Canadian Prime Ministers*

Julie Nguyen

5 March 2023

0.1 Introduction

This paper serves as a practice of gathering data through web scraping. The practice aims at reproducing the list of Canadian Prime Ministers from 1815 to present using rvest (Wickham (2022)) and Selector Gadget.

0.2 Data and Findings

The information of Canadian Prime Ministers is gathered from Wikipedia using Selector Gadget. Other packages are also utilizes such as janitor (Firke (2021)), tidyverse (Wickham et al. (2019)) for cleaning and kableExtra (Zhu (2021)) for visualization.

I used set.seed() and sample() to simulate the list to see how the table will look like.

```
# A tibble: 10 x 4
   prime_minister birth_year death_year party
   <chr>               <int>      <int> <chr>
 1 Ida                  1783       1874 Conservative
 2 Florence             1786       1850 Other
 3 Joyce                1822       1913 Conservative
 4 Amanda               1828       1902 Progressive Conservative
 5 Paul                 1854       1936 Other
 6 William              1859       1943 Other
 7 Jerry                1914       1973 Progressive Conservative
 8 Lillian              1962       2055 Other
 9 Susan                1966       2035 Conservative
10 Scott                1978       2033 Liberal
```

The data is first scraped from the site using xpath, and rvest (Wickham (2022)) is used to download and save the input.

```
library(rvest)
library(tidyverse)
library(xml2)

raw_data <-
  read_html("https://en.wikipedia.org/wiki/List_of_prime_ministers_of_Canada")

write_html(raw_data, "C:/Users/anjul/OneDrive/Documents/iSchool/Winter 2023/INF312/Canadian-Prime-
```

Since the table consists of merge cells with notes, I removed those cells to choose the most important information that I need, including name, birth and death year, term of office and political party. I cleaned each category separately then merge them together to produce the final table.

*Code and data are available at:

```
library(rvest)
library(tidyverse)
library(janitor)
library(kableExtra)
library(knitr)

raw_data <- read_html(here::here("inputs/pms.html"))

parse_data <-
  raw_data |>
  html_element(xpath = '//*[@id="mw-content-text"]/div[1]/table[2]') |>
  html_table()

# Clean table
cleaned_data <- parse_data[-c(2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40,42,44,46,48
  clean_names() %>%
  mutate(no = rep(1:28))

# Extract birth and death year
dob <- cleaned_data %>%
  separate(
    name_birth_death,
    into = c("name", "date"),
    sep = "\\(",
    extra = "merge") %>%
  mutate(born = str_extract(date, "[[:digit:]]{4}-[[:digit:]]{4}"),
         alive = str_extract(date, "b.[[:space:]][[:digit:]]{4}")) %>%
  select(name, born, alive)

cleaned_dob <- dob %>%
  separate(born, into = c("birth", "died"),
           sep = "-") %>%
  mutate(born = str_remove_all(alive, "b.[[:space:]]"),
         birth = if_else(!is.na(alive), born, birth)) %>%
  select(-c(born, alive))

# Clean office term
office_term_clean <- cleaned_data %>%
  mutate(term_of_office = str_extract(term_of_office, "[[:digit:]]{4}"),
         term_of_office_2 = str_extract(term_of_office_2, "[[:digit:]]{4}")
  ) %>%
  rename("office_start" = term_of_office, "office_end" = term_of_office_2) %>%
  select(office_start, office_end)

# Clean political party
party_clean <- cleaned_data %>%
  mutate(political_party = str_remove(political_party, "Ldr.[[:space:]][[:digit:]]{4}"),
         political_party = str_remove(political_party, "[()]")) %>%
  select(political_party)
```

Then, I visualized the table using kableExtra (Zhu (2021)) with 6 columns showing name, birth and death year, terms of office, and the political party.

Prime Minister	Birth year	Death year	Took office	Left office	Political Party
John A. Macdonald	1815	1891	1867	1873	Liberal–Conservative
Alexander Mackenzie	1822	1892	1873	1878	Liberal
John A. Macdonald	1815	1891	1878	1891	Liberal–Conservative
John Abbott	1821	1893	1891	1892	Liberal–Conservative
John Thompson	1845	1894	1892	1894	Liberal–Conservative
Mackenzie Bowell	1823	1917	1894	1896	Conservative
Charles Tupper	1821	1915	1896	1896	Conservative
Wilfrid Laurier	1841	1919	1896	1911	Liberal
Robert Borden	1854	1937	1911	1920	Government Unionist
Arthur Meighen	1874	1960	1920	1921	Conservative
William Lyon Mackenzie King	1874	1950	1921	1926	Liberal
Arthur Meighen	1874	1960	1926	1926	Conservative
William Lyon Mackenzie King	1874	1950	1926	1930	Liberal
R. B. Bennett	1870	1947	1930	1935	Conservative
William Lyon Mackenzie King	1874	1950	1935	1948	Liberal
Louis St. Laurent	1882	1973	1948	1957	Liberal
John Diefenbaker	1895	1979	1957	1963	Progressive Conservative
Lester B. Pearson	1897	1972	1963	1968	Liberal
Pierre Trudeau	1919	2000	1968	1979	Liberal
Joe Clark	1939	NA	1979	1980	Progressive Conservative
Pierre Trudeau	1919	2000	1980	1984	Liberal
John Turner	1929	2020	1984	1984	Liberal
Brian Mulroney	1939	NA	1984	1993	Progressive Conservative
Kim Campbell	1947	NA	1993	1993	Progressive Conservative
Jean Chrétien	1934	NA	1993	2003	Liberal
Paul Martin	1938	NA	2003	2006	Liberal
Stephen Harper	1959	NA	2006	2015	Conservative
Justin Trudeau	1971	NA	2015	NA	Liberal

0.3 Reflections

For me, this web scraping practice was both fun and challegnging. It provides me an alternative way to gather the data that I need besides API or download directing the dataset. The string extract and remove are definitely the hardest parts to deal with since I do not know the formula for the pattern. It took me a long time to separate de birth and death year from the name. I also find removing parenthesis and brackets in a string difficult. In the future, I hope to improve the terms of office, separating the day, month, and year, so that I can count the days the PMs run the office. I would also want to know more about how I can deal with merge cells.

References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Wickham, Hadley. 2022. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.