Scrapping

Ryan Flake

11/10/2020

Scrapping Practice

```
hot100page <- "https://www.billboard.com/charts/hot-100"
hot100 <- read_html(hot100page)

hot100
```

```
## {html_document}
## <html class="" lang="">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body class="chart-page chart-page-" data-trackcategory="Charts-TheHot100 ...
```

```
str(hot100)
```

```
## List of 2
##  $ node:<externalptr>
##  $ doc :<externalptr>
##  - attr(*, "class")= chr [1:2] "xml_document" "xml_node"
```

```
body_nodes <- hot100 %>%
  html_node("body") %>%
  html_children()
body_nodes
```

```
## {xml_nodeset (35)}
##  [1] <div class="header-wrapper ">\n<header id="site-header" class="site-head ...
##  [2] <div class="site-header__placeholder"></div>
##  [3] <script>\n        var PGM = window.PGM || {};\n        PGM.config = PGM. ...
##  [4] <main id="main" class="page-content"><div id="charts" data-page-title="T ...
##  [5] <div class="ad_desktop dfp-ad dfp-ad-promo " data-position="promo" data- ...
##  [6] <div class="ad-container footerboard footerboard--bottom">\n    <div cla ...
##  [7] <footer id="site-footer" class="site-footer"><div class="container foote ...
##  [8] <div class="biz-modal">\n    <div class="biz-modal__content">\n        < ...
##  [9] <script>\n    window.CLARITY = window.CLARITY || [];\n</script>
## [10] <div class="ad_clarity" data-out-of-page="true" style="display: none;"><  ...
## [11] <script>\n    var darkMatterCMD = function() {\n        this.darkMatterC ...
## [12] <script src="https://www.billboard.com/assets/1603122088/js/vendors_/art ...
## [13] <script src="https://www.billboard.com/assets/1603122088/js/vendors_/clo ...
## [14] <script src="https://www.billboard.com/assets/1603122088/js/vendors_/rea ...
## [15] <script src="https://www.billboard.com/assets/1603122088/js/vendors_/rea ...
## [16] <script src="https://www.billboard.com/assets/1603122088/js/vendors_/rea ...
## [17] <script src="https://www.billboard.com/assets/1603122088/js/vendors_/rea ...
## [18] <script src="https://www.billboard.com/assets/1603122088/js/default_/art ...
## [19] <script src="https://www.billboard.com/assets/1603122088/js/default_/rea ...
```

```
## [20] <script src="https://www.billboard.com/assets/1603122088/js//react-chart ...
## ...
```

```
body_nodes %>%
  html_children()
```

```
## {xml_nodeset (9)}
## [1] <header id="site-header" class="site-header " role="banner"><div class="s ...
## [2] <div class="header-wrapper__secondary-header">\n<nav class="site-header-l ...
## [3] <div id="charts" data-page-title="THE HOT 100" data-chart-code="HSI" data ...
## [4] <div class="footerboard-wrapper">\n        <div class="ad_desktop_placeho ...
## [5] <div class="container footer-content">\n\t\t\t\t\t\t<div class="cover-image ...
## [6] <div class="container">\n\t\t<p class="copyright__paragraph">© 2020 Billb ...
## [7] <div class="container">\n\t\t<p class="station-identification">\n\t\t\tBi ...
## [8] <div class="container">\n\t\t\n\n\n    <div class="ad_desktop dfp-ad dfp- ...
## [9] <div class="biz-modal__content">\n        <button class="biz-modal__close ...
```

```
rank <- hot100 %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all("//span[contains(@class, 'chart-element__rank__number')]") %>%
  rvest::html_text()

artist <- hot100 %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all("//span[contains(@class, 'chart-element__information__artist')]") %>%
  rvest::html_text()

title <- hot100 %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all("//span[contains(@class, 'chart-element__information__song')]") %>%
  rvest::html_text()

chart_df <- data.frame(rank, artist, title)
knitr::kable(
  chart_df %>% head(10)
)
```

rank	artist	title
1	24kGoldn Featuring iann dior	Mood
2	Ariana Grande	Positions
3	Drake Featuring Lil Durk	Laugh Now Cry Later
4	The Weeknd	Blinding Lights
5	Gabby Barrett Featuring Charlie Puth	I Hope
6	Jawsh 685 x Jason Derulo	Savage Love (Laxed - Siren Beat)
7	Internet Money & Gunna Featuring Don Toliver & NAV	Lemonade
8	Ariana Grande	34+35
9	Bad Bunny & Jhay Cortez	Dakiti
10	Cardi B Featuring Megan Thee Stallion	WAP

Real Scrapping

```
heritage <- "https://www.heritage.org/voterfraud-print/search"

heritage_fraud <- read_html(heritage)
```

```
heritage_fraud
```

```
## {html_document}
## <html lang="en" dir="ltr" prefix="content: http://purl.org/rss/1.0/modules/content/  dc: http://purl
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body class="path-voterfraud-print">\n    <!-- Google Tag Manager (noscri ...
```

```
str(heritage_fraud)
```

```
## List of 2
##  $ node:<externalptr>
##  $ doc :<externalptr>
##  - attr(*, "class")= chr [1:2] "xml_document" "xml_node"
```

```
State <- heritage_fraud %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all(
    "//span[
      contains(
    @class, 'views-field views-field-field-fraud-state-administrative-area'
      )
    ]"
  ) %>%
  rvest::html_text()

Year <- heritage_fraud %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all(
    "//span[
      contains(
        @class, 'views-field views-field-field-year-of-disposition'
      )
    ]"
  ) %>%
  rvest::html_text()

Name <- heritage_fraud %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all(
    "//span[
      contains(
        @class, 'views-field views-field-name'
      )
    ]"
  ) %>%
  rvest::html_text()


Type_of_Case <- heritage_fraud %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all(
    "//span[
      contains(
        @class, 'views-field views-field-field-case-type'
```

```
      )
    ]"
  ) %>%
  rvest::html_text()

Type_of_Fraud <- heritage_fraud %>%
  rvest::html_nodes('body') %>%
  xml2::xml_find_all("//span[contains(@class, 'views-field views-field-field-fraud-type')]") %>%
  rvest::html_text()

heritage_chart_df <- data.frame(State, Year, Name, Type_of_Case, Type_of_Fraud)

knitr::kable(
  heritage_chart_df %>% head(10)
)
```

State	Year	Name	Type_of_Case	Type_of_Fraud
State	Year	Name	Type of Case	Type of Fraud
StateArizona	Year2020	NameRandy Allen Jumper	Case TypeCriminal Conviction	Fraud TypeFraudulent Use Of Absentee Ballots, Duplicate Voting
StateCalifornia	Year2020	NameApril Atilano	Case TypeCriminal Conviction	Fraud TypeFalse Registrations
StateCalifornia	Year2020	NameJentry Jasperson	Case TypeCriminal Conviction	Fraud TypeBallot Petition Fraud
StateCalifornia	Year2020	NameBradley Jasperson	Case TypeCriminal Conviction	Fraud TypeBallot Petition Fraud
StateCalifornia	Year2020	NameNorman Hall	Case TypeCriminal Conviction	Fraud TypeBallot Petition Fraud
StateCalifornia	Year2020	NameRichard Howard	Case TypeCriminal Conviction	Fraud TypeBuying Votes, Ballot Petition Fraud
StateCalifornia	Year2020	NameLouis Wise	Case TypeCriminal Conviction	Fraud TypeBuying Votes, Ballot Petition Fraud
StateCalifornia	Year2020	NameChristopher Williams	Case TypeCriminal Conviction	Fraud TypeBuying Votes, Ballot Petition Fraud
StateCalifornia	Year2020	NameNickey Huntley	Case TypeCriminal Conviction	Fraud TypeBuying Votes, Ballot Petition Fraud

Data Tidying

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------- tidyvers
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.3.1
## -- Conflicts ---------------------------------------------------------------------- tidyverse_con
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()
```

```
## x XML::xml()              masks rvest::xml()
```

```
heritage_chart_df_2 <-heritage_chart_df %>%
  slice(2:959)
```

```
heritage_chart_df_3 <- heritage_chart_df_2 %>%
  separate(
    col = State,
    into = combine("Trash", "State"),
    sep = 5,
    convert = TRUE
  ) %>%
  separate(
    col = Year,
    into = combine("Trash2", "Year"),
    sep = 4,
    convert = TRUE
  ) %>%
  separate(
    col = Name,
    into = combine("Trash3", "Name"),
    sep = 4,
    convert = TRUE
  ) %>%
  separate(
    col = Type_of_Case,
    into = combine("Tash4", "Type_of_Case"),
    sep = 9,
    convert = TRUE
  ) %>%
  separate(
    col = Type_of_Fraud,
    into = combine("Trash5", "Type_of_Fraud"),
    sep = 10,
    convert = TRUE
  )
```

```
## Warning: `combine()` is deprecated as of dplyr 1.0.0.
## Please use `vctrs::vec_c()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
Voter_Fraud_Data_Clean <- heritage_chart_df_3 %>%
  select(State, Year, Name, Type_of_Case, Type_of_Fraud)
write_csv(Voter_Fraud_Data_Clean, "Voter_Fraud_Data.csv")
```