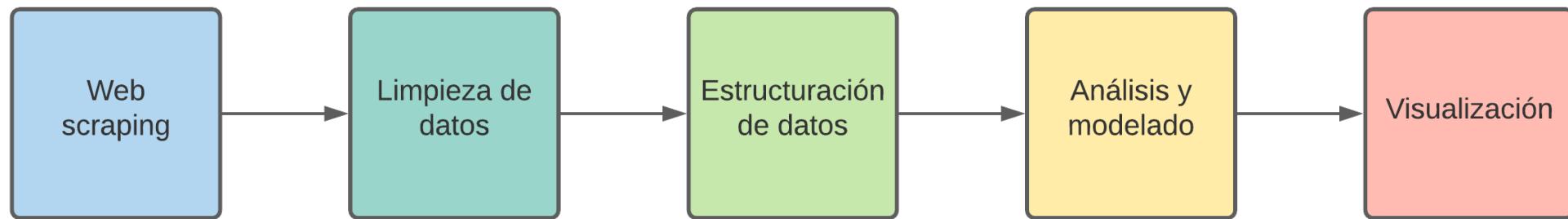


Ejemplos prácticos de web scraping y limpieza de datos

Javier Mtz

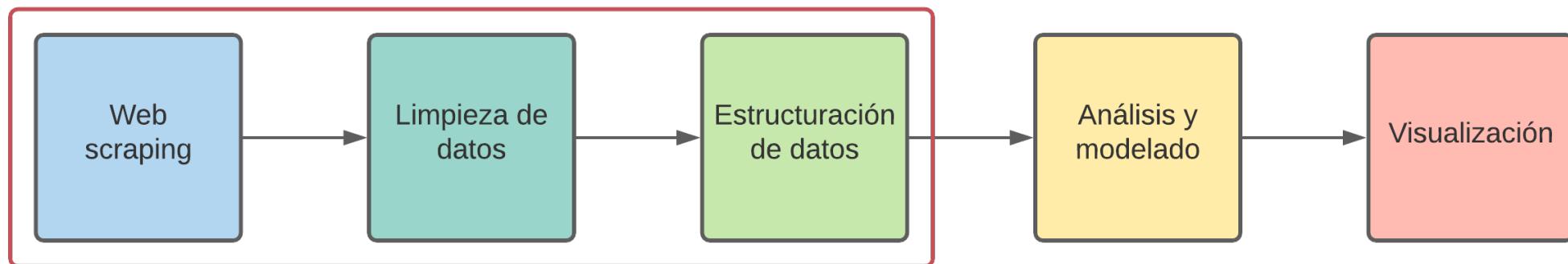
¿Qué vamos a ver?

Proceso de análisis



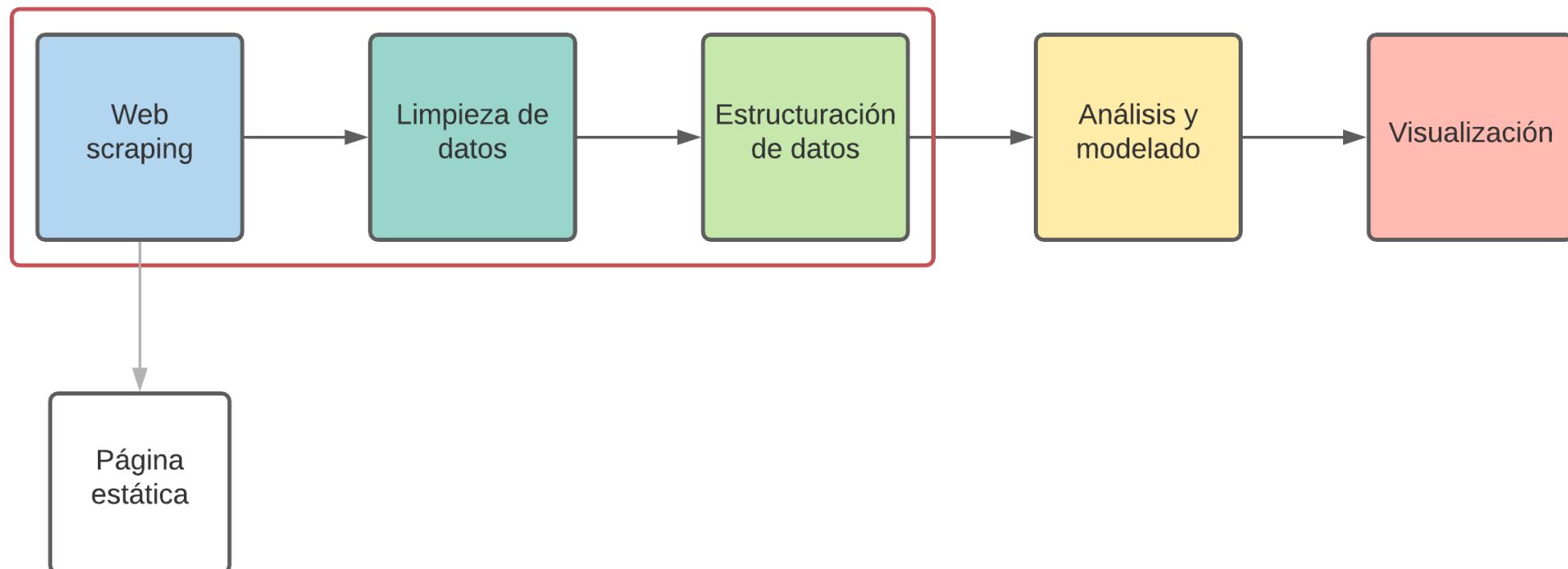
¿Qué vamos a ver?

Proceso de análisis



¿Qué vamos a ver?

Proceso de análisis



Análisis de las mañaneras

Análisis de las mañaneras

Objetivo:

- Analizar el discurso del Presidente frente a diferentes temas.



Análisis de las mañaneras

Identificación de fuentes de información

- lopezobrador.org.mx

AMLO CORONAVIRUS COVID-19 SALA DE PRENSA CONFERENCIA EN VIVO GABINETE BIOGRAFÍA

Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador
agosto 16, 2021
Buenos días. Animo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo Sheffield va a informarnos; luego, veremos los videos, como todos los lunes, de [...]

Versión estenográfica. Agua Saludable para La Laguna, desde Lerdo, Durango
agosto 15, 2021
Amigas, amigos de La Laguna, productores, representantes de los sectores productivos de La Laguna, ambientalistas, autoridades municipales, ciudadanos gobernadores: Nos da mucho [...]

Versión estenográfica. Visita a la región de la Presa El Zapotillo, en Cañas de Obregón, Jalisco
agosto 14, 2021
Amigas, amigos de Jalisco: Nos da mucho gusto estar aquí en Zapotillo, donde se construyó una presa que, aun cuando todavía no se termina, ya ha significado una inversión de alrededor [...]

- gob.mx/presidencia

GOBIERNO DE MÉXICO Registro para vacunación Información sobre COVID-19 Trámites Gobierno

Comunicados ¿Qué hacemos? Directorio Transparencia

lunes, 16 de agosto de 2021
Versión estenográfica. Conferencia de prensa del presidente Andrés Manuel López Obrador del 16 de agosto de 2021
continuar leyendo

domingo, 15 de agosto de 2021
Versión estenográfica. Agua Saludable para La Laguna
continuar leyendo

sábado, 14 de agosto de 2021
Versión estenográfica. Visita a la región de la Presa 'El Zapotillo'
continuar leyendo

MÉDIO REVISTA DE ENTRADA DE LA COMANDANCIA DEL EJÉRCITO MEXICANO

Análisis de las mañaneras

¿Qué queremos extraer de la página?



Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador

· agosto 16, 2021

Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de [...]



Versión estenográfica. Agua Saludable para La Laguna, desde Lerdo, Durango

· agosto 15, 2021

Amigas, amigos de La Laguna, productores, representantes de los sectores productivos de La Laguna, ambientalistas, autoridades municipales, ciudadanos gobernadores: Nos da mucho [...]

Análisis de las mañaneras

¿Qué queremos extraer de la página?

- Título/evento



Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador

· agosto 16, 2021

Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de [...]



Versión estenográfica. Agua Saludable para La Laguna, desde Lerdo, Durango

· agosto 15, 2021

Amigas, amigos de La Laguna, productores, representantes de los sectores productivos de La Laguna, ambientalistas, autoridades municipales, ciudadanos gobernadores: Nos da mucho [...]

Análisis de las mañaneras

¿Qué queremos extraer de la página?

- Título/evento
- Fecha del evento



Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador

agosto 16, 2021

Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de [...]



Versión estenográfica. Agua Saludable para La Laguna, desde Lerdo, Durango

agosto 15, 2021

Amigas, amigos de La Laguna, productores, representantes de los sectores productivos de La Laguna, ambientalistas, autoridades municipales, ciudadanos gobernadores: Nos da mucho [...]



Análisis de las mañaneras

¿Qué queremos extraer de la página?

- Título/evento
- Fecha del evento
- Url

https://lopezobrador.org.mx/2021/08/16/version-estenografica-de-la-conferencia-de-prensa-matutina-del-presidente-Andrés-Manuel-López-Obrador

AMLO CORONAVIRUS COVID-19

Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador

agosto 16, 2021

2021: Año de la Independencia

PRESIDENTE ANDRÉS MANUEL LÓPEZ OBRADOR: Buenos días. Ánimo.

Bueno, vamos a iniciar la semana con el *Quién es quién en los precios*, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de cómo vamos en las obras; al final, la licenciada Laura Velázquez, coordinadora nacional de Protección Civil, va a informarles sobre la ayuda humanitaria que estamos enviando a Haití, a los hermanos de Haití.

Entonces, vamos con Ricardo.

RICARDO SHEFFIELD PADILLA, PROCURADOR FEDERAL DEL CONSUMIDOR: Buenos días, señor presidente; buenos días a todas y a todos ustedes.

Análisis de las mañaneras

¿Qué queremos extraer de la página?

- Título/evento
- Fecha del evento
- Url
- Cuerpo del texto

The screenshot shows a news article from the website of Andrés Manuel López Obrador. The title of the article is "Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador". The date of the conference is listed as "agosto 16, 2021". The text of the speech is presented in a large orange box, with some parts highlighted in blue. The text discusses the start of the week, mentioning Ricardo Sheffield and Laura Velázquez. The footer of the page includes the text "2021: Año de la Independencia".

https://lopezobrador.org.mx/2021/08/16/version-estenografica-de-la-conferenci

AMLO CORONAVIRUS COVID-19

Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador

agosto 16, 2021

2021: Año de la Independencia

PRESIDENTE ANDRÉS MANUEL LÓPEZ OBRADOR: Buenos días. Ánimo.

Bueno, vamos a iniciar la semana con el *Quién es quién en los precios*, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de cómo vamos en las obras; al final, la licenciada Laura Velázquez, coordinadora nacional de Protección Civil, va a informarles sobre la ayuda humanitaria que estamos enviando a Haití, a los hermanos de Haití.

Entonces, vamos con Ricardo.

RICARDO SHEFFIELD PADILLA, PROCURADOR FEDERAL DEL CONSUMIDOR: Buenos días, señor presidente; buenos días a todas y a todos ustedes.

Análisis de las mañaneras

¿Cómo extraemos los elementos?

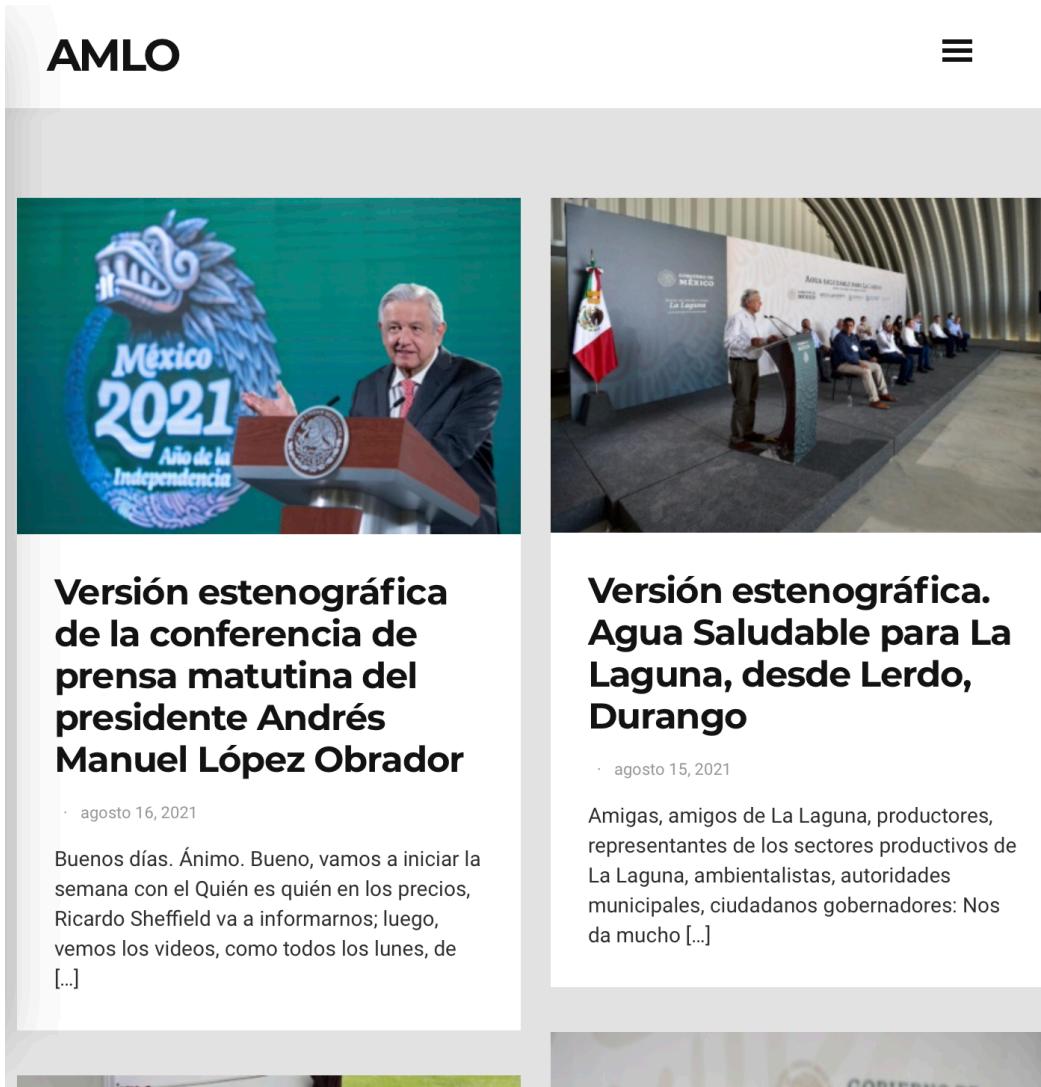
1. Crear un proceso para la extracción del título, fecha y url
2. Repetir este proceso para cada página con información

1 2 3 4 5 6 7 ... 209 210 211

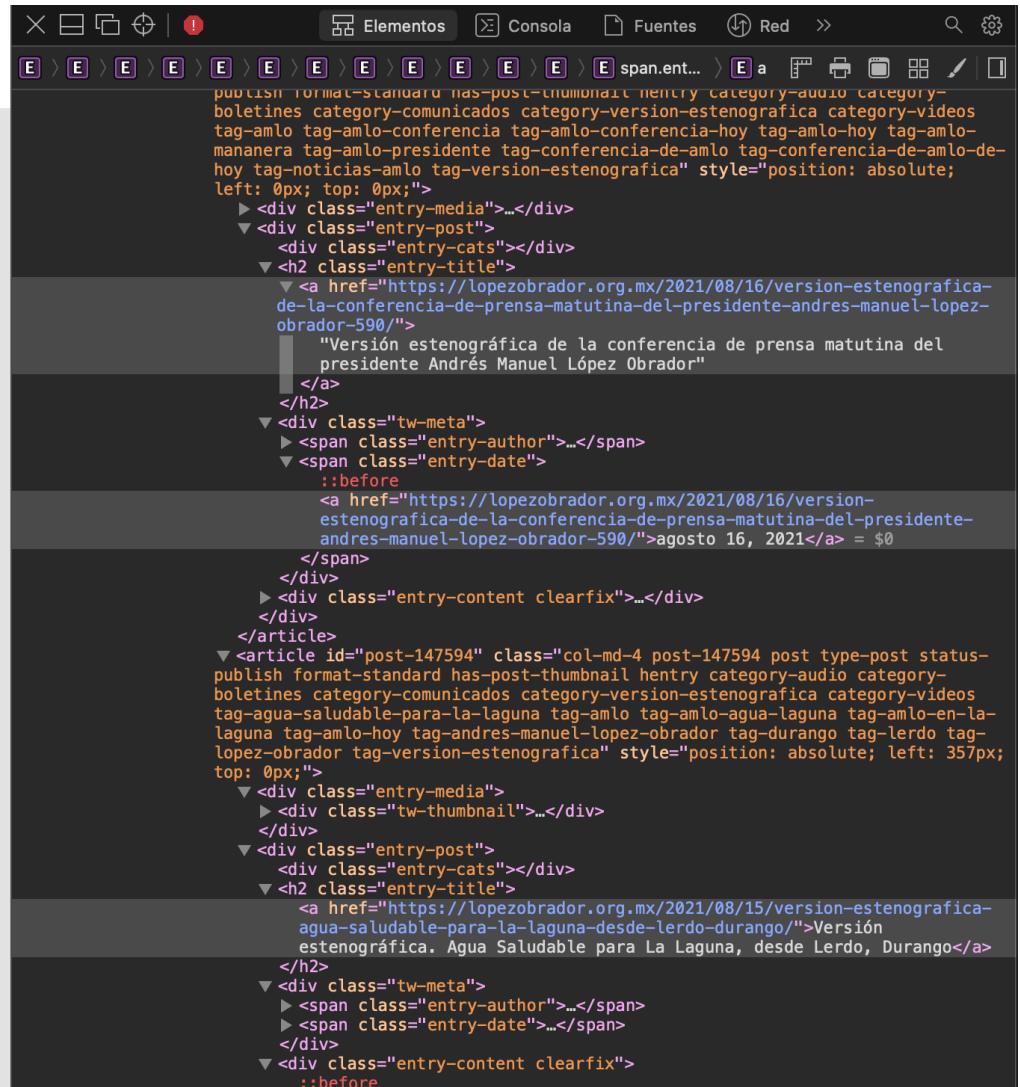
3. Extracción del texto de cada publicación a partir de la url



1. Crear un proceso para la extracción del título, fecha y url



The screenshot shows a news article from the official website of the Mexican government. At the top left is the logo "AMLO". The main image is a photo of President Andrés Manuel López Obrador speaking at a podium. Below the image is the title "Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador". The date "agosto 15, 2021" is listed below the title. The text of the speech begins with "Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de [...]".



The screenshot shows the browser's developer tools with the DOM inspector open. The current element selected is a heading with the text "Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador". The browser's address bar shows the URL "https://lopezobrador.org.mx/2021/08/16/version-estenografica-de-la-conferencia-de-prensa-matutina-del-presidente-andres-manuel-lopez-obrador-590/". The developer tools interface includes tabs for "Elementos", "Consola", "Fuentes", and "Red". The DOM tree shows various HTML elements like

,

, , and tags, along with their corresponding CSS classes and styles.



14 / 41

1. Crear un proceso para la extracción del título, fecha y url

Paquetes a utilizar

R

```
pacman::p_load(tidyverse, # Manejo de dataframe  
                rvest,      # Web scraping de  
                páginas estáticas  
                stringr)    # Manejo de texto
```

Python

```
import pandas # Manejo de dataframe  
import requests # Web scraping de páginas  
estáticas  
from bs4 import BeautifulSoup  
import re      # Manejo de texto  
# Otras librerías para análisis y manejo de texto  
import nltk  
import string
```



1. Crear un proceso para la extracción del título, fecha y url

Extraer información de página inicial

R

```
url <-  
"https://lopezobrador.org.mx/transcripciones/"  
  
pagina <- html_session(url)  
  
list_trans_evento <- pagina %>%  
  html_nodes("h2") %>%  
  html_text()  
  
list_trans_fecha <- pagina %>%  
  html_nodes("span.entry-date a") %>%  
  html_text()  
  
list_trans_link <- pagina %>%  
  html_nodes(".entry-title a") %>%  
  html_attr('href')
```

Python

```
page =  
requests.get("https://lopezobrador.org.mx/transcripciones/")  
  
soup = BeautifulSoup(page.content, 'html.parser')  
  
list_trans_evento = [x.text for x in  
soup.find_all('h2')]  
  
list_trans_fecha = [x.text for x in  
soup.find_all(class_ ='entry-date')]  
  
list_trans_link = [x.find_all("a")[0]["href"]  
for x in soup.find_all('h2', class_ ='entry-title')]
```



1. Crear un proceso para la extracción del título, fecha y url

Dataframe único

R

```
total_eventos <- tibble(evento =  
list_trans_evento,  
                      fecha = list_trans_fecha,  
                      link = list_trans_link)
```

Python

```
total_eventos = pd.DataFrame({'evento':  
list_trans_evento,  
'fecha':  
list_trans_fecha,  
'link':  
list_trans_link})
```



1. Crear un proceso para la extracción del título, fecha y url

Dataframe único

evento	fecha	link
Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-estenografica-de-la-conferencia-de-prensa-matutina-del-presidente-andres-manuel-lopez-obrador-590/
Versión estenográfica. Agua Saludable para La Laguna, desde Lerdo, Durango	agosto 15, 2021	https://lopezobrador.org.mx/2021/08/15/version-estenografica-agua-saludable-para-la-laguna-desde-lerdo-durango/
Versión estenográfica. Visita a la región de la Presa El Zapotillo, en Cañadas de Obregón, Jalisco	agosto 14, 2021	https://lopezobrador.org.mx/2021/08/14/version-estenografica-visita-a-la-region-de-la-presa-el-zapotillo-en-canadas-de-obregon-jalisco/
Versión estenográfica. Ceremonia de revista de entrada de la Comandancia del Ejército Mexicano, en Campo Marte	agosto 13, 2021	https://lopezobrador.org.mx/2021/08/13/version-estenografica-ceremonia-de-revista-de-entrada-de-la-comandancia-del-ejercito-mexicano-en-campo-marte/
Discurso del presidente Andrés Manuel López Obrador en los 500 años de Resistencia Indígena. 1521, México-Tenochtitlan	agosto 13, 2021	https://lopezobrador.org.mx/2021/08/13/discurso-del-presidente-andres-manuel-lopez-obrador-en-los-500-anos-de-resistencia-indigena-1521-mexico-tenochtitlan/
Versión estenográfica de la conferencia de prensa matutina del presidente Andrés Manuel López Obrador	agosto 13, 2021	https://lopezobrador.org.mx/2021/08/13/version-estenografica-de-la-conferencia-de-prensa-matutina-del-presidente-andres-manuel-lopez-obrador-589/



2. Repetir este proceso para cada página con información

Crear loop del proceso

R

```
for (pag in 2:211) {  
  transcripciones <- html_session(paste0("https://lopezobrador.org.mx/transcripciones/",  
                                     "page/", as.character(pag), "/"))  
  
  list_trans_evento <- pagina %>%  
    html_nodes("h2") %>%  
    html_text()  
  
  list_trans_fecha <- pagina %>%  
    html_nodes("span.entry-date a") %>%  
    html_text()  
  
  list_trans_link <- pagina %>%  
    html_nodes(".entry-title a") %>%  
    html_attr('href')  
  
  total_eventos <- bind_rows(total_eventos,  
                               tibble(evento = list_trans_evento,  
                                     fecha = list_trans_fecha,  
                                     link = list_trans_link))  
}
```



2. Repetir este proceso para cada página con información

Crear loop del proceso

Python

```
for x in range(1, 210):

    page = requests.get('https://lopezobrador.org.mx/transcripciones/page/{}/'.format(x))

    soup = BeautifulSoup(page.content, 'html.parser')

    list_trans_evento_loop = [x.text for x in soup.find_all('h2')]

    list_trans_fecha_loop = [x.text for x in soup.find_all(class_ ='entry-date')]

    list_trans_link_loop = [x.find_all("a")[0]["href"] for x in soup.find_all('h2', class_ ='entry-title')]

    total_eventos_loop = pd.DataFrame({'evento': list_trans_evento_loop,
                                         'fecha': list_trans_fecha_loop,
                                         'link': list_trans_link_loop})

    total_evetos = pd.concat([total_eventos_loop, total_evetos])
```



3. Extracción del texto de cada publicación a partir de la url

R

```
texto_transcripciones <- map(total_eventos$link,  
  function(x){html_session(x) %>%  
    html_nodes(".entry-content") %>%  
    html_text() })
```

Python

```
texto_transcripciones = []  
  
for x in total_eventos["link"]:  
  
    page = requests.get(x)  
  
    soup = BeautifulSoup(page.content,  
    'html.parser')  
  
    texto_transcripciones_loop =  
    soup.find_all(class_ = "entry-content")  
    [0].get_text()  
  
    texto_transcripciones +=  
    [texto_transcripciones_loop]
```



3. Extracción del texto de cada publicación a partir de la url

Lista de transcripciones

x

Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de cómo vamos en las obras; al final, la licenciada Laura Velázquez, coordinadora nacional de Protección Civil, va a informarle [...]

Amigas, amigos de La Laguna, productores, representantes de los sectores productivos de La Laguna, ambientalistas, autoridades municipales, ciudadanos gobernadores: Nos da mucho gusto estar de nuevo aquí, en Lerdo, para hablar sobre este proyecto, que es muy importante porque se trata de la salud de [...]

Amigas, amigos de Jalisco: Nos da mucho gusto estar aquí, en Zapotillo, donde se construyó una presa que, aun cuando todavía no se termina, ya ha significado una inversión de alrededor de seis mil millones de pesos. Esta presa no se concluyó por la oposición de tres comunidades que, con razón, no [...]

Amigas, amigos: Agradezco mucho la presencia en este importante acto cultural e histórico de la representante del pueblo mohawk, que nos acompaña y que ha expresado su sentimiento y lo que están haciendo hermanos nuestros en Estados Unidos y en Canadá. También agradezco mucho a Jamescita Mae, s [...]

ncia en este importante acto cultural e histórico de la representante del pueblo Mohawk, que nos acompaña y que ha expresado su sentimiento y lo que están haciendo hermanos nuestros en Estados Unidos y en Canadá. También agradezco mucho a Jamescita Mae, senadora de Arizona. Y nos da mucho gusto [...]

Buenos días. Vamos el día de hoy a presentar un informe sobre una acción que se está llevando a cabo. Consiste en limpiar las bodegas de todos los objetos, mercancías decomisadas. Durante mucho tiempo se creó esta forma de actuar en el gobierno, de llevar a cabo decomisos en aduanas y almacenar, y se [...]

Pero antes de unirlo con el dataframe, vamos a procesarlo.



Limpieza y ordenar texto

Eliminación de saltos, textos y dobles espacios innecesarios

R

Python

```
num_transcrip = 0
transcripcion =
texto_transcripciones[num_transcrip]
transcripcion = transcripcion.replace("\n", " ")
transcripcion = transcripcion.replace("\xa0", " ")
transcripcion = re.sub(' +', ' ', transcripcion)
transcripcion = transcripcion.replace("2021: Año
de la Independencia ", "")
transcripcion =
transcripcion.replace(r"\\\+\\\\+\\\\+\\\\+", "")
```



Limpieza y ordenar texto

Eliminación de saltos, textos y dobles espacios innecesarios

x

PRESIDENTE ANDRÉS MANUEL LÓPEZ OBRADOR: Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo Sheffield va a informarnos; luego, vemos los videos, como todos los lunes, de cómo vamos en las obras; al final, la licenciada Laura Velázquez, coordinadora nacional de Protección Civil, va a informarles sobre la ayuda humanitaria que estamos enviando a Haití, a los hermanos de Haití. Entonces, vamos con Ricardo. RICARDO SHEFFIELD PADILLA, PROCURADOR FEDERAL DEL CONSUMIDOR: Buenos días, señor presidente; buenos días a todas y a todos ustedes. Quién es quién en el precio de los combustibles. Vamos a ver primeramente las gasolineras y el diésel. Aquí podemos ver el comportamiento de los precios promedio, 20 pesos 57 centavos por litro para la gasolina regular, 22 pesos 39 centavos para por litro para la Premium y 21 pesos con 73 centavos para el diésel. Vemos el comportamiento que ha tenido tanto estos tres combustibles como el petróleo, que están íntim



Limpieza y ordenar texto

El problema:

Estas transcripciones incluyen la participación de distintos actores.

¿Qué hacer al respecto?

Estructurar estas participaciones.



Limpieza y ordenar texto

Estructurar intervenciones

R

```
intervensor <- transcripcion %>%
  str_extract_all("(:space:)[:upper:]+)*
  (:space:)[:upper:]+)(\\:)|([:space:]
  [:upper:]+)*,(:space:)[:upper:]+)*(:space:
  [:upper:]+)(\\:)" %>%
  .[[1]]
  unique()
%>%
  str_trim(side = "both") %>%
  .[order(nchar(.), .)]
```

Python

```
def sin_acento(x):
    output =
x.replace('á','a').replace('é','e').replace('í','
í').replace('ó','o').replace('ú','u')\
.replace('Á','A').replace('É','E').replace('Í','I
').replace('Ó','O').replace('Ú','U')
    return output

intervensor = sin_acento(transcripcion)
intervensor = re.findall("((\\s[A-Z]+)*)(\\:)|
((\\s[A-Z]+)*),((\\s[A-Z]+)+)(\\:)", 
intervensor)
intervensor = np.unique(intervensor)
intervensor = intervensor.strip()
intervensor = sorted(intervensor, key=len)
intervensor = [x for x in intervensor if x not
in {'', ':'}]
```



Limpieza y ordenar texto

Estructurar intervenciones

x

PREGUNTA:

VOZ MUJER:

VOZ HOMBRE:

INTERLOCUTOR:

INTERVENCIÓN:

INTERLOCUTORA:

ROCÍO NAHLE GARCÍA:

LAURA VELÁZQUEZ ALZÚA:

ANDRÉS MANUEL LÓPEZ OBRADOR:

PRESIDENTE ANDRÉS MANUEL LÓPEZ OBRADOR:

ROCÍO NAHLE GARCÍA, SECRETARIA DE ENERGÍA:

RICARDO SHEFFIELD PADILLA, PROCURADOR FEDERAL DEL CONSUMIDOR:

LAURA VELÁZQUEZ ALZÚA, COORDINADORA NACIONAL DE PROTECCIÓN CIVIL:



Limpieza y ordenar texto

Separación por intervención

R

```
intervensor_text <- paste0("~~~", intervensor)

names(intervensor_text) <- intervensor

transcripcion_final <- transcripcion %>%
  str_replace_all(intervensor_text) %>%
  str_split(pattern = "~~~") %>% .[[1]] %>%
  as_tibble() %>%
  rename(texto = value) %>%
  mutate(orador = str_extract(texto, paste0(intervensor,
                                             collapse = "|"))),
  orador = str_remove(orador, "\\:"),  

  orador = str_replace_all(orador,  

                          "ANDRÉS MANUEL LÓPEZ OBRADOR PRESIDENTE|PRESIDENTE ANDRÉS MANUEL LÓPEZ  

OBRADOR|ANDRÉS MANUEL LÓPEZ OBRADOR", "AMLO"),
  palabras_amlo = ifelse(str_detect(orador, "AMLO"), 1, 0),
  texto = str_remove_all(texto, paste0(intervensor,
                                         collapse = "|")),
  texto = str_remove_all(texto, "PRESIDENTE"),
  texto = str_trim(texto, side = "both"),
  num = num_transcrip)
```



Limpieza y ordenar texto

Separación por intervención

Python

```
intervensor_text = ["~~~"+x for x in intervensor]
transcripcion = sin_acento(transcripcion)

transcripcion_final = pd.DataFrame([transcripcion])[0].replace(intervensor,intervensor_text,regex=True)[0]
transcripcion_final = transcripcion_final.split("~~~")
transcripcion_final = pd.DataFrame({'texto': transcripcion_final})

intervensor_2 = [x+": " for x in intervensor]

transcripcion_final['orador'] = transcripcion_final['texto'].str.extract(r"((\s[A-Z]+)*)(\:)|((\s[A-Z]+)*),\n((\s[A-Z]+)+)(\:)")[0]
transcripcion_final['texto'] = transcripcion_final['texto'].replace(intervensor_2,"",regex=True)
transcripcion_final['num'] = num_transcrip
```



Limpieza y ordenar texto

Separación por intervenciones

texto	orador	palabras_amlo	num
Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo S	AMLO	1	1
Buenos días, señor presidente; buenos días a todas y a todos ustedes. Quién es quién en el precio de	RICARDO SHEFFIELD PADILLA, PROCURADOR FEDERAL DEL CONSUMIDOR	0	1
Vamos con los videos. (INICIA VIDEO)	AMLO	1	1
La Secretaría de la Defensa Nacional informa los avances en la construcción del Aeropuerto Internaci	VOZ MUJER	0	1
Hoy es viernes 13 de agosto y vamos a mandar el reporte de Dos Bocas desde la Ciudad de Pyeongtaek,	ROCÍO NAHLE GARCÍA, SECRETARIA DE ENERGÍA	0	1
Esta semana se realizaron visitas de supervisión a los talleres donde se fabrican algunos módulos de	VOZ HOMBRE	0	1
En estamos en Hyosung, Corea del Sur, donde están fabricando 30 módulos del paquete 2 y 3 a cargo de	ROCÍO NAHLE GARCÍA	0	1
En la Ciudad Ulsan, se supervisó el avance y la fabricación de equipos críticos.	VOZ HOMBRE	0	1



Limpieza y ordenar texto

Realizar proceso de limpieza a los demás textos

Funciones + Loops

R

```
limpieza_conferencias <- function(num_texto) {  
  # Estructurar intervenciones  
  # Separación por intervenciones  
  # Integrar a un dataframe  
}  
  
texto_transcripciones_final <-  
map_df(1:length(texto_transcripciones),  
  
  limpieza_conferencias) %>%  
  left_join(total_eventos %>%  
            mutate(num = row_number()),  
            by = "num")
```

Python

```
def limpieza_conferencias(num_texto):  
  # Estructurar intervenciones; Separación por  
  # intervenciones ;Integrar a un dataframe  
  
  texto_transcripciones_final = []  
  
  for x in range(0, len(texto_transcripciones)):  
    texto_transcripciones_final +=  
      limpieza_conferencias(x)  
  
  total_eventos["num"] = ['C'] =  
  np.arange(len(total_eventos))  
  texto_transcripciones_final =  
  pd.merge(texto_transcripciones_final,  
           total_eventos,  
           on = "num")
```



Limpieza y ordenar texto

Realizar proceso de limpieza a los demás textos

texto	orador	palabras_amlo	num	evento	fecha	link
Buenos días. Ánimo. Bueno, vamos a iniciar la semana con el Quién es quién en los precios, Ricardo S	AMLO		1	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est
Buenos días, señor presidente; buenos días a todas y a todos ustedes. Quién es quién en el precio de	RICARDO SHEFFIELD PADILLA, PROCURADOR FEDERAL DEL CONSUMIDOR		0	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est
Vamos con los videos. (INICIA VIDEO)	AMLO		1	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est
La Secretaría de la Defensa Nacional informa los avances en la construcción del Aeropuerto Internaci	VOZ MUJER		0	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est
Hoy es viernes 13 de agosto y vamos a mandar el reporte de Dos Bocas desde la Ciudad de Pyeongtaek,	ROCÍO NAHLE GARCÍA, SECRETARIA DE ENERGÍA		0	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est
Esta semana se realizaron visitas de supervisión a los talleres donde se fabrican algunos módulos de	VOZ HOMBRE		0	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est
En estamos en Hyosung, Corea del Sur, donde están fabricando 30 módulos del paquete 2 y 3 a cargo de	ROCÍO NAHLE GARCÍA		0	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est
En la Ciudad Ulsan, se supervisó el avance y la fabricación de equipos críticos.	VOZ HOMBRE		0	Versión estenográfica de la conferencia de prensa	agosto 16, 2021	https://lopezobrador.org.mx/2021/08/16/version-est



Ejemplos de análisis

Catorce sexenios: los discursos de toma de posesión de Cárdenas a López Obrador

Manuel Toral



DESARMAR LA CORRUPCIÓN

CORRÓMPEME OTRA VEZ

ARTÍCULOS DE INVESTIGACIÓN

CONCURSO

Catorce
sexenios: los
discursos de
toma de
posesión de
Cárdenas a
López Obrador

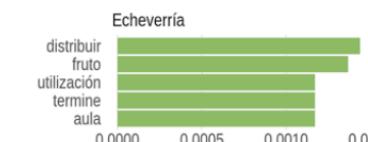
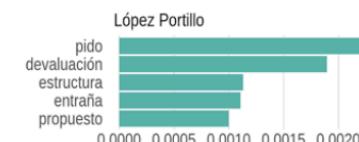
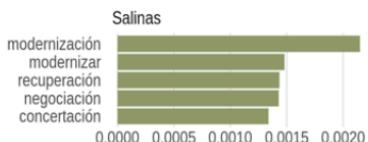
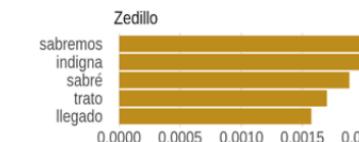
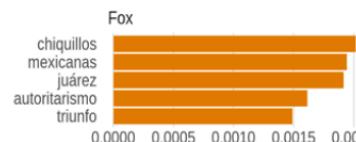
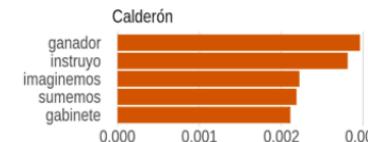
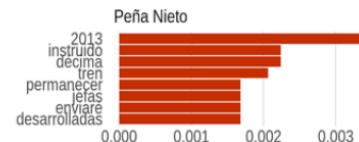
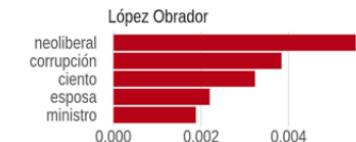
0

Compartir

Twittear

Las 5 palabras más características

por cada discurso inaugural



¿De qué habla el presidente cuando habla de progreso social?

Katia Guzmán

nexos

TALLER DE DATOS CONTEXTO DEBATE HALLAZGOS DATA CÍVICA MÉXICO, ¿CÓMO VAMOS? PUNTO DECIMAL

F

¿De qué habla el presidente cuando habla de progreso social?

0

 Compartir

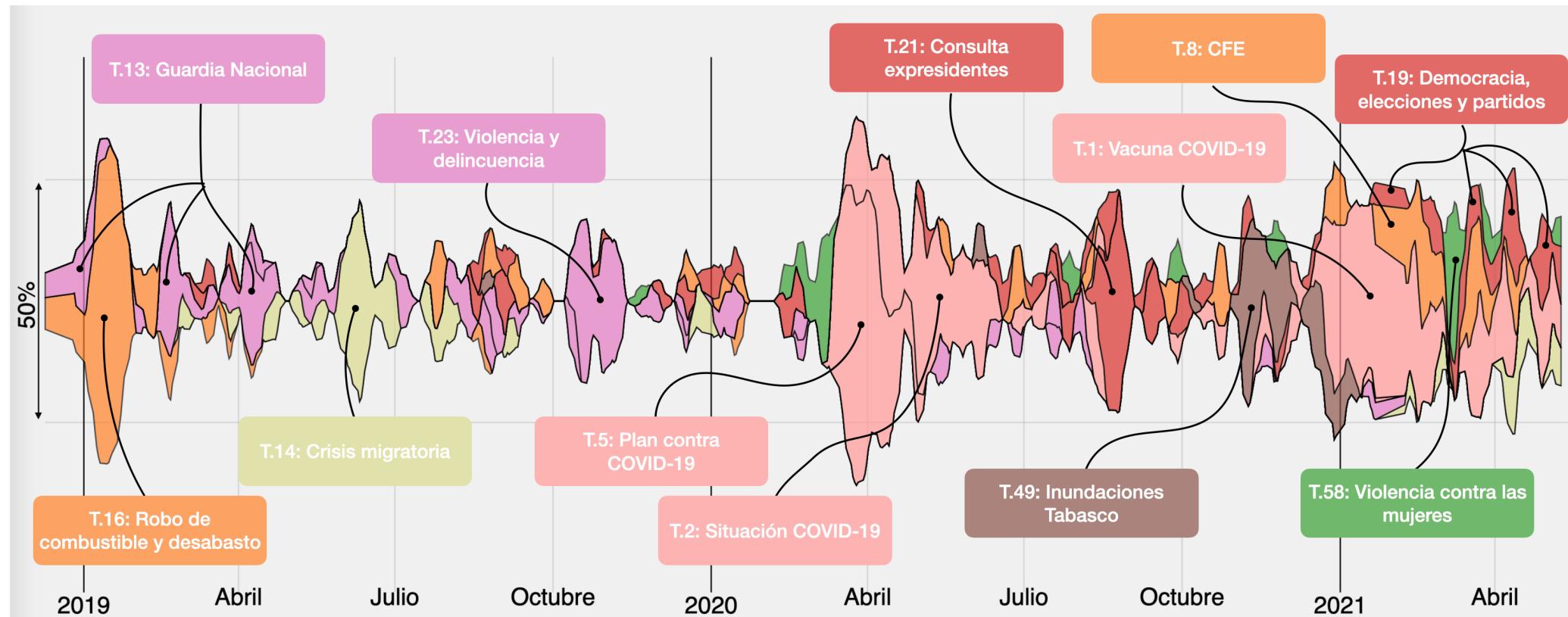
 Twittear

	Nutrición y cuidados médicos básicos	Agua y saneamiento	Vivienda	Seguridad personal
Necesidades Humanas Básicas	UNIDOS ECONOMÍA PAÍSES PANDEMIA CRISIS DESARROLLO CRECIMIENTO RELACIÓN MEXICANOS EMPLEOS	CENTAVOS PRECIO LITRO MARGEN PRECIOS QUIÉN GAS SEMANA VERACRUZ GASOLINA	AEROPUERTO OBRA TREN PROYECTO OBRAS VIDEO SURESTE MAYA CAMPECHE SANTA	GUARDIA MARINA ELEMENTOS ROBO DEFENSA SECRETARIO PROTECCIÓN POLICÍA BAJA PAZ
Fundamentos del Bienestar	Acceso a conocimientos básicos SALUD HOSPITALES MÉDICOS MEDICAMENTOS SISTEMA HOSPITAL ATENCIÓN MAESTROS SEGURO MEDICINAS	Acceso a la información y comunicaciones NADIE IMAGÍNENSE NUNCA MAL RESPETO CAMPÀÑA VECES CONSERVADORES CAMBIO COMUNICACIÓN	Salud y bienestar PERSONAS CIENTO CASOS MOMENTO GRACIAS SEMANA SOCIAL ATENCIÓN EPIDEMIA BUENOS	Calidad medioambiental PEMEX PETRÓLEO PRODUCCIÓN COMISIÓN EMPRESAS ELECTRICIDAD EMPRESA BARRILES GAS ELÉCTRICA
Oportunidades	Derechos personales LEY FISCALÍA INVESTIGACIÓN JUSTICIA REFORMA CONSTITUCIÓN HUMANOS PROCESO NADIE DEBE	Libertad personal y de elección PROGRAMA BIENESTAR MAYORES ADULTOS PRESUPUESTO PENSIÓN SABEN BECAS MENSUALES VIVA	Inclusión HISTORIA MUJERES CIUDAD GRACIAS PUEBLOS TRANSFORMACIÓN JUÁREZ CONSTITUCIÓN REVOLUCIÓN GRAN	Acceso a educación superior PRESUPUESTO EMPRESAS IMPUESTOS TRABAJADORES RECURSOS HACIENDA DEUDA AVIÓN ARRIBA AUSTERIDAD



LOS TEMAS DEL PRESIDENTE EN LAS CONFERENCIAS MATUTINAS

Humberto González



Posibles análisis

Posibles análisis

- ¿Qué dice Claudia Sheinbaum sobre la seguridad en la CdMx?

Selecciona un tipo de publicación

- Boletines
- Discurso**
- Entrevista
- Noticias
- Síntesis informativa
- Versiones

Notas por tipo de publicación

Filtrar por

Año

Todos

Mes

Todos los meses

Día

Todos

Mostrando 10 de 732 publicaciones - página 1 de 74

Mensaje de la Jefa de Gobierno, Claudia Sheinbaum Pardo; y del secretario de Obras y Servicios, Jesús...

JEFA DE GOBIERNO, CLAUDIA SHEINBAUM PARDO (CSP): Muy buenas tardes. Me da mucho gusto estar aquí, en Venustiano Carranza; saludo al alcalde Manuel Ballesteros; y, a la próxima alcaldesa, qué bueno que están aquí con...



Posibles análisis

- ¿Qué dice Silvano Aureoles sobre la seguridad en Michoacán?



Ciudadanos Gobierno Prensa Entérate Transparencia Contacto



Mensajes del Gobernador



lunes 9, agosto, 2021

No hay condiciones para regreso a clases presenciales en Michoacán: Silvano

Prensa Mensajes del Gobernador



lunes 2, agosto, 2021

Tercera ola, la más grave en toda la epidemia: Silvano Aureoles

Prensa Mensajes del Gobernador



**Gobierno
del Estado
de Michoacán**

jueves 22, julio, 2021

Rebrote de COVID-19 amenaza a Michoacán: Silvano

Prensa Mensajes del Gobernador



Posibles análisis

- ¿Qué dice Francisco Domínguez sobre la seguridad en Querétaro?

 **QRO** ORGULLO DE MX y mayores logros. ¡Claro que lo lograremos!: Francisco Domínguez

Buscar en queretaro.gob.mx

[Ver nota completa >](#)

[Inicio](#)



MENSAJE POLÍTICO DEL GOBERNADOR FDS

 domingo, 15 de agosto de 2021

[Ver nota completa >](#)



SEXTO INFORME DE GOBIERNO DE FDS

 domingo, 15 de agosto de 2021

[Ver nota completa >](#)



Posibles análisis

- ¿Qué dijeron los senadores al respecto de la seguridad?
- ¿Qué ha discutido la comisión de Seguridad pública?

Versiones Estenográficas
¿Qué es? | ¿Qué contiene? | ¿Cómo se consulta?

LXIV Legislatura (2018-2021)

2021

Agosto

Calendario de versiones estenográficas

Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Dom	Lun	Mar	Mier	Jue	Vie	Sab					
1	2	3	4	5	6	7					
8	9	10	11	12	13	14					
15	16	17	18	19	20	21					
22	23	24	25	26	27	28					
29	30	31									

Última versión estenográfica

LXIV Legislatura
Lunes 09 de agosto de 2021
Segundo Periodo Ordinario

LXIV
2018-2021

LXIII
2015-2018

LXII
2012-2015

LXI
2009-2012

[Facebook](#) [Twitter](#) [WhatsApp](#) [Email](#)