Introduction to Web Scraping with R

An Introductory Case Study

Simon Munzert | IPSDS

Data on the web

Data on the web

borough's mayors make up the council of mayors (*Rat der Bürgermeister*), which is led by the city's Governing Mayor and advises the Senate. The neighborhoods have no local government bodies.

Twin towns – sister cities [edit]

See also: List of twin towns and sister cities in Germany

Berlin maintains official partnerships with 17 cities.[100] Town twinning between Berlin and other cities began with its sister city Los Angeles in 1967. East Berlin's partnerships were canceled at the time of German reunification but later partially reestablished. West Berlin's partnerships had previously been restricted to the borough level. During the Cold War era, the partnerships had reflected the different power blocs, with West Berlin partnering with capitals in the Western World, and East Berlin mostly partnering with cities from the Warsaw Pact and its allies.

There are several joint projects with many other cities, such as Beirut, Belgrade, São Paulo, Copenhagen, Helsinki, Johannesburg, Mumbai, Oslo, Shanghai, Seoul, Sofia, Sydney, New York City and Vienna. Berlin participates in international city associations such as the Union of the Capitals of the European Union, Eurocities, Network of European Cities of Culture, Metropolis, Summit Conference of the World's Major Cities, and Conference of the World's Capital Cities. Berlin's official sister cities are:[100]

- 1967 Los Angeles, United States
- 1987 Paris, France
- 1988 Madrid, Spain
- 1989 Istanbul, Turkey
- 1991 Warsaw, Poland[101]
- 1991 Moscow, Russia
- 1992 Brussels, Belgium
- 1992 Budapest, Hungary[102]
- 1993 Tashkent, Uzbekistan
- 1993 Mexico City, Mexico
- 1993 Jakarta, Indonesia
- 1994 Beijing, China
- 1994 Tokyo, Japan
- 1994 Buenos Aires, Argentina
- 1995 Prague, Czech Republic[103]
- 2000 Windhoek, Namibia
- 2000 London, United Kingdom

Capital city [edit]

Berlin is the capital of the Federal Republic of Germany. The President of Germany, whose functions are mainly ceremonial under the German constitution, has his official residence in Schloss Bellevue.[104] Berlin is the seat of the German executive, housed in the Chancellery, the *Bundeskanzleramt*. Facing the Chancellery is the Bundestag, the German Parliament, housed in the renovated Reichstag building since the government relocated to Berlin in 1998. The Bundesrat ("federal council", performing the function of an upper house) is the representation of the Federal States (*Bundesländer*) of Germany and has its

Data on the web

borough's mayors make up the council of mayors (*Rat der Bürgermeister*), which is led by the city's Governing Mayor and advises the Senate. The neighborhoods have no local government bodies.

Twin towns – sister cities [edit]

See also: List of twin towns and sister cities in Germany

Berlin maintains official partnerships with 17 cities.[100] Town twinning between Berlin and other cities began with its sister city Los Angeles in 1967. East Berlin's partnerships were canceled at the time of German reunification but later partially reestablished. West Berlin's partnerships had previously been restricted to the borough level. During the Cold War era, the partnerships had reflected the different power blocs, with West Berlin partnering with capitals in the Western World, and East Berlin mostly partnering with cities from the Warsaw Pact and its allies.

There are several joint projects with many other cities, such as Beirut, Belgrade, São Paulo, Copenhagen, Helsinki, Johannesburg, Mumbai, Oslo, Shanghai, Seoul, Sofia, Sydney, New York City and Vienna. Berlin participates in international city associations such as the Union of the Capitals of the European Union, Eurocities, Network of European Cities of Culture, Metropolis, Summit Conference of the World's Major Cities, and Conference of the World's Capital Cities. Berlin's official sister cities are:[100]

- 1967 Los Angeles, United States
- 1987 Paris, France
- 1988 Madrid, Spain
- 1989 Istanbul, Turkey
- 1991 Warsaw, Poland[101]
- 1991 Moscow, Russia
- 1992 Brussels, Belgium
- 1992 Budapest, Hungary[102]
- 1993 Tashkent, Uzbekistan
- 1993 Mexico City, Mexico
- 1993 Jakarta, Indonesia
- 1994 Beijing, China
- 1994 Tokyo, Japan
- 1994 Buenos Aires, Argentina
- 1995 Prague, Czech Republic[103]
- 2000 Windhoek, Namibia
- 2000 London, United Kingdom

Capital city [edit]

Berlin is the capital of the Federal Republic of Germany. The President of Germany, whose functions are mainly ceremonial under the German constitution, has his official residence in Schloss Bellevue.[104] Berlin is the seat of the German executive, housed in the Chancellery, the *Bundeskanzleramt*. Facing the Chancellery is the Bundestag, the German Parliament, housed in the renovated Reichstag building since the government relocated to Berlin in 1998. The Bundesrat ("federal council", performing the function of an upper house) is the representation of the Federal States (*Bundesländer*) of Germany and has its

Let's grab these data!

Step 1: Load packages

R code

```
library(rvest)
library(stringr)
```

end

Let's grab these data!

Step 2: Parse page source

R code

```
library(rvest)
library(stringr)
parsed_url <- read_html("https://en.wikipedia.org/wiki/Berlin")
```

end

Let's grab these data!

Step 3: Extract information

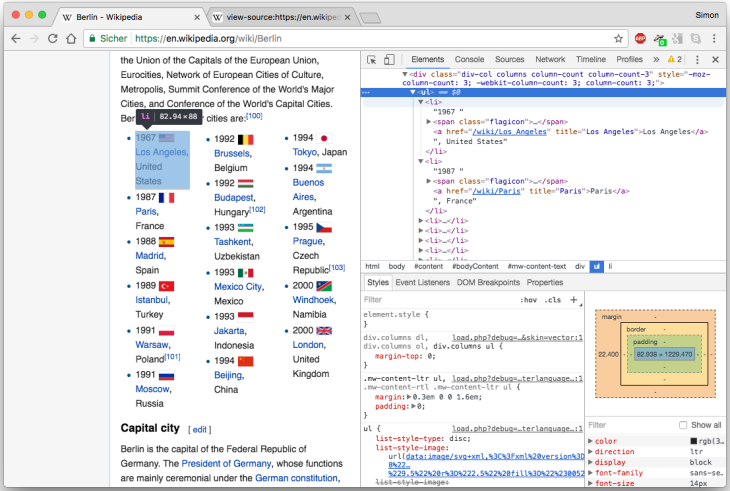
R code

```
library(rvest)
library(stringr)
parsed_url <- read_html("https://en.wikipedia.org/wiki/Berlin")
parsed_nodes <- html_nodes(parsed_url, xpath = "//div[contains(@class, 'column-count-3')]//li"
)
cities <- html_text(parsed_nodes)
cities[1:10]
```

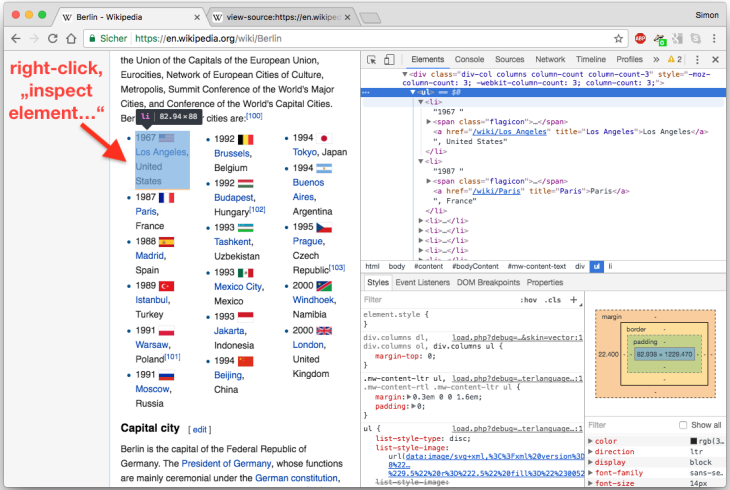
```
 [1] "1967  Los Angeles, United States" "1987  Paris, France"
 [3] "1988  Madrid, Spain"              "1989  Istanbul, Turkey"
 [5] "1991  Warsaw, Poland[103]"        "1991  Moscow, Russia"
 [7] "1992  Brussels, Belgium"          "1992  Budapest, Hungary[104]"
 [9] "1993  Tashkent, Uzbekistan"       "1993  Mexico City, Mexico"
```

end

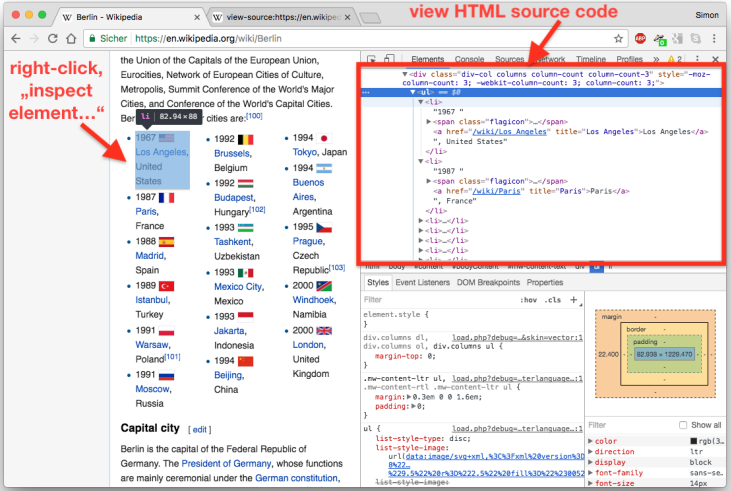
Why was this so easy?



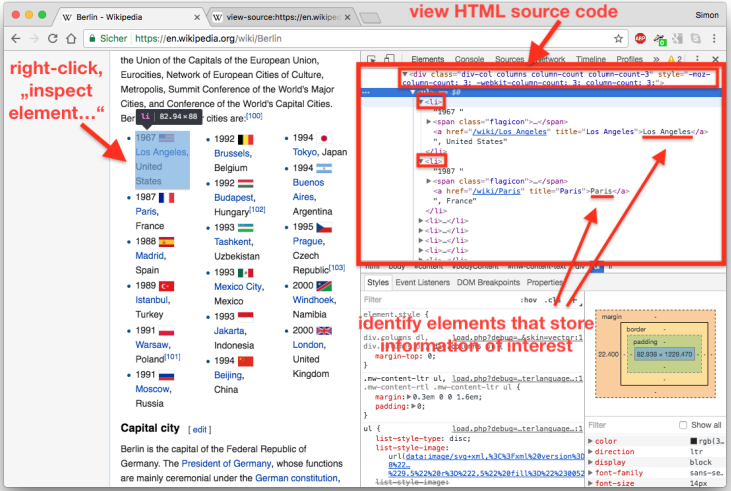
Why was this so easy?



Why was this so easy?



Why was this so easy?



Let's clean up these data!

R code

```
cities[1:10]
```

```
 [1] "1967  Los Angeles, United States" "1987  Paris, France"
 [3] "1988  Madrid, Spain"              "1989  Istanbul, Turkey"
 [5] "1991  Warsaw, Poland[103]"        "1991  Moscow, Russia"
 [7] "1992  Brussels, Belgium"          "1992  Budapest, Hungary[104]"
 [9] "1993  Tashkent, Uzbekistan"       "1993  Mexico City, Mexico"
```

end

Let's clean up these data!

R code

```
cities[1:10]
```

```
 [1] "1967  Los Angeles, United States" "1987  Paris, France"
 [3] "1988  Madrid, Spain"              "1989  Istanbul, Turkey"
 [5] "1991  Warsaw, Poland[103]"        "1991  Moscow, Russia"
 [7] "1992  Brussels, Belgium"          "1992  Budapest, Hungary[104]"
 [9] "1993  Tashkent, Uzbekistan"       "1993  Mexico City, Mexico"
```

end

Step 1: Remove footnotes with a regular expression

R code

```
cities <- str_replace(cities, "\\[\\d+\\]", "")
cities[1:10]
```

```
 [1] "1967  Los Angeles, United States" "1987  Paris, France"
 [3] "1988  Madrid, Spain"              "1989  Istanbul, Turkey"
 [5] "1991  Warsaw, Poland"             "1991  Moscow, Russia"
 [7] "1992  Brussels, Belgium"          "1992  Budapest, Hungary"
 [9] "1993  Tashkent, Uzbekistan"       "1993  Mexico City, Mexico"
```

end

Let's clean up these data!

Step 2: Extract data with regular expressions

R code

```
year <- str_extract(cities, "\\d{4}")
city <- str_extract(cities, "[[:alpha:] ]+") %>% str_trim
country <- str_extract(cities, "[[:alpha:] ]+$") %>% str_trim
year[1:10]
 [1] "1967" "1987" "1988" "1989" "1991" "1991" "1992" "1992" "1993" "1993"
city[1:10]
 [1] "Los Angeles" "Paris"       "Madrid"      "Istanbul"    "Warsaw"
 [6] "Moscow"      "Brussels"    "Budapest"    "Tashkent"    "Mexico City"
country[1:10]
 [1] "United States" "France"        "Spain"         "Turkey"
 [5] "Poland"        "Russia"        "Belgium"       "Hungary"
 [9] "Uzbekistan"    "Mexico"
```

end

Let's clean up these data!

Step 3: Put everything into data frame

R code

```
cities_df <- data.frame(year, city, country)
head(cities_df)
```

```
  year        city       country
1 1967 Los Angeles United States
2 1987       Paris        France
3 1988      Madrid         Spain
4 1989    Istanbul        Turkey
5 1991      Warsaw        Poland
6 1991      Moscow        Russia
```

end

Let's map these data!

Step 1: Load necessary packages

R code

```
library(ggmap)
library(maps)
```

end

Let's map these data!

Step 2: Geocode cities with the Google Maps API

R code

```
library(ggmap)
library(maps)
cities_coords <- geocode(paste0(cities_df$city, ", ", cities_df$country))
cities_df$lon <- cities_coords$lon
cities_df$lat <- cities_coords$lat
cities_df$lon[1:10]
 [1] -118.243685    2.352222   -3.703790   28.978359          NA
 [6]   37.617300          NA   19.040235          NA  -99.133208
cities_df$lat[1:10]
 [1] 34.05223 48.85661 40.41678 41.00824       NA 55.75583       NA
 [8] 47.49791       NA 19.43261
```

end

Let's map these data!

Step 3: Plot world map, add coordinates

R code

```
map_world <- borders("world", colour = "gray50", fill = "white")
ggplot() + map_world + geom_point(aes(x = cities_df$lon, y = cities_df$lat), color = "red",
size = 1) + theme_void()
```

end

Let's map these data!

Step 3: Plot world map, add coordinates

R code

```
map_world <- borders("world", colour = "gray50", fill = "white")
ggplot() + map_world + geom_point(aes(x = cities_df$lon, y = cities_df$lat), color = "red",
size = 1) + theme_void()
```

end