

Ciencia de Datos para Políticas Públicas

Clase 09 - Web scraping y otros

Pablo Aguirre Hormann

07/10/2020

¿Qué veremos hoy?

- Introducción a Web scraping
- Ejemplos de *Machine Learning* para Políticas Públicas

Web scraping

¿Qué es?

Objetivo: obtener información de la web

Distintas formas:

- Click y descargar
- API: *Application Programming Interface*
 - Interactuar directamente con la API
 - Usar paquetes que facilitan su uso desde R
- **Web scraping**

¿Cómo?

The image shows a web browser displaying the 'Observatorio Social' website. The header includes the logo of the 'Ministerio de Desarrollo Social y Familia' and 'Gobierno de Chile'. A navigation bar contains links: 'Inicio', 'Encuesta CASEN', 'Encuestas MDS', 'Compromisos Internacionales', and 'Indicadores Territoriales'. Below the navigation bar, the page title is 'Porcentaje de personas en situación de pobreza'. A browser developer tools window is open, showing the 'Elements' tab. The HTML structure is visible, including the head and body tags. The body contains a div with the id 'content'. The 'Styles' panel on the right shows the default Bootstrap box-sizing styles.

Ministerio de Desarrollo Social y Familia
Gobierno de Chile

Observatorio Social

Inicio Encuesta CASEN Encuestas MDS Compromisos Internacionales Indicadores Territoriales

Inicio / Encuesta CASEN »

Porcentaje de personas en situación de pobreza

```
<!DOCTYPE html>
<html>
  <head>
    <title></title>
    <meta http-equiv="content-type" content="text/html; charset=utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1, minimum-scale=1, maximum-scale=1, user-scalable=0">
    <script src="https://apis.google.com/_/scs/apps-static/_/js/k=oz.gapi.es.cvqdi.d=1/ed=1/am=wQE/rs=AGLTcCO9z1lRdXskTlVlVRw0Y09UcAjMxA/cb=gapi.loaded_0" nonce async></script>
    <script nonce="https://ssl.gstatic.com/accounts/o/2231879498-postmessagerelay.js"></script>
  </head>
  <body data-new-gr-c-s-loaded="true">...</body>
</html>
```

html:10-is body:page div#content

Styles Computed Event Listeners »

Filter :hov .cls +

element.style { }

#content { margin: 0 auto 25px; main.css:1517 }

* { -webkit-box-sizing: border-box; -moz-box-sizing: border-box; box-sizing: border-box; bootstrap.min.css:1019 }

Un poco de ayuda

```
vignette("selectorgadget")
```



chrome web store

[Inicio](#) > [Extensiones](#) > SelectorGadget



SelectorGadget

Ofrecido por: selectorgadget.com

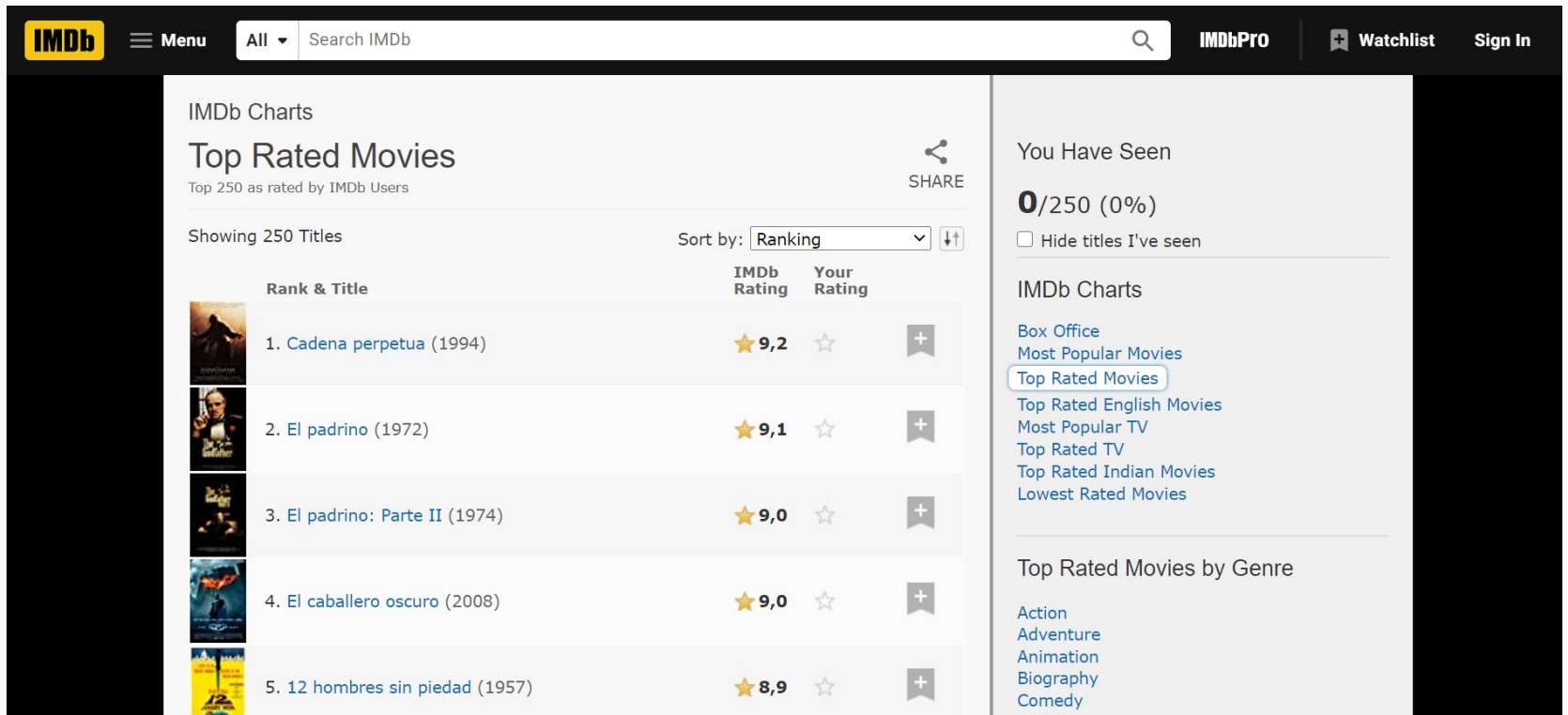
★★★★★ 84

[Herramientas para desarrolladores](#)

 100.000+ usuarios

Demo - Web scraping

- `web_scraping.R`
- `rvest`
- Ejemplo no "muy políticas públicas" pero bueno para demostrar el punto
 - **IMDb**: *Internet Movie Database*



The screenshot shows the IMDb website's 'Top Rated Movies' page. The header includes the IMDb logo, a menu, a search bar, and links to IMDbPro, Watchlist, and Sign In. The main content area is titled 'IMDb Charts' and 'Top Rated Movies', with a subtitle 'Top 250 as rated by IMDb Users'. It shows a list of 250 titles, sorted by Ranking. The first five titles are:

Rank & Title	IMDb Rating	Your Rating
1. Cadena perpetua (1994)	★ 9,2	☆
2. El padrino (1972)	★ 9,1	☆
3. El padrino: Parte II (1974)	★ 9,0	☆
4. El caballero oscuro (2008)	★ 9,0	☆
5. 12 hombres sin piedad (1957)	★ 8,9	☆

The right sidebar contains a 'You Have Seen' section showing 0/250 (0%) and a 'Hide titles I've seen' checkbox. Below this is the 'IMDb Charts' section with links to Box Office, Most Popular Movies, Top Rated Movies (highlighted), Top Rated English Movies, Most Popular TV, Top Rated TV, Top Rated Indian Movies, and Lowest Rated Movies. At the bottom is the 'Top Rated Movies by Genre' section with links to Action, Adventure, Animation, Biography, Comedy, and Crime.

Tengan en cuenta...

- El ejemplo que vimos lo hace ver bastante fácil. Muchos casos lo son
- Pero... algunas *web* son más difíciles de "scrapear" que otras
- Existen otros paquetes para cosas más complejas
 - `RSelenium`

Casos de ML en Políticas Públicas

Educación

Objetivo: Incrementar tasas de graduación del colegio

- 1 de cada 5 estudiantes de USA no se gradúa de *high-school* a tiempo
- Existen programas de intervención para ayudar a los estudiantes que estén en riesgo de no graduarse a tiempo
- **¿Cómo identificar a estos estudiantes? ¿Cómo diferenciar cuáles tienen mayor o menor riesgo?**
- Distritos escolares de *North Carolina*, *Virginia*, y *Washington* han desarrollado modelos predictivos con el fin de estimar el riesgo de que un alumno no se gradue a tiempo

Educación (cont)

$$NoGraduadoATiempo = \hat{f}(X)$$

$$NoGraduadoATiempo = \{1, 0\}$$

$$X = \{notas, NSE, N_detenciones, ActExtPrg, \dots\}$$

- Se puede ocupar información del pasado (por ejemplo, 2010 a 2019) para entrenar y testear distintos modelos
- Con el mejor modelo se pueden tomar las características de los alumnos actuales (X 's) y predecir $NoGraduadoATiempo(Y)$
- Como la variable a predecir, $NoGraduadoATiempo$, es binaria (1/0), lo que se obtiene como resultado es un valor de "riesgo (probabilidad) de no graduarse a tiempo" (entre 0 y 1).
- El valor de "riesgo de no graduarse a tiempo" sirve como un indicador con el cual rankear de mayor a menor los estudiantes a los cuales apoyar.

Medición de pobreza

Objetivo: tener indicadores de pobreza actualizados con mayor frecuencia

- Encuestas dan información muy valiosa pero son costosas
- Lo anterior significa que se realizan cada cierto número de años (Ej: CASEN cada 2 años)
- **¿Existe una forma de tener información actualizada con mayor frecuencia?**

Medición de pobreza

Idea: buscar variables con alto poder predictivo de pobreza pero que se puedan obtener con mayor frecuencia

- Existen distintos estudios que vinculan la intensidad lumínica nocturna con actividad económica (*PIB*) y recientemente también con posesión de activos y pobreza

$$Pobreza = \hat{f}(IntensidadLuminicaNocturna, \dots)$$

- Intensidad lumínica es una variable que se puede obtener a través de imágenes satelitales ("mapa") y que se puede discretizar ocupando, por ejemplo, los límites comunales de Chile.
- Ejemplo de aplicación: tomar la información histórica de CASEN y entrenar un modelo predictivo de pobreza comunal usando intensidad lumínica (y otras variables) como predictor
- Trabajo necesario: obtener la información de intensidad lumínica a nivel comunal (imágenes satelitales)
- Teniendo un modelo podemos luego obtener información de intensidad lumínica, por ejemplo, cada 6 meses para estimar pobreza de forma más frecuente

Medio ambiente

Objetivo: disminuir el potencial impacto ambiental o de la salud de las personas a través de un plan de fiscalizaciones efectivo

- En USA existen 300.000 establecimientos regulados por la *Clean Water Act* y que, por ende, la **EPA** (*Environmental Protection Agency*) debe fiscalizar
- Debido a las capacidades institucionales, se fiscalizan aproximadamente 28.000 establecimiento cada año (>10%) y durante 2012 y 2016 solo el 25% fue inspeccionado al menos una vez
- Las inspecciones son caras por lo que es necesario que cada una de estas sea ocupada de la forma más eficiente posible
- **¿Cómo escoger una lista de establecimientos a fiscalizar que maximice el uso de los recursos de la institución?**
- Hace algunos años la **EPA** ha estado testeando modelos predictivos para calcular la probabilidad de que una fiscalización encuentre un incumplimiento

Medio ambiente (cont)

Idea:

$$\text{Incumplimiento} = \hat{f}(\text{Características Establecimiento}, \text{Ubicación}, \dots)$$

- En la Superintendencia del Medio Ambiente estamos trabajando en algo similar a esto
- También estamos trabajando en un modelo predictivo de calidad del aire para la zona de Concón, Quintero, y Puchuncaví

Comentarios

- Muchos problemas de políticas públicas no son de predicción
 - Importante dedicarle tiempo a pensar si estas herramientas son o no aplicables

_ Cuando se identifica un problema de predicción se tiene la ventaja, en general, de tener muchos datos administrativos disponibles

- Queda mucho por avanzar todavía. Primer paso, tener datos disponibles de forma simple.

Próxima clase

- Última clase
- Presentaciones
 - Máximo 8 minutos de presentación
 - Máximo 4 minutos de preguntas
 - 15 minutos de break
- Informe final: sábado 17 de octubre 23:59