Introduction aux mégadonnées en sciences sociales (FAS 1001)

Cours 8: Web scraping

Nadjim Fréchet

12 mars 2024

Retour sur le TP3 😮

- Pas encore terminé de corriger. Vous aurez les notes ce soir ou demain.
- Pour les travaux de sessions. Je vais tout faire pour terminer les corrections pour le début de la semaine prochaine.
- Pour le TP4. Je vous donne une semaine de plus! 🤴

ATTENTION

RAPPEL DE LA DATE DE REMISE: lundi 25 mars avant minuit. Et oui, une semaine de plus.

Invité de la semaine 😇

- Benjamin Guinaudeau
- As du codage. Spécialisation en analyses textuelles automatisées, Machine learning et data science.
- Article avec des données de TikTok: APSA ITP Best Journal Article Award
- MAINTENANT PROFESSEUR À L'UNIVERSITÉ LAVAL À QUÉBEC!

Lecture de la semaine 📽 🔞



- Chapitre 7 du <u>livre</u> de Alexander (2023).
- Introduction aux APIs. C'est quoi un API (ou Application Programming Interface) selon vous?
- Certains dans cette classe ont utilisé un API (sous forme d'un fichier JSON).
- Accès (ou appel) à des données (entreprises, groupes, particuliers, etc.) disponibles sous certaines conditions (clairement explicitées).
- Souvent accessible avec des packages (R, Python ou autres) développés par des programmeurs ou les propriétaires des données.

Exemple d'API 🖊 TheyWorkForYou

- Site pour chercher des données parlementaires britanniques.
- Exemple de clé
- Package de disponible pour télécharger les données

Exemple d'API 🖊 TheyWorkForYou

```
1 # Load package
2
3 library(twfy)
4 library(tidyverse)
5
6 # Set API key
7
8 my_key <- "key"
9
10 set_api_key(my_key)
11
12 # Chercher les données de David Cameron
13
14 # getDebates(type ="commons", person = 10777)</pre>
```

C'est bien quand c'est accessible...

- Qu'est-ce qu'on fait quand les données ne sont pas accessibles?
- Web scraping direct?
- Discussion sur la légalité du scraping. Est-ce une pratique sans limite?

Web scraping

- Plusieurs packages de disponibles dans différentes langues de programmation. Exemple BeautifulSoup dans Python.
- Dans R pour le scraping de page statique, le package <u>rvest</u> est un excellent outil de travail.
- Pour *scraper* des données tirées du web, un peu d'exploration de page s'impose. On peut le faire de différentes manières, notamment avec <u>SelectorGadget</u>.
- SelectorGadget permet de chercher des éléments HTML de la page web. D'autres manières possibles, notamment de regarder directement le code source de la page.

Exemple avec une page Wikipedia

- Pratiquez-vous sur une page *Wikipédia*! Les éléments d'une page *Wikipédia* restent plutôt stables à travers le temps.
- Exemple avec la page Wikipédia suivante.

```
1 #### Scraping de données de sondages canadiens ####
2
3 # 1 - Libraries ----
4
5 library(tidyverse)
6 library(lubridate) # Package pour gérer des données temporelles
7 library(rvest) # Package pour le scraping
8
9 # 2 - Scraping ----
10
11 Polls_data <- read_html("https://en.wikipedia.org/wiki/Opinion_poll</pre>
```

Exemple avec une page Wikipedia

• On cherche les éléments de la page qui seront nécessaires pour notre travail.

```
1 Polls_data_1 <- Polls_data |>
2  ### Chercher l'élément "table" ###
3  html_elements("table") |>
4  ### Chercher le deuxième élément (ou table) de la liste ###
5  pluck(2) |>
6  ### Transformation en matrice de données ###
7  html_table(fill = T)
```

Importance de nettoyer ses données

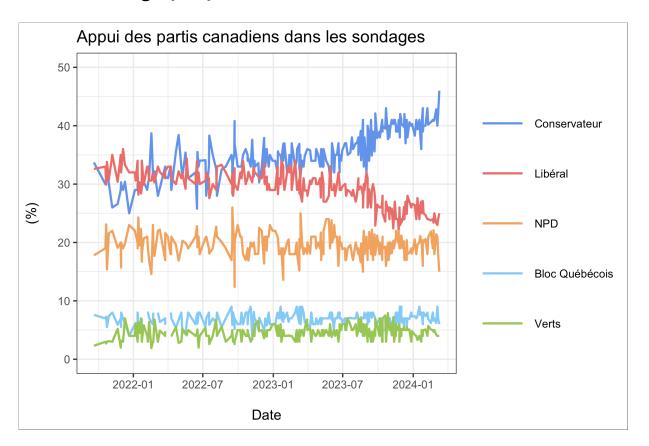
```
1 Polls_data_f <- Polls_data_1 |>
     ### renommer les variables pertinentes ###
3
     rename(firm
                             = "Polling firm",
 4
             date
                             = "Last dateof polling[a]",
 5
             pred_cpc
                             = CPC,
             pred_lpc
 6
                             = LPC,
 7
             pred_ndp
                             = NDP,
8
             pred_bq
                             = BQ
9
                             = PPC,
             pred_ppc
             pred gpc
                             = GPC,
10
             error_margin = "Marginof error[c]",
11
             sample_size = "Samplesize[d]",
poll_method = "Polling method[e]",
12
13
             leading_margin = Lead) |>
14
15
     ### Nettoyage des variables ###
     filter(firm != "") |>
16
     # Enlever des rangées inutiles #
17
18
     slice(-c(1,235,236)) |>
19
     # Retirer les variable de trop #
```

Importance de nettoyer ses données

```
1 # Visualisation de nos données
2
3 glimpse(Polls_data_f)
```

```
Rows: 233
    Columns: 12
                                 <chr> "Mainstreet Research", "Abacus Data", "Angus
    $ firm
    Reid", "Na...
                                 <date> 2024-03-09, 2024-03-06, 2024-03-04, 2024-03-
    $ date
    01, 2024-0...
                                <dbl> 46.0, 42.0, 40.0, 42.8, 41.0, 40.8, 41.0,
    $ pred cpc
    40.6, 40.2, 4...
    $ pred_lpc
                                <dbl> 25.0, 24.0, 23.0, 23.3, 25.0, 23.4, 24.0,
    23.8, 24.0, 2...
                                <dbl> 15.0, 18.0, 21.0, 21.4, 18.0, 22.1, 19.0,
    $ pred_ndp
    21.9, 20.9, 1...
                                <dbl> 6.0, 7.0, 9.0, 6.2, 7.0, 6.6, 8.0, 6.2, 6.9,
    $ pred_bq
                       Session Build Debug Profile Tools Window Help
                                                                           • ♣ ♣ ® ⑤ 🖵 🕏 • • • Tue Mar 12 3:26 AM
🛂 🕶 🕷 🚭 🕶 🔜 📑 🛮 📥 🖟 Go to file/function 🔠 🖺 🕶 Addins 🕶
                                                                                                      split_french_debate_texts
←⇒ 📠 🔚 🔲 Source on Save 🔍 🎢 🔻 📗
                                                                            R 4.3.2 ~/Library/CloudStorage/Dropbox/UdeM/Fas-1001/_data/spli
                                               Run Source -
                                                                           Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publication
     ibrary(pdftools)
                                                                           Type 'demo()' for some demos, 'help()' for on-line help, o
      brary(tidyverse)
     library(lubridate)
                                                                            'help.start()' for an HTML browser interface to help.
     # Texte du débat Macron VS Lepen
                                                                           Files Plots Packages Help Viewer Presentation
     text_raw <- data.frame(text = paste0(suppressMessages(pdf_text("_data/macron_vs_lepen.pdf</pre>
      Dropbox > UdeM > Fas-1001 > _data > split_french_debate_tex ...
            text = str_replace_all(text, c("Emmanuel Macron :
                                    "Marine Le Pen :" = "Mme Le Pen :",
"Christophe Jakubyszyn :" = "M. Jakubyszyn :",
                                                                            data __data
53:36 📕 1.3 - Fusion des deux variables clés et création de la base de données finale
                                                                   R Script
                                                                            ■ Mail Rhistory
                                                                                                       23.5 KB
                                                                                                               Feb 17, 2024
Environment History Connections Tutorial
                                                                            code_split_text.R
                                                                                                      2.3 KB
                                                                                                               Feb 14, 2024,
                                                                ≣ List - | C -
📹 🔚 🔛 Import Dataset 🗸 🔇 146 MiB 🗸 🧳
                                                                            split_french_debate_texts.Rproj 205 B
                                                                                                               Mar 12, 2024,
R - Global Environment -
                              Environment is empty
```

Visualisation graphique



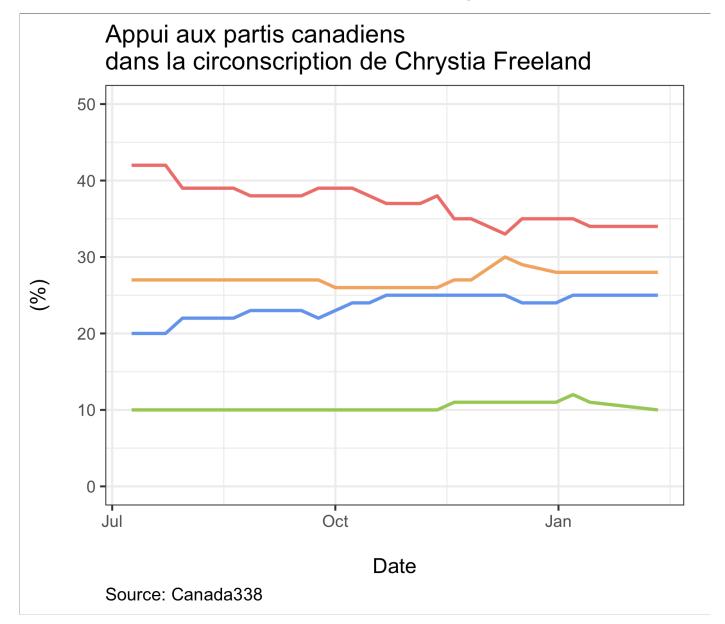
Scraping plus complexe

Autres scraping possible de pages PDF

- Voir le <u>chapitre 7</u> de Alexander (<u>2023</u>) pour différents exemples.
- Travail mené par certaines d'entre vous.

Scraping de pages PDF

Possibilité d'automatiser le processus de scraping 📈



Bibliographie

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in R.* Chapman; Hall/CRC.

Error	×