%
  count(Location, word, sort = TRUE) %>%
  group_by(Location) %>%
  top_n(10, n)

library(ggplot2)
ggplot(counts,
       aes(x = reorder(word, n), y = n,
           fill = Location)) +
  geom_col() +
  facet_wrap(~ Location, scales = "free_y") +
  labs(x = "Word", y = "Frequency") +
  coord_flip() +
  theme_minimal() +
  ggtitle("Most Common Words by Location")
```