UE Visualisation

2019-2020

Dr. Maxime Wack

AHU Informatique médicale

Hôpital Européen Georges Pompidou,
Université de Paris

Web scraping

Utilisation de `httr` et `rvest`

httr

Permet de faire des requêtes réseau

→ interroger et télécharger directement depuis R

rvest

Extraction de données depuis des pages HTML

httr

Télécharger une page wikipedia

```
GET("https://en.wikipedia.org/wiki/Comparison_of_operating_systems")
```

```
## Response [https://en.wikipedia.org/wiki/Comparison_of_operating_systems]
##   Date: 2019-11-21 22:05
##   Status: 200
##   Content-Type: text/html; charset=UTF-8
##   Size: 241 kB
## <!DOCTYPE html>
## <html class="client-nojs" lang="en" dir="ltr">
## <head>
## <meta charset="UTF-8"/>
## <title>Comparison of operating systems - Wikipedia</title>
## <script>document.documentElement.className="client-js";RLCONF={"wgBreakFra
## "Articles with unsourced statements from May 2018","Articles with unsource
## "wgNoticeProject":"wikipedia","wgWikibaseItemId":"Q3345986","wgCentralAuth
## "ext.gadget.watchlist-notice","ext.gadget.DRN-wizard","ext.gadget.charinse
## <script>(RLQ=window.RLQ||[]).push(function(){mw.loader.implement("user.tok
## ...
```

Parsing HTML

```
wiki %>%
  read_html -> wiki_html
```

```
## {html_document}
## <html class="client-nojs" lang="en" dir="ltr">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UT
## [2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subje
```

Sélecteurs CSS

W3Schools

Selecteurs permettant d'identifier un **nœud** précis dans le **DOM** (Document Object Model) d'une page HTML

Permet de sélectionner par identifiant, classe, position dans la hiérarchie, position entre élements d'un même niveau, ou relativement entre élements

Utiliser l'**inspecteur** des outils de développement du navigateur pour identifier les éléments à capturer

Sélecteurs CSS

```
wiki_html %>%
  html_nodes(".wikitable")
```

```
## {xml_nodeset (4)}
## [1] <table class="wikitable sortable" style="font-size: smaller; text-alig
## [2] <table class="wikitable sortable" style="font-size: smaller; text-alig
## [3] <table class="wikitable" style="font-size: smaller; text-align: center
## [4] <table class="wikitable" style="font-size: smaller; text-align: center
```

```
wiki_html %>%
  html_node("div + .wikitable")
```

```
## {html_node}
## <table class="wikitable sortable" style="font-size: smaller; text-align: c
## [1] <tbody>\n<tr>\n<th>Name\n</th>\n<th>Creator\n</th>\n<th abbr="Initial
```

Extraction d'une table

```
wiki_html %>%
  html_node("div + .wikitable") %>%
  html_table -> wikitable
```

Name	Creator	Initial public release	Predecessor	Current stable version	Rel d
AIX	IBM	1986	UNIX System V Release 3	7.2	2015 Octo
Android	Android, Inc., Google	2008	None	10	2019 Sept 3

Exercices

Transformer cette table en forme normale

Extraire la table avec les informations techniques

Identifier les OS libres fonctionnant avec un microkernel