727 HW3

Jay Kim

Web Scraping

```r
url <- "https://en.wikipedia.org/wiki/Grand_Boulevard,_Chicago"
page <- read_html(url)

# Extract the census population table
census_table <- page %>%
  html_element("table.us-census-pop") %>%
  html_table()

census_table <- census_table[1:10,c(1:2,4)]

str(census_table)
```

```
tibble [10 x 3] (S3: tbl_df/tbl/data.frame)
 $ Census: chr [1:10] "1930" "1940" "1950" "1960" ...
 $ Pop.  : chr [1:10] "87,005" "103,256" "114,557" "80,036" ...
 $ %±   : chr [1:10] "-" "18.7%" "10.9%" "-30.1%" ...
```

```r
print(census_table)
```

```
# A tibble: 10 x 3
   Census Pop.    `%±`
   <chr>  <chr>   <chr>
 1 1930   87,005  -
 2 1940   103,256 18.7%
 3 1950   114,557 10.9%
 4 1960   80,036  -30.1%
 5 1970   80,166  0.2%
```

 6 1980   53,741 -33.0%
 7 1990   35,897 -33.2%
 8 2000   28,006 -22.0%
 9 2010   21,929 -21.7%
10 2020   24,589 12.1%

Expanding to More Pages

```
src <- read_html(url)
nds <- html_nodes(src,
                  xpath = '//*[contains(concat( " ", @class, " " ), concat( " ", "navbox-odd"
names <- html_text(nds)
names[[1]]
```

```
[1] "Armour Square, Chicago\nDouglas, Chicago\nOakland, Chicago\n\n\n\nFuller Park, Chicago\n
```

```
# Extracted text
names_text <- names[[1]]

# Split by newline and clean up
neighborhood_names <- strsplit(names_text, "\n")[[1]] %>%
  trimws() %>%  # Remove leading/trailing whitespace
  .[. != ""]    # Remove empty strings

# Convert to URL format (replace spaces with underscores)
url_suffixes <- gsub(" ", "_", neighborhood_names)

print(url_suffixes)
```

```
[1] "Armour_Square,_Chicago"   "Douglas,_Chicago"
[3] "Oakland,_Chicago"         "Fuller_Park,_Chicago"
[5] "Grand_Boulevard,_Chicago" "Kenwood,_Chicago"
[7] "New_City,_Chicago"        "Washington_Park,_Chicago"
[9] "Hyde_Park,_Chicago"
```

```
# Fix the Washington Park URL
url_suffixes[url_suffixes == "Washington_Park,_Chicago"] <- "Washington_Park_(community_area)

# Base URL
```

```r
base_url <- "https://en.wikipedia.org/wiki/"

# Initialize an empty list to store tables
all_tables <- list()

# Loop through each neighborhood
for (i in seq_along(url_suffixes)) {
  full_url <- paste0(base_url, url_suffixes[i])

  cat("Scraping:", neighborhood_names[i], "\n")

  page <- read_html(full_url)

  # Extract the census population table
  census_table <- page %>%
    html_element("table.us-census-pop") %>%
    html_table()

  # Select first 10 rows and columns 1, 2, and 4
  census_table <- census_table[1:10, c(1:2, 4)]

  # Rename columns to include neighborhood name
  if (i == 1) {
    colnames(census_table) <- c("Year", paste0(neighborhood_names[i], c("_Pop", "_Change")))
  } else {
    census_table <- census_table[, 2:3]
    colnames(census_table) <- paste0(neighborhood_names[i], c("_Pop", "_Change"))
  }

  all_tables[[i]] <- census_table

  Sys.sleep(1)
}
```

```
Scraping: Armour Square, Chicago
Scraping: Douglas, Chicago
Scraping: Oakland, Chicago
Scraping: Fuller Park, Chicago
Scraping: Grand Boulevard, Chicago
Scraping: Kenwood, Chicago
Scraping: New City, Chicago
Scraping: Washington Park, Chicago
```

```
Scraping: Hyde Park, Chicago
```

```r
# Combine all tables side-by-side
combined_census_table <- do.call(cbind, all_tables)

print(combined_census_table)
```

```
   Year Armour Square, Chicago_Pop Armour Square, Chicago_Change
1  1930                     21,450                             -
2  1940                     18,472                        -13.9%
3  1950                     23,294                         26.1%
4  1960                     15,783                        -32.2%
5  1970                     13,063                        -17.2%
6  1980                     12,475                         -4.5%
7  1990                     10,801                        -13.4%
8  2000                     12,032                         11.4%
9  2010                     13,391                         11.3%
10 2020                     13,890                          3.7%
   Douglas, Chicago_Pop Douglas, Chicago_Change Oakland, Chicago_Pop
1                50,285                       -               13,763
2                53,124                    5.6%               16,540
3                78,745                   48.2%               14,962
4                52,325                  -33.6%               14,500
5                43,731                  -16.4%               24,464
6                35,700                  -18.4%               24,378
7                30,652                  -14.1%               18,291
8                26,470                  -13.6%               16,748
9                18,238                  -31.1%                8,197
10               20,291                   11.3%                6,110
   Oakland, Chicago_Change Fuller Park, Chicago_Pop Fuller Park, Chicago_Change
1                        -                   14,437                           -

2                    20.2%                   15,094                        4.6%
3                    -9.5%                   17,174                       13.8%
4                    -3.1%                   12,181                      -29.1%
5                    68.7%                    7,354                      -39.6%
6                    -0.4%                    5,832                      -20.7%
7                   -25.0%                    4,364                      -25.2%
8                    -8.4%                    3,420                      -21.6%
9                   -51.1%                    2,876                      -15.9%
10                  -25.5%                    2,567                      -10.7%
   Grand Boulevard, Chicago_Pop Grand Boulevard, Chicago_Change
```

1	87,005	-
2	103,256	18.7%
3	114,557	10.9%
4	80,036	-30.1%
5	80,166	0.2%
6	53,741	-33.0%
7	35,897	-33.2%
8	28,006	-22.0%
9	21,929	-21.7%
10	24,589	12.1%

	Kenwood, Chicago_Pop	Kenwood, Chicago_Change	New City, Chicago_Pop
1	26,942	-	87,103
2	29,611	9.9%	80,725
3	35,705	20.6%	75,917
4	41,533	16.3%	67,428
5	26,890	-35.3%	60,747
6	21,974	-18.3%	55,860
7	18,178	-17.3%	53,226
8	18,363	1.0%	51,721
9	17,841	-2.8%	44,377
10	19,116	7.1%	43,628

	New City, Chicago_Change	Washington Park, Chicago_Pop
1	-	44,016
2	-7.3%	52,736
3	-6.0%	56,856
4	-11.2%	43,690
5	-9.9%	46,024
6	-8.0%	31,935
7	-4.7%	19,425
8	-2.8%	14,146
9	-14.2%	11,717
10	-1.7%	12,707

	Washington Park, Chicago_Change	Hyde Park, Chicago_Pop
1	-	48,017
2	19.8%	50,550
3	7.8%	55,206
4	-23.2%	45,577
5	5.3%	33,531
6	-30.6%	31,198
7	-39.2%	28,630
8	-27.2%	29,920
9	-17.2%	25,681
10	8.4%	29,456

```
   Hyde Park, Chicago_Change
1                          -
2                       5.3%
3                       9.2%
4                     -17.4%
5                     -26.4%
6                      -7.0%
7                      -8.2%
8                       4.5%
9                     -14.2%
10                     14.7%
```

Scraping and Analyzing Text Data

```
src <- read_html(url)
nds <- html_nodes(src,
                  xpath = '//p')
textbody <- html_text(nds)
textbody <- textbody %>%
  paste(collapse = ' ') %>%
  gsub("\\s+", " ", .) %>%  # Replace multiple spaces with single space
  trimws()
textbody
```

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

```
# Initialize vectors to store data
locations <- c()
descriptions <- c()


# Loop through each neighborhood
for (i in seq_along(url_suffixes)) {
  full_url <- paste0(base_url, url_suffixes[i])

  cat("Scraping:", neighborhood_names[i], "\n")

  tryCatch({
    page <- read_html(full_url)
```

```
    # Extract all paragraph text
    textbody <- page %>%
      html_nodes(xpath = '//p') %>%
      html_text() %>%
      paste(collapse = ' ') %>%
      gsub("\\s+", " ", .) %>%
      trimws()

    # Store the data
    locations <- c(locations, neighborhood_names[i])
    descriptions <- c(descriptions, textbody)

  }, error = function(e) {
    cat("Error scraping", neighborhood_names[i], ":", e$message, "\n")
  })

  Sys.sleep(1)
}
```

```
Scraping: Armour Square, Chicago
Scraping: Douglas, Chicago
Scraping: Oakland, Chicago
Scraping: Fuller Park, Chicago
Scraping: Grand Boulevard, Chicago
Scraping: Kenwood, Chicago
Scraping: New City, Chicago
Scraping: Washington Park, Chicago
Scraping: Hyde Park, Chicago
```

```
# Create tibble
community_areas <- tibble(
  location = locations,
  description = descriptions
)

# 1. CREATE TOKENS - one token per row
community_areas_tokens <- community_areas %>%
  unnest_tokens(word, description)

print("Tokens created - first few rows:")
```

```
[1] "Tokens created - first few rows:"
```

```r
print(head(community_areas_tokens, 20))
```

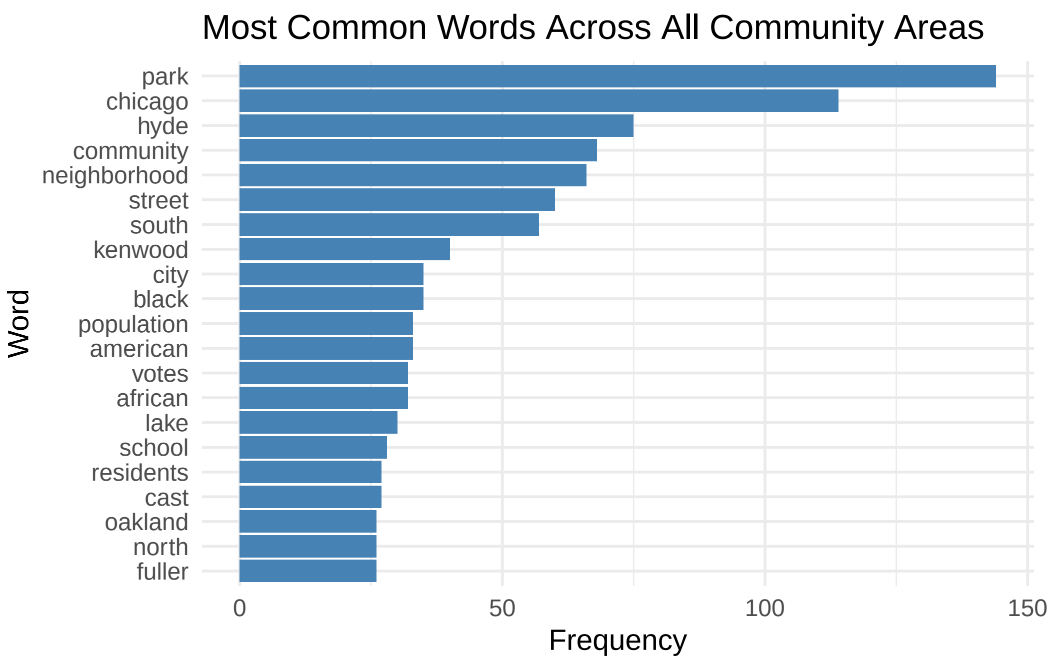
```
# A tibble: 20 x 2
   location                word
   <chr>                   <chr>
 1 Armour Square, Chicago armour
 2 Armour Square, Chicago square
 3 Armour Square, Chicago is
 4 Armour Square, Chicago a
 5 Armour Square, Chicago chicago
 6 Armour Square, Chicago neighborhood
 7 Armour Square, Chicago on
 8 Armour Square, Chicago the
 9 Armour Square, Chicago city's
10 Armour Square, Chicago south
11 Armour Square, Chicago side
12 Armour Square, Chicago as
13 Armour Square, Chicago well
14 Armour Square, Chicago as
15 Armour Square, Chicago a
16 Armour Square, Chicago larger
17 Armour Square, Chicago officially
18 Armour Square, Chicago defined
19 Armour Square, Chicago community
20 Armour Square, Chicago area
```

```r
# 2. REMOVE STOP WORDS
data("stop_words") # Call in stopword dataset
community_areas_clean <- community_areas_tokens %>%
  anti_join(stop_words, by = "word") %>%
  filter(!str_detect(word, "^\\d+$"))  # Remove numbers

# 3. MOST COMMON WORDS OVERALL
top_words_overall <- community_areas_clean %>%
  count(word, sort = TRUE) %>%
  top_n(20, n)

# Plot most common words overall
ggplot(top_words_overall, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "steelblue") +
```

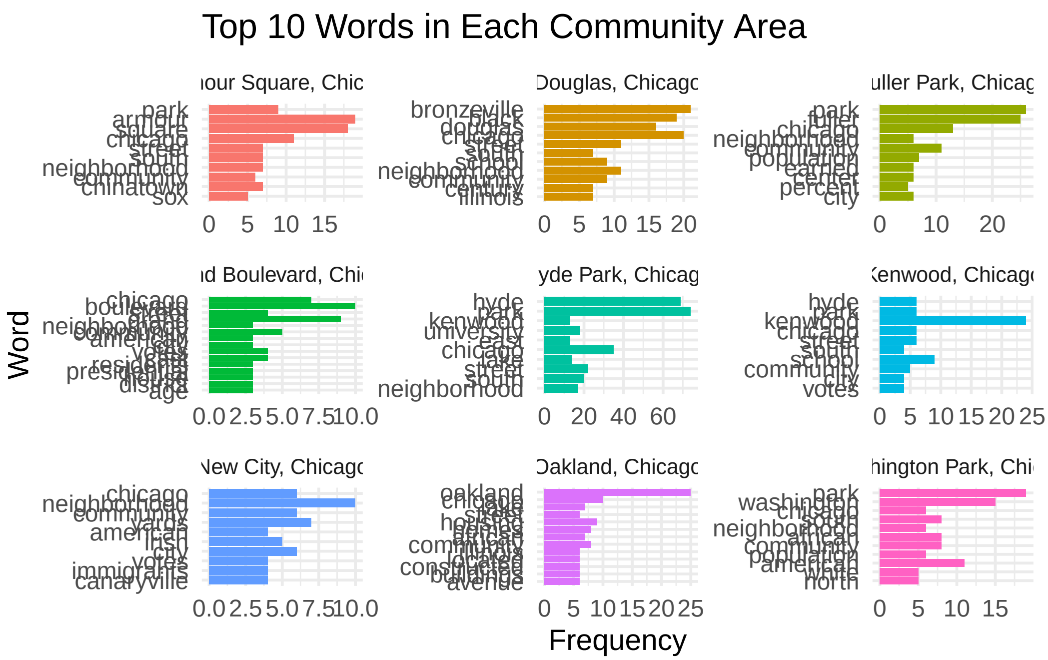
```
  coord_flip() +
  labs(title = "Most Common Words Across All Community Areas",
       x = "Word",
       y = "Frequency") +
  theme_minimal()
```




```
# 4. MOST COMMON WORDS WITHIN EACH LOCATION
top_words_by_location <- community_areas_clean %>%
  group_by(location) %>%
  count(word, sort = TRUE) %>%
  top_n(10, n) %>%
  ungroup()

# Plot most common words by location
ggplot(top_words_by_location, aes(x = reorder(word, n), y = n, fill = location)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  facet_wrap(~location, scales = "free") +
  labs(title = "Top 10 Words in Each Community Area",
       x = "Word",
       y = "Frequency") +
  theme_minimal() +
```

```
  theme(strip.text = element_text(size = 8))
```




```
# 5. SIMILARITIES ANALYSIS
# Find words that appear in multiple locations
word_locations <- community_areas_clean %>%
  distinct(location, word) %>%
  count(word) %>%
  arrange(desc(n))

common_words <- word_locations %>%
  filter(n >= 5)  # Words appearing in 5+ locations

print("Words appearing in most locations (similarities):")
```

```
[1] "Words appearing in most locations (similarities):"
```

```
print(common_words)
```

```
# A tibble: 100 x 2
   word             n
   <chr>        <int>
```

```
 1 boulevard        9
 2 chicago          9
 3 city             9
 4 community        9
 5 district         9
 6 neighborhood     9
 7 north            9
 8 park             9
 9 railroad         9
10 residents        9
# i 90 more rows
```

```r
# 6. DIFFERENCES ANALYSIS - Unique words per location
unique_words <- community_areas_clean %>%
  group_by(location) %>%
  count(word) %>%
  arrange(desc(n)) %>%
  slice(1:5) %>%
  ungroup()

print("Top unique/distinctive words per location:")
```

```
[1] "Top unique/distinctive words per location:"
```

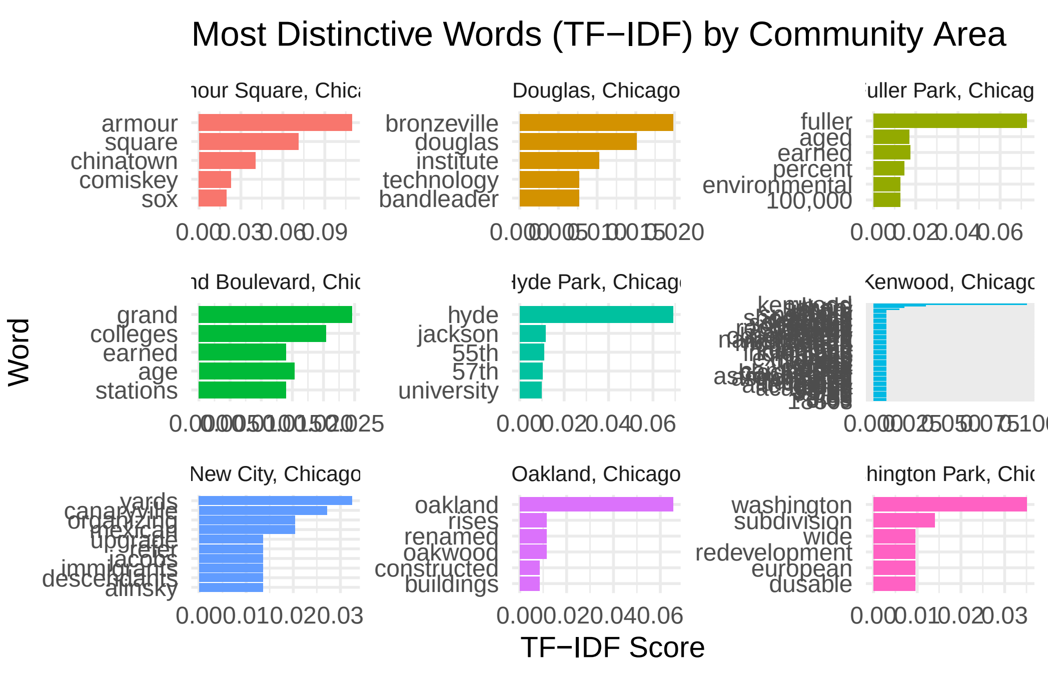
```r
print(unique_words)
```

```
# A tibble: 45 x 3
   location                word             n
   <chr>                   <chr>        <int>
 1 Armour Square, Chicago armour          19
 2 Armour Square, Chicago square          18
 3 Armour Square, Chicago chicago         11
 4 Armour Square, Chicago park             9
 5 Armour Square, Chicago chinatown        7
 6 Douglas, Chicago       bronzeville     21
 7 Douglas, Chicago       chicago         20
 8 Douglas, Chicago       black           19
 9 Douglas, Chicago       douglas         16
10 Douglas, Chicago       neighborhood    11
# i 35 more rows
```

```r
# TF-IDF Analysis for differences
community_areas_tfidf <- community_areas_clean %>%
  count(location, word) %>%
  bind_tf_idf(word, location, n) %>%
  arrange(desc(tf_idf))


# Plot TF-IDF
top_tfidf <- community_areas_tfidf %>%
  group_by(location) %>%
  top_n(5, tf_idf) %>%
  ungroup()

ggplot(top_tfidf, aes(x = reorder(word, tf_idf), y = tf_idf, fill = location)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  facet_wrap(~location, scales = "free") +
  labs(title = "Most Distinctive Words (TF-IDF) by Community Area",
       x = "Word",
       y = "TF-IDF Score") +
  theme_minimal() +
  theme(strip.text = element_text(size = 8))
```



```r
# 7. SUMMARY OF SIMILARITIES AND DIFFERENCES
cat("\n=== SIMILARITIES ===\n")
```

```
=== SIMILARITIES ===
```

```r
cat("Common themes across locations:\n")
```

```
Common themes across locations:
```

```r
print(common_words %>% head(10))
```

```
# A tibble: 10 x 2
   word            n
   <chr>       <int>
 1 boulevard       9
 2 chicago         9
 3 city            9
 4 community       9
 5 district        9
 6 neighborhood    9
 7 north           9
 8 park            9
 9 railroad        9
10 residents       9
```

```r
cat("\n=== DIFFERENCES ===\n")
```

```
=== DIFFERENCES ===
```

```r
cat("Distinctive characteristics (high TF-IDF):\n")
```

```
Distinctive characteristics (high TF-IDF):
```

```r
community_areas_tfidf %>%
  group_by(location) %>%
  top_n(3, tf_idf) %>%
  select(location, word, tf_idf) %>%
  print(n = 30)
```

```
# A tibble: 32 x 3
# Groups:   location [9]
   location                  word         tf_idf
   <chr>                     <chr>         <dbl>
 1 Armour Square, Chicago   armour      0.110
 2 Kenwood, Chicago         kenwood     0.0988
 3 Fuller Park, Chicago     fuller      0.0726
 4 Armour Square, Chicago   square      0.0711
 5 Hyde Park, Chicago       hyde        0.0690
 6 Oakland, Chicago         oakland     0.0654
 7 Armour Square, Chicago   chinatown   0.0404
 8 Washington Park, Chicago washington  0.0351
 9 Kenwood, Chicago         hyde        0.0338
10 New City, Chicago        yards       0.0325
11 New City, Chicago        canaryville 0.0271
12 Grand Boulevard, Chicago grand       0.0246
13 Grand Boulevard, Chicago colleges    0.0204
14 New City, Chicago        mexican     0.0203
15 New City, Chicago        organizing  0.0203
16 Kenwood, Chicago         school      0.0198
17 Douglas, Chicago         bronzeville 0.0198
18 Fuller Park, Chicago     earned      0.0174
19 Fuller Park, Chicago     aged        0.0170
20 Grand Boulevard, Chicago age         0.0153
21 Douglas, Chicago         douglas     0.0151
22 Washington Park, Chicago subdivision 0.0140
23 Hyde Park, Chicago       jackson     0.0117
24 Oakland, Chicago         oakwood     0.0115
25 Oakland, Chicago         renamed     0.0115
26 Oakland, Chicago         rises       0.0115
27 Hyde Park, Chicago       55th        0.0110
28 Douglas, Chicago         institute   0.0102
29 Washington Park, Chicago dusable     0.00960
30 Washington Park, Chicago european    0.00960
# i 2 more rows
```