Tutorial 7*

Web Scraping Prime Ministers of Canada from Wikipedia

Samita Prabhasavat

25 February 2022

1 About the Data

1.1 Data Source

In this paper, I am interested in looking into how long the prime ministers of Canada lived, based on the year they were born. The data I use in this paper is from the list of prime ministers of Canada on Wikipedia. The page contains information, including name, year of birth, and year of death of every prime minister of Canada.

1.2 Data Gathering Process

The first step is to find the source of the data. As I want to know how long the prime minister of Canada lived, I need to find the website or the source that contains all information I want and those information should also be reliable and correct. Wikipedia is a popular encyclopedia often used as a tertiary source in papers. The Wikipedia page about the prime ministers of Canada is suitable to use in this paper as it provides all information I am interested in table format.

The second step is to download the page using rvest package as html file. SelctorGadget is also used to choose the information I want from the whole page. In this case, I choose the Name column which contains both name and birth-death year of each prime minister.

The third step is to clean the data to. There are many ways to clean the data mostly depending on the data itself and how we want the data to look like. In this paper, I start from filtering out the blank lines in the data frame so that I don't have to deal with unnecessary attributes. As the names and the years are now in the same column, I have to separate them in order to get three columns which are name, birth year, and death year. After getting all columns I need ready, I add another column to store the age of the prime ministers when they died. The values in this column are calculated by subtracting birth year from death year. The last thing I do is to check if all data are correct and in the right place. There are some values missing form the data frame, so I replace them manually.

*Code and data are available at: https://github.com/PSamita/Tutorial-Week-7.git

Table 1: Canada Prime Minister, by how old they were when they died

Prime Minister	Birth Year	Death Year	Age at Death
Sir John A. Macdonald	1815	1891	76
Alexander Mackenzie	1822	1892	70
Sir John Abbott	1821	1893	72
Sir John Thompson	1845	1894	49
Sir Mackenzie Bowell	1823	1917	94
Sir Charles Tupper	1821	1915	94
Sir Wilfrid Laurier	1841	1919	78
Sir Robert Borden	1854	1937	83
Arthur Meighen	1874	1960	86
William Lyon Mackenzie King	1874	1950	76
R. B. Bennett	1870	1947	77
Louis St. Laurent	1882	1973	91
John Diefenbaker	1895	1979	84
Lester B. Pearson	1897	1972	75
Pierre Trudeau	1919	2000	81
Joe Clark	1939	NA	NA
John Turner	1929	2020	91
Brian Mulroney	1939	NA	NA
Kim Campbell	1947	NA	NA
Jean Chrétien	1934	NA	NA
Paul Martin	1938	NA	NA
Stephen Harper	1959	NA	NA
Justin Trudeau	1971	NA	NA

2 About the Table

2.1 Findings

According to Table 1, Canada had 22 former prime ministers, and the current prime minister, Justin Trudeau, is the 23rd prime minister of Canada. Among 22 former prime ministers, 6 of them; Joe Clark, Brian Mulroney, Kim Campbell, Jean Chrétien, Paul Martin, and Stephen Harper, are still alive. Sir John Thompson died at the youngest age of 49 while Sir Mackenzie Bowell and Sir Charles Tupper died at the oldest age of 94. Seven former prime ministers died in their 70s, four former prime ministers died in their 80s, and four other former prime ministers died in their 90s.

3 Reflection

3.1 What took you longer than expected?

The data cleaning process took longer time than I thought. Even though the data on Wikipedia page looks somewhat similar to the case study we went through in the class, small differences make the data cleaning process different. It took me some times to decide on what to do with the data as there are many different ways that can lead to the same result. Some methods I tried didn't give the result I wanted so I had to change to another method. Moreover, the data from web scraping is much messier than the data from Toronto open data portal I previously worked with. The data from Toronto open data portal doesn't require much cleaning, I only changed the names of the columns and deal with some null values. On the other hand, the data from web scraping has to be cleaned starting from the format of the data itself which requires longer time to do.

3.2 When did it become fun?

The first part I found very fun is the web scraping part where I gather data from the Wikipedia page. I think it is interesting to explore how much I can do on the website. I spent some times using SelctorGadget to explore ways I can do to select the data I'm interested in and to see the limitations of how far I can explore the page.

Another part I enjoyed doing is the data cleaning process. As I mentioned earlier, there are many different ways to clean the data the can lead to the same result. I find it fun to explore each way and to try adjusting the code to make it less messy. I think finding the ways to clean the data is similar to trying to solve a math problem. There are many ways to approach one problem, and each approach or method has its own advantages and disadvantages. I don't think there are strict data cleaning rules or steps that can be used with every data in this world because every data is different. In order to choose an appropriate way to clean the data, you not only need to understand the nature of it but also need to understand what you want. The fun part for me is where I can explore those methods and choose the one I think fits the data the most.

3.3 What would you do differently next time you do this?

The next time I work with this data, I would like to explore more data cleaning methods to substitute with the part I manually replaced data. Even though the code works well, I think there should be a better way to write this code. I'm lucky that this data only has 23 rows so it was easy to replace values manually. However, if I have to work with a bigger data, replacing values manually might be a bad and unproductive way to deal with the data.

For my future web scraping projects, I would like to explore other website format that are different from the case studies we went through in class and try to clean those data. It might take longer time than writing this paper but I think it should be a good practice on dealing with data.