727HW3

Yuchen Ding and Kailin Xu

2024-10-15

Github

https://github.com/Asuka-nn/727-HW3.git

```
library(xml2)
library(rvest)
library(tidyverse)
library(robotstxt)
library(dplyr)
library(tidytext)
library(scales)
```

Web Scraping

```
paths_allowed('https://en.wikipedia.org/wiki/Grand_Boulevard,_Chicago')
```

```
##  en.wikipedia.org
```

```
## [1] TRUE
```

```
gbc <- read_html('https://en.wikipedia.org/wiki/Grand_Boulevard,_Chicago')
hp <- html_nodes(gbc, xpath ='//*[@id="mw-content-text"]/div[1]/table[2]')
str(hp)
```

```
## List of 1
##  $ :List of 2
##   ..$ node:<externalptr>
##   ..$ doc :<externalptr>
##   ..- attr(*, "class")= chr "xml_node"
##  - attr(*, "class")= chr "xml_nodeset"
```

```
his_p <- html_table(hp)
his_p2 <- as.data.frame(his_p)
hp_clean <- his_p2[-nrow(his_p2),-3]
names(hp_clean) <- c("year","population","percent change")
hp_clean$area <- "Grand_Boulevard"
print(hp_clean)
```

```
##    year population percent change            area
## 1  1930     87,005               - Grand_Boulevard
## 2  1940    103,256          18.7% Grand_Boulevard
## 3  1950    114,557          10.9% Grand_Boulevard
## 4  1960     80,036         -30.1% Grand_Boulevard
## 5  1970     80,166           0.2% Grand_Boulevard
## 6  1980     53,741         -33.0% Grand_Boulevard
## 7  1990     35,897         -33.2% Grand_Boulevard
## 8  2000     28,006         -22.0% Grand_Boulevard
## 9  2010     21,929         -21.7% Grand_Boulevard
## 10 2020     24,589          12.1% Grand_Boulevard
```

Expanding to More Pages

```
adjacent <- html_nodes(gbc, xpath =
                         '//*[@id="mw-content-text"]/div[1]/div[13]/table/tbody/tr[2]/td/div/table')
adjacent2 <- html_table(adjacent)
adjacent3 <- as.data.frame(adjacent2)
adjacent_east <- adjacent3[,3]
adjacent_east <- adjacent_east[adjacent_east != ""]
adjacent_east
```

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

```
east <- gsub(" ", "_", adjacent_east)
east
```

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

```
pops <- hp_clean
```

```
for(i in east) {
  url <- paste0("https://en.wikipedia.org/wiki/", i)
  print(url)
}
```

```
## [1] "https://en.wikipedia.org/wiki/Oakland,_Chicago"
## [1] "https://en.wikipedia.org/wiki/Kenwood,_Chicago"
## [1] "https://en.wikipedia.org/wiki/Hyde_Park,_Chicago"
```

```
for(i in east) {
  url <- paste0("https://en.wikipedia.org/wiki/", i)
  src <- read_html(url)

  hispp <- html_nodes(src, xpath ='//*[@id="mw-content-text"]/div[1]/table[2]')
  hispp2 <- html_table(hispp)
  hispp3 <- as.data.frame(hispp2)
  hispp_clean <- hispp3[-nrow(hispp3),-3]
  names(hispp_clean) <- c("year","population","percent change")
```

```
  hispp_clean <- hispp_clean %>% mutate(area = i)
  pops <- bind_rows(pops,hispp_clean)

}
print(pops)
```

```
##    year population percent change              area
## 1  1930     87,005              -   Grand_Boulevard
## 2  1940    103,256          18.7%   Grand_Boulevard
## 3  1950    114,557          10.9%   Grand_Boulevard
## 4  1960     80,036         -30.1%   Grand_Boulevard
## 5  1970     80,166           0.2%   Grand_Boulevard
## 6  1980     53,741         -33.0%   Grand_Boulevard
## 7  1990     35,897         -33.2%   Grand_Boulevard
## 8  2000     28,006         -22.0%   Grand_Boulevard
## 9  2010     21,929         -21.7%   Grand_Boulevard
## 10 2020     24,589          12.1%   Grand_Boulevard
## 11 1910     13,763              -  Oakland,_Chicago
## 12 1920     16,540          20.2%  Oakland,_Chicago
## 13 1930     14,962          -9.5%  Oakland,_Chicago
## 14 1940     14,500          -3.1%  Oakland,_Chicago
## 15 1950     24,464          68.7%  Oakland,_Chicago
## 16 1960     24,378          -0.4%  Oakland,_Chicago
## 17 1970     18,291         -25.0%  Oakland,_Chicago
## 18 1980     16,748          -8.4%  Oakland,_Chicago
## 19 1990      8,197         -51.1%  Oakland,_Chicago
## 20 2000      6,110         -25.5%  Oakland,_Chicago
## 21 2010      5,918          -3.1%  Oakland,_Chicago
## 22 2020      6,799          14.9%  Oakland,_Chicago
## 23 1930     26,942              -  Kenwood,_Chicago
## 24 1940     29,611           9.9%  Kenwood,_Chicago
## 25 1950     35,705          20.6%  Kenwood,_Chicago
## 26 1960     41,533          16.3%  Kenwood,_Chicago
## 27 1970     26,890         -35.3%  Kenwood,_Chicago
## 28 1980     21,974         -18.3%  Kenwood,_Chicago
## 29 1990     18,178         -17.3%  Kenwood,_Chicago
## 30 2000     18,363           1.0%  Kenwood,_Chicago
## 31 2010     17,841          -2.8%  Kenwood,_Chicago
## 32 2020     19,116           7.1%  Kenwood,_Chicago
## 33 1930     48,017              - Hyde_Park,_Chicago
## 34 1940     50,550           5.3% Hyde_Park,_Chicago
## 35 1950     55,206           9.2% Hyde_Park,_Chicago
## 36 1960     45,577         -17.4% Hyde_Park,_Chicago
## 37 1970     33,531         -26.4% Hyde_Park,_Chicago
## 38 1980     31,198          -7.0% Hyde_Park,_Chicago
## 39 1990     28,630          -8.2% Hyde_Park,_Chicago
## 40 2000     29,920           4.5% Hyde_Park,_Chicago
## 41 2010     25,681         -14.2% Hyde_Park,_Chicago
## 42 2020     29,456          14.7% Hyde_Park,_Chicago
```

Use cbind

```
pops2 <- data.frame(matrix(NA, nrow = 12, ncol = 0))
loc <- c("Grand_Boulevard,_Chicago", east)

for (i in loc) {
  url2 <- paste0("https://en.wikipedia.org/wiki/", i)
  src2 <- read_html(url2)

  hispp_2 <- html_nodes(src2, xpath ='//*[@id="mw-content-text"]/div[1]/table[2]')
  hispp2_2 <- html_table(hispp_2)
  hispp3_2 <- as.data.frame(hispp2_2)
  hispp_clean_2 <- hispp3_2[-nrow(hispp3_2), -3]
  names(hispp_clean_2) <- c("year", "population", "percent change")

  current_rows <- nrow(hispp_clean_2)
  if(current_rows < 12) {
    missing_rows <- 12 - current_rows
  hispp_clean_2 <- rbind(hispp_clean_2,
                         setNames(data.frame(matrix(NA, nrow = missing_rows, ncol = 3)),
                                  names(hispp_clean_2)))}
  hispp_clean_2$area <- i

  pops2 <- cbind(pops2, hispp_clean_2)
}

  print(pops2)
```

```
##    year population percent change                    area year population
## 1  1930     87,005              - Grand_Boulevard,_Chicago 1910     13,763
## 2  1940    103,256          18.7% Grand_Boulevard,_Chicago 1920     16,540
## 3  1950    114,557          10.9% Grand_Boulevard,_Chicago 1930     14,962
## 4  1960     80,036         -30.1% Grand_Boulevard,_Chicago 1940     14,500
## 5  1970     80,166           0.2% Grand_Boulevard,_Chicago 1950     24,464
## 6  1980     53,741         -33.0% Grand_Boulevard,_Chicago 1960     24,378
## 7  1990     35,897         -33.2% Grand_Boulevard,_Chicago 1970     18,291
## 8  2000     28,006         -22.0% Grand_Boulevard,_Chicago 1980     16,748
## 9  2010     21,929         -21.7% Grand_Boulevard,_Chicago 1990      8,197
## 10 2020     24,589          12.1% Grand_Boulevard,_Chicago 2000      6,110
## 11 <NA>       <NA>           <NA> Grand_Boulevard,_Chicago 2010      5,918
## 12 <NA>       <NA>           <NA> Grand_Boulevard,_Chicago 2020      6,799
##    percent change             area year population percent change
## 1               - Oakland,_Chicago 1930     26,942              -
## 2           20.2% Oakland,_Chicago 1940     29,611           9.9%
## 3           -9.5% Oakland,_Chicago 1950     35,705          20.6%
## 4           -3.1% Oakland,_Chicago 1960     41,533          16.3%
## 5           68.7% Oakland,_Chicago 1970     26,890         -35.3%
## 6           -0.4% Oakland,_Chicago 1980     21,974         -18.3%
## 7          -25.0% Oakland,_Chicago 1990     18,178         -17.3%
## 8           -8.4% Oakland,_Chicago 2000     18,363           1.0%
## 9          -51.1% Oakland,_Chicago 2010     17,841          -2.8%
## 10         -25.5% Oakland,_Chicago 2020     19,116           7.1%
## 11          -3.1% Oakland,_Chicago <NA>       <NA>           <NA>
```

```
## 12          14.9% Oakland,_Chicago <NA>       <NA>           <NA>
##                  area year population percent change              area
## 1  Kenwood,_Chicago 1930     48,017              - Hyde_Park,_Chicago
## 2  Kenwood,_Chicago 1940     50,550           5.3% Hyde_Park,_Chicago
## 3  Kenwood,_Chicago 1950     55,206           9.2% Hyde_Park,_Chicago
## 4  Kenwood,_Chicago 1960     45,577         -17.4% Hyde_Park,_Chicago
## 5  Kenwood,_Chicago 1970     33,531         -26.4% Hyde_Park,_Chicago
## 6  Kenwood,_Chicago 1980     31,198          -7.0% Hyde_Park,_Chicago
## 7  Kenwood,_Chicago 1990     28,630          -8.2% Hyde_Park,_Chicago
## 8  Kenwood,_Chicago 2000     29,920           4.5% Hyde_Park,_Chicago
## 9  Kenwood,_Chicago 2010     25,681         -14.2% Hyde_Park,_Chicago
## 10 Kenwood,_Chicago 2020     29,456          14.7% Hyde_Park,_Chicago
## 11 Kenwood,_Chicago <NA>       <NA>           <NA> Hyde_Park,_Chicago
## 12 Kenwood,_Chicago <NA>       <NA>           <NA> Hyde_Park,_Chicago
```

Scraping and Analyzing Text Data

```
wenzi <- html_nodes(gbc, xpath="//p")
wenzi2 <- html_text(wenzi)
descrip <- wenzi2 %>% paste(collapse = ' ')
print(descrip)
```

```
## [1] "\n Grand Boulevard on the South Side of Chicago, Illinois, is one of the city's Community Areas
King College in Englewood. A high school diploma had been earned by 85.5% of Grand Boulevard residents a
```

```
location_data <- tibble(Location = character(), Description = character())

locations <- c("Armour Square, Chicago",    "Douglas, Chicago", "Oakland, Chicago",
               "Fuller Park, Chicago",  "Grand Boulevard, Chicago", "Kenwood, Chicago",
               "New City, Chicago", "Washington Park, Chicago", "Hyde Park, Chicago")
locations <- gsub(" ", "_", locations)
location_data <- tibble(Location = character(), Description = character())

for(i in locations) {
  url <- paste0("https://en.wikipedia.org/wiki/", i)
  src <- read_html(url)

  des <- html_nodes(src, xpath="//p")
  des2 <- html_text(des)
  des3 <- des2 %>% paste(collapse = ' ')

  description_1 <- tibble(Location = i, Description = des3)
  location_data <- rbind(location_data, description_1)
}

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~
```

"Park" is the most common words used overall.

```
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
```

Similarities

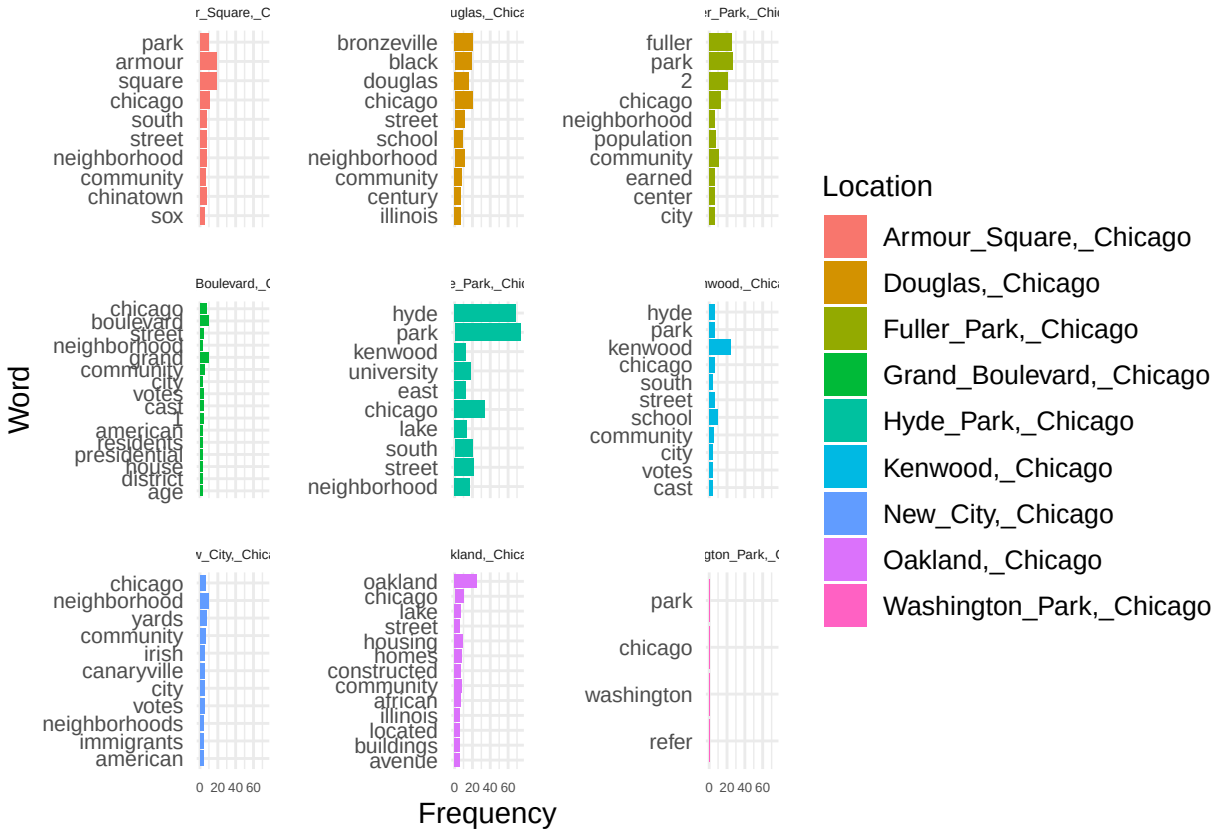
"Park" appears as one of the most frequent words across locations. "Chicago" is another common word across all locations, reflecting that all these neighborhoods are part of the broader Chicago area. "Community" and "Neighborhood" are frequently mentioned, indicating a focus on communal living.

Differiences

Place-specific terms: For example, "Hyde" in Hyde Park, "Kenwood" in Kenwood, and "Oakland" in Oakland are unique to their respective locations.
Size and Focus: Some areas, like Hyde Park and Grand Boulevard, show more diverse frequent words, which might suggest a richer historical or demographic narrative compared to smaller or less documented areas like Fuller Park or Oakland.

```
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() +
  theme(
    panel.spacing = unit(1, "lines"),
    axis.text.x = element_text(size = 5),
    axis.text.y = element_text(size = 8),
    strip.text = element_text(size = 5),
    legend.text = element_text(size = 10))
```




```
  ggtitle("Most Common Words by Location")
```

```
## $title
## [1] "Most Common Words by Location"
```

```
##
## attr(,"class")
## [1] "labels"
```

```
description_count <- location_words %>%
  count(Location, word, sort = TRUE) %>%
  group_by(Location) %>%
  mutate(proportion = n / sum(n)) %>%
  select(-n) %>%
  pivot_wider(names_from = Location, values_from = proportion) %>%
  pivot_longer(`Fuller_Park,_Chicago`:`Washington_Park,_Chicago`,
               names_to = "Location", values_to = "proportion")

head(description_count)
```

```
## # A tibble: 6 x 4
##   word  `Hyde_Park,_Chicago` Location                 proportion
##   <chr>                <dbl> <chr>                         <dbl>
## 1 park                0.0473 Fuller_Park,_Chicago        0.0429
## 2 park                0.0473 Oakland,_Chicago            0.00616
## 3 park                0.0473 Kenwood,_Chicago            0.0214
## 4 park                0.0473 Douglas,_Chicago            0.00447
## 5 park                0.0473 Armour_Square,_Chicago      0.0212
## 6 park                0.0473 Grand_Boulevard,_Chicago    0.00418
```

```
ggplot(description_count, aes(x = proportion, y = `Hyde_Park,_Chicago`,
                      color = abs(`Hyde_Park,_Chicago` - proportion))) +
  geom_abline(color = "gray40", lty = 2) +
  geom_jitter(alpha = 0.3, size = 1.5, width = 0.3, height = 0.3) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5, size = 2.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  scale_color_gradient(low = "blue", high = "red") +
  facet_wrap(~Location, ncol = 2) +
  theme(legend.position="none") +
  labs(y = "Hyde_Park,_Chicago", x = NULL)
```


Armour_Square,_Chicago
Douglas,_Chicago
Fuller_Park,_Chicago
Grand_Boulevard,_Chicago
Kenwood,_Chicago
New_City,_Chicago
Oakland,_Chicago
Washington_Park,_Chicago
Hyde_Park,_Chicago
10.0%
1.0%
0.1%
park
chicago
street
neighborhood
east
community
african
west
black
line
boulevard
school
homes
plan
cast
6
housing
5
catholic
university
south
lake
house
north
u
chicago's
illinois
center
city
votes
due
century
home
constructed
population
residents
percent
2
white
47th
selection
1
michigan
shore
kenwood
votes
home
public
located
barack
9
2012
7
historic
apartments
10
avenue
washington