HW3

KailinXu

2024-10-04

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

```
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")
print(hp_clean)
```

```
##    year population percent change
## 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%
```

```
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")

  pops <- rbind(pops,hispp_clean)
}
str(pops)
```

```
## 'data.frame':    42 obs. of  3 variables:
##  $ year          : chr  "1930" "1940" "1950" "1960" ...
##  $ population    : chr  "87,005" "103,256" "114,557" "80,036" ...
##  $ percent change: chr  "-" "18.7%" "10.9%" "-30.1%" ...
```

```
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
```

```
descrip2 <- tibble(location = 'Grand_Boulevard', description = descrip )
description_final <- descrip2

for(i in east) {
  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)
  description_final <- rbind(description_final, description_1)
}
```

park is the most common words used overall.

```
description_tokens <- description_final %>%
  unnest_tokens(word, description)
description_clean <- description_tokens %>% anti_join(stop_words)
```

```
## Joining with `by = join_by(word)`
```

```
description_clean %>% count(word, sort = TRUE) %>% head(5)
```

```
## # A tibble: 5 x 2
##   word        n
##   <chr>   <int>
## 1 park       85
## 2 hyde       75
## 3 chicago    57
## 4 kenwood    40
## 5 street     38
```

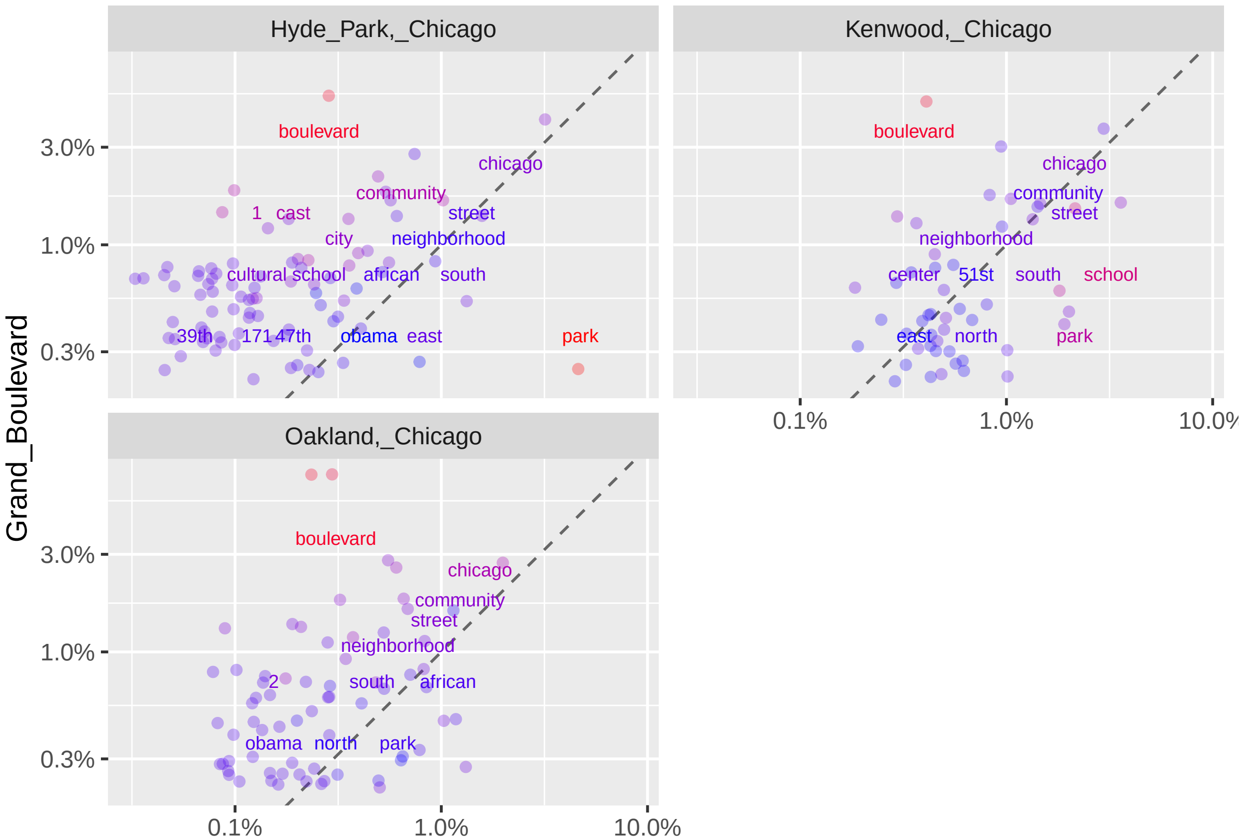
```
description_count <- description_clean %>%
  count(location, word, sort = TRUE) %>%
  group_by(location) %>%
  mutate(proportion = n / sum(n)) %>%
  select(-n) %>%
  pivot_wider(names_from = location, values_from = proportion) %>%
  pivot_longer(`Hyde_Park,_Chicago`:`Kenwood,_Chicago`,
               names_to = "location", values_to = "proportion")

head(description_count)
```

```
## # A tibble: 6 x 4
##   word  Grand_Boulevard location           proportion
##   <chr>           <dbl> <chr>                   <dbl>
## 1 park          0.00418 Hyde_Park,_Chicago    0.0473
## 2 park          0.00418 Oakland,_Chicago      0.00616
## 3 park          0.00418 Kenwood,_Chicago      0.0214
## 4 hyde         NA       Hyde_Park,_Chicago    0.0441
## 5 hyde         NA       Oakland,_Chicago     NA
## 6 hyde         NA       Kenwood,_Chicago      0.0214
```

```
ggplot(description_count, aes(x = proportion, y = `Grand_Boulevard`,
                      color = abs(`Grand_Boulevard` - 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 = "Grand_Boulevard", x = NULL)
```

```
## Warning: Removed 3218 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 3218 rows containing missing values or values outside the scale range
## (`geom_text()`).
```


Hyde_Park,_Chicago
Kenwood,_Chicago
Oakland,_Chicago
Grand_Boulevard
3.0%
1.0%
0.3%
0.1%
1.0%
10.0%
boulevard
chicago
community
1
cast
street
city
neighborhood
cultural
school
african
south
39th
171
47th
obama
east
park
center
51st
north
2