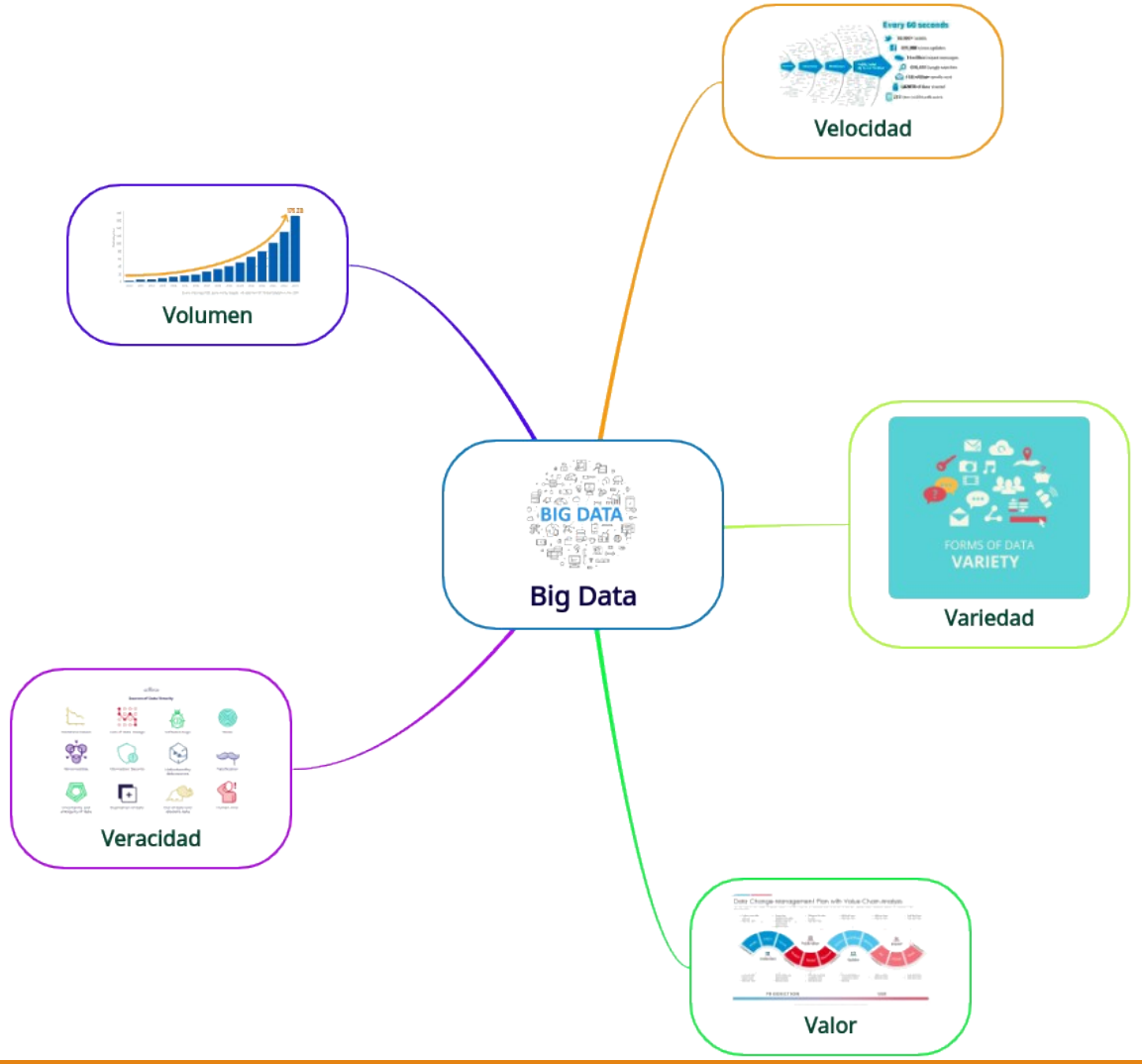
Programa Ingeniería de Sistemas

Herramientas de Software Libre para la obtención de datos Big Data

Juan Francisco Mendoza Moreno, PhD


BIG DATA

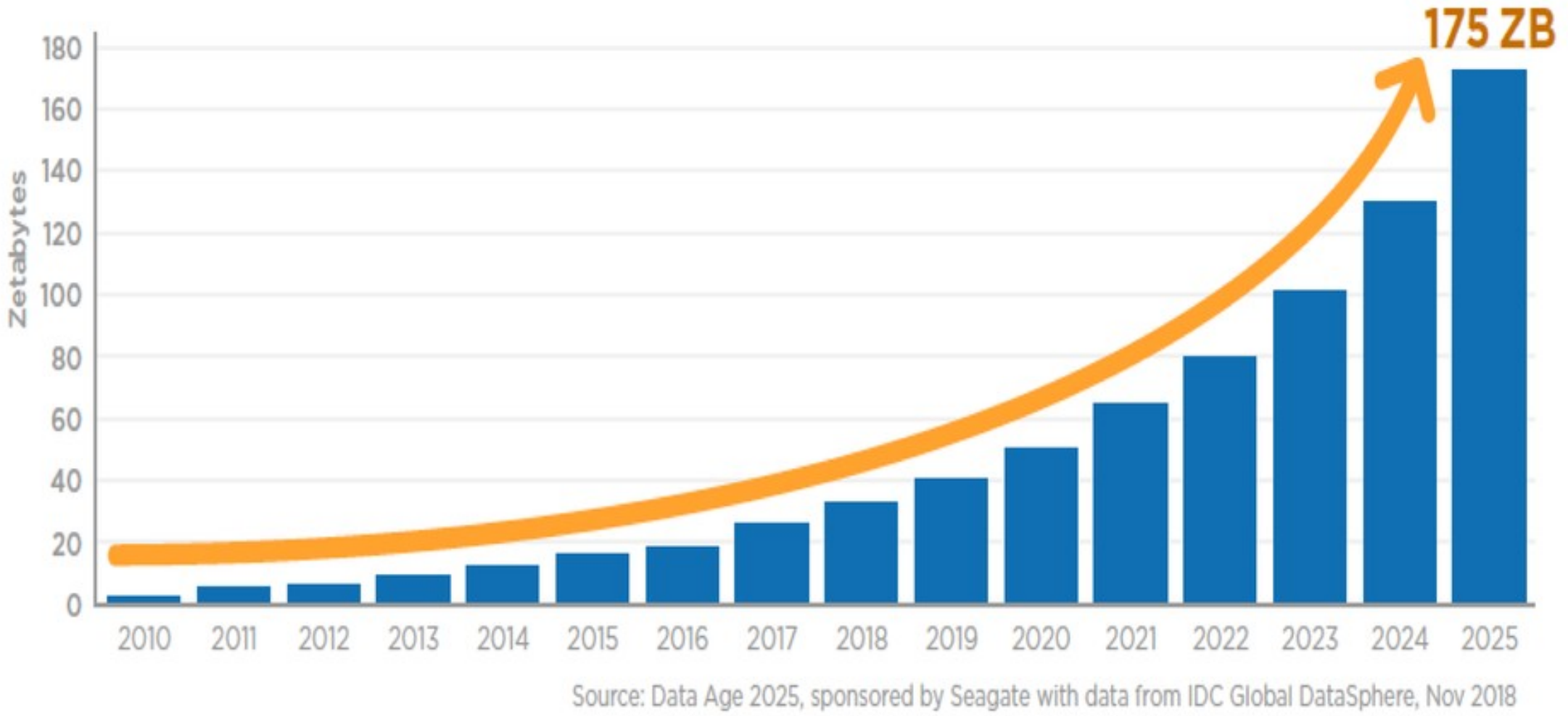
¿Qué es Big Data?



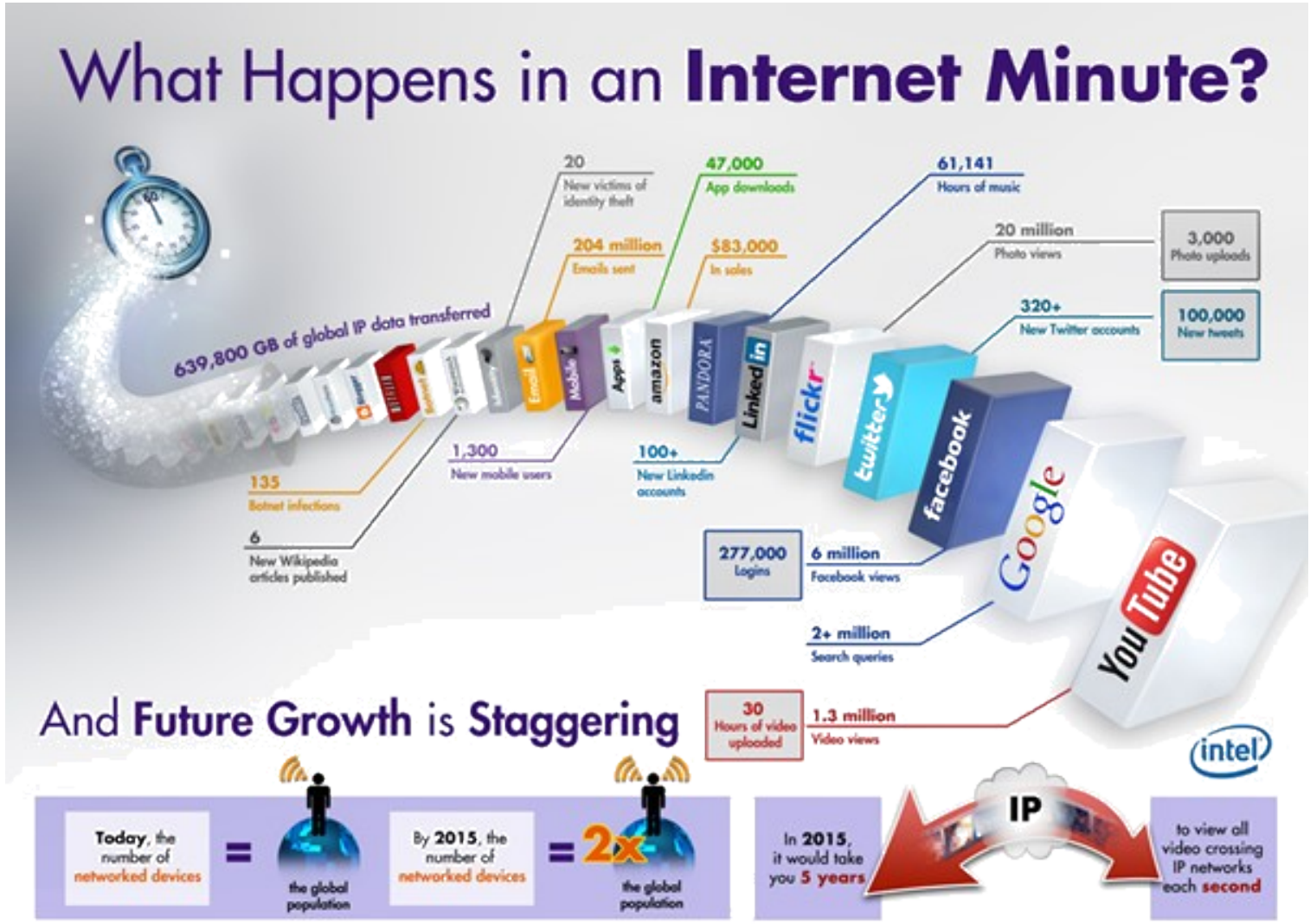
Datos: El petróleo de nuestra era

Fuente Imagen: www.OleoShop.com

Volumen

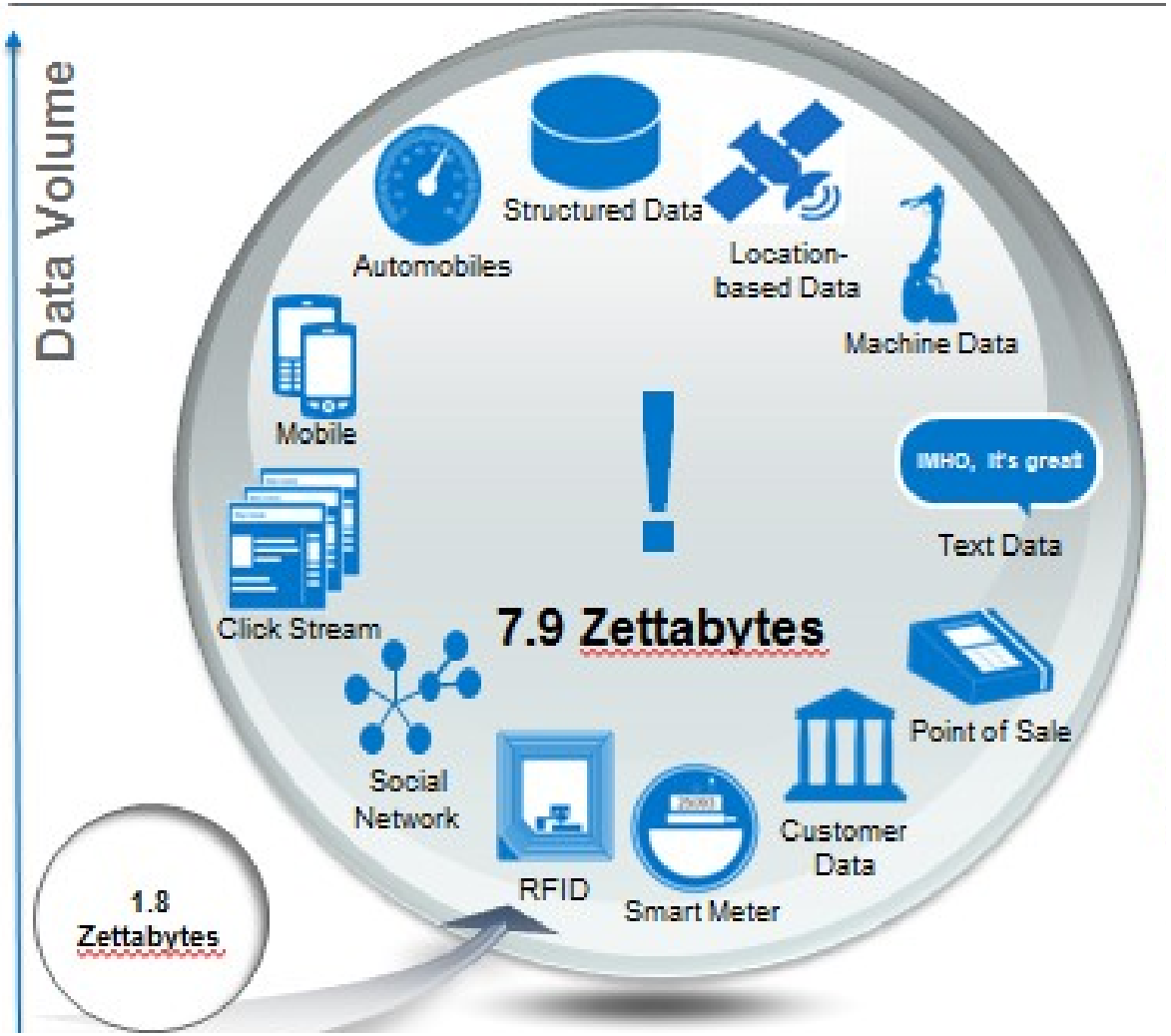

175 ZB
Zetabytes
180
160
140
120
100
80
60
40
20
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Velocidad



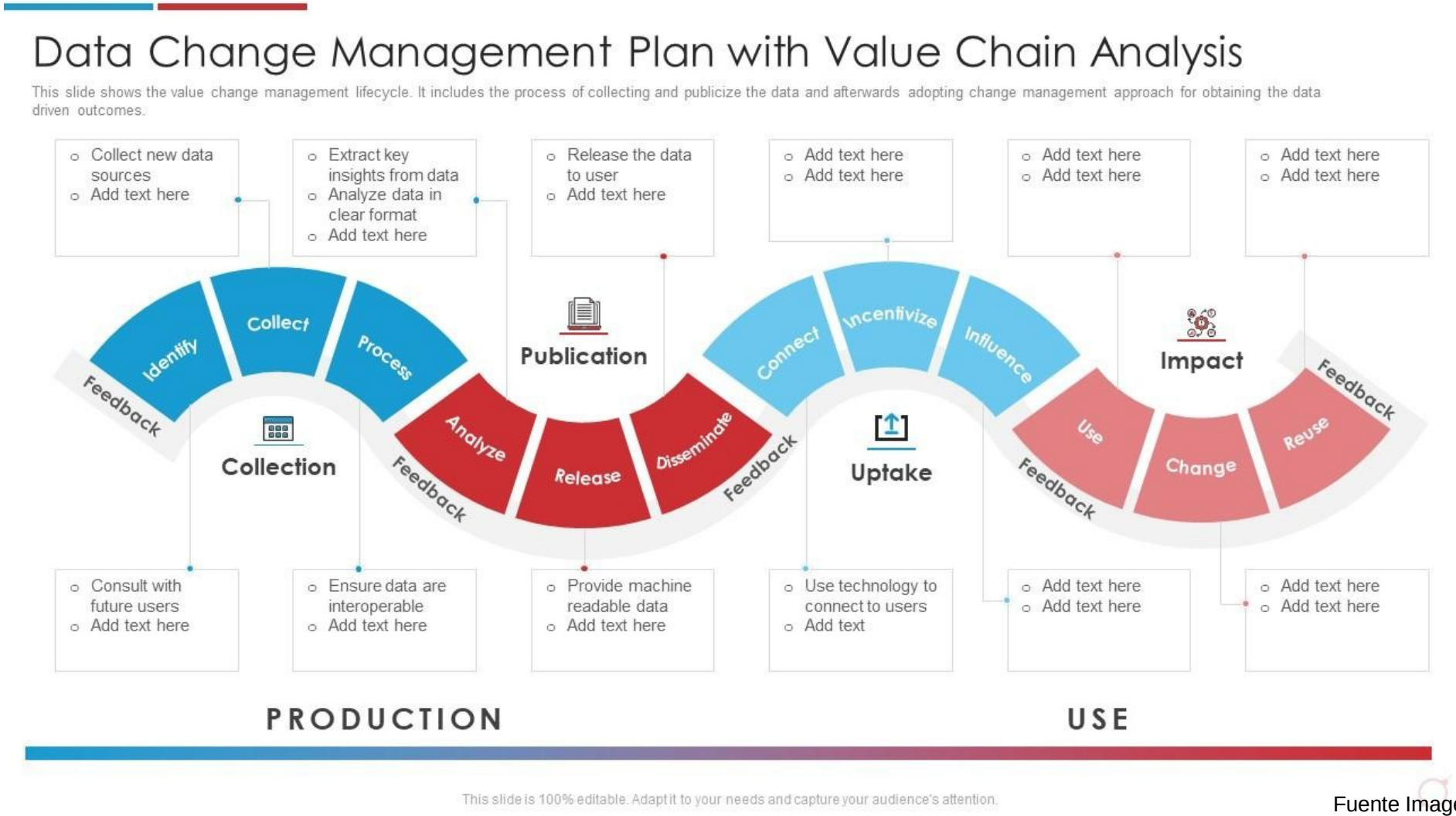
Fuente Imagen: www.LifeHach.org

Variedad



Fuente Imagen: https://insightextractor.com/

Valor



Fuente Imagen: www.SlideTeam.com/

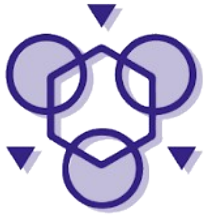
Veracity

Statistical biases

Lack of data lineage

Software bugs

Noise



Abnormalities

Information Security

Untrustworthy data sources

Falsification

Uncertainty and ambiguity of data

Duplication of data

Out of date and obsolete data

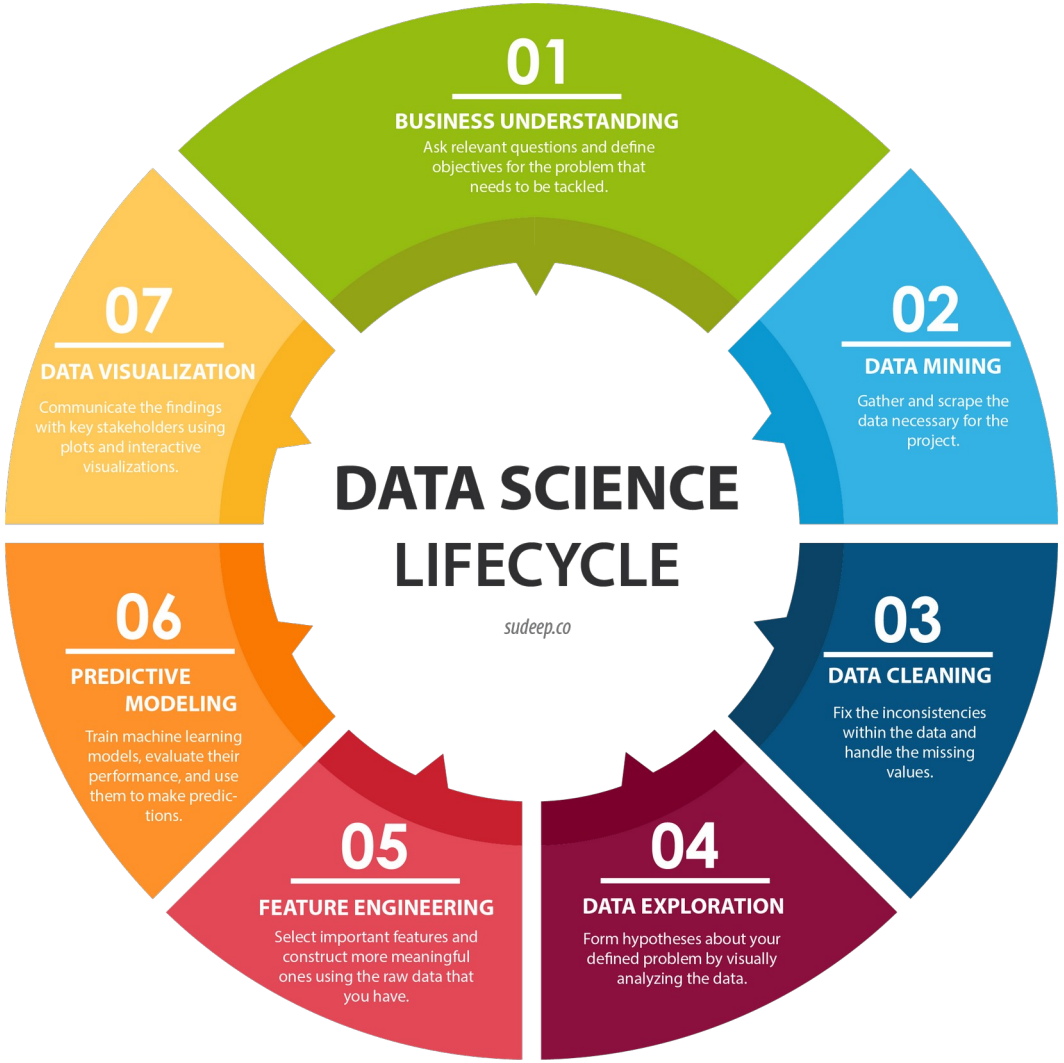
Human error

Fuente Imagen: www.SciForce.com/

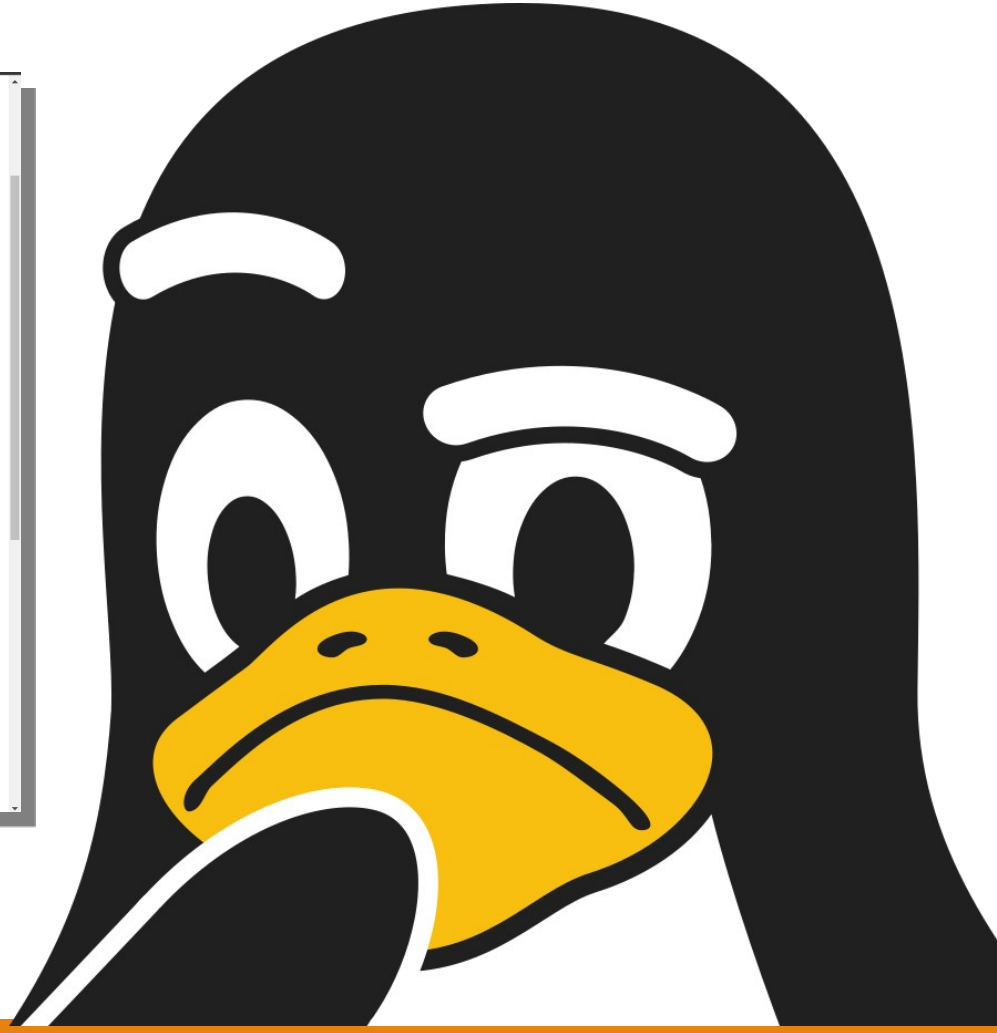
Fuentes de datos

Fuente Imagen: www.SlideTeam.com/

Ciclo de Vida de Big Data



Comprensión del negocio



Minería de datos

WebScraping

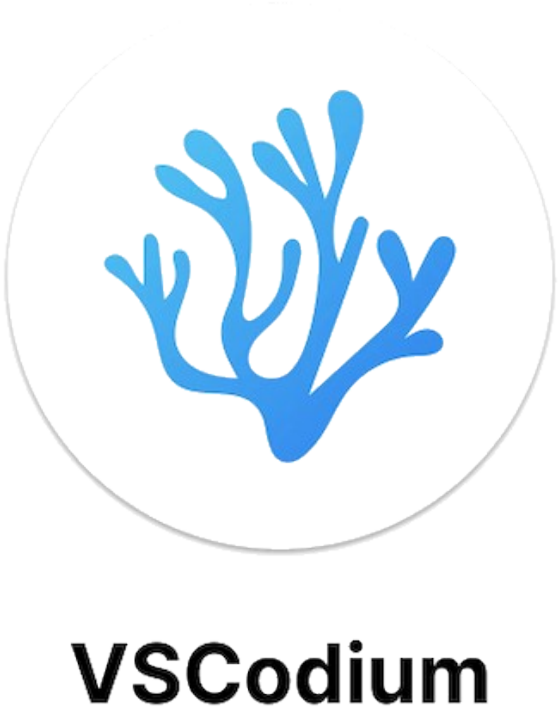
- https://youtu.be/6bXjEeVq0q4 - Infraestructure as Code in Cloud using Terraform
- https://youtu.be/9vTJ9e9ub-M - Software y hardware libre para una cultura maker
- https://youtu.be/JsSpfTgEqnw - Programa de Comunidad Wikimedia Colombia "
- https://youtu.be/1uobKpVuEsQ - Software en el contexto de las elecciones de Colombia
- https://youtu.be/5oFFcDRWziQ - Proyecto FotoE14
- https://youtu.be/G60yT8HFFfM - Datos abiertos y experiencias de uso de software libre en Bogotá
- https://youtu.be/DiDLAYlp4zw - Espacios de aprendizaje seguros - WWCode Bogotá
- https://youtu.be/fT224vy3uOo - Mapeemos tu casa en OpenStreetMap - Taller
- https://youtu.be/oznrPSTXy2I - El valor de las certificaciones Linux y Open Source en un Mercado Global
- https://youtu.be/ndEwoMSoIa4 - Privacidad y Software Libre en tiempos de persecución
- https://youtu.be/ur4YhzkPS_4 - Flutter, Un solo código para muchas plataformas
- https://youtu.be/ebDI4KbM7Rc - Despliegue de aplicaciones con GitLab CI
- https://youtu.be/0fEUBGWlcRI - Como aprender a desarrollar en Ubuntu Touch con web technologies sin morir en el intento.
- https://youtu.be/nf9xFlW3Bdw - A path forward for Latin America
- https://youtu.be/w8xlxPWXXHU - Tú eres software libre, yo soy software libre, todos somos software libre.
- https://youtu.be/w37Dz12wPZU - Escritura musical para principiantes con MuseScore
- https://youtu.be/jqr5s_LdIEU - Zoomorfismo aplicado a la Robótica
- https://youtu.be/qnwotgqFKsg - FIWARE la plataforma para el desarrollo y despliegue de aplicaciones de Internet del Futuro
- https://youtu.be/vOYStvGx404 - Blockchain criptomonedas, NFTs y play to earn
- https://youtu.be/s01CSxDmvcY - Matrix Chat Seguro, Autónomo y Fedarado para Organizaciones
- https://youtu.be/S7Ap3EsUtBg - Processing, arte digital.
- https://youtu.be/5L2ZLgh2iX8 - Computadora en una caja
- https://youtu.be/LqQf0Wa36j0 - Liberación de software y legalidad
- https://youtu.be/DbHgNEb-nJY - Fotomanipulacion con Gimp

Fuente Imagen: www.DailyTech.com/

Python

- **Función**: Python es un lenguaje de programación que nos permite trabajar ágilmente e integrar sistemas de forma eficiente.
- **Instalación**: https://www.python.org/downloads/
- **Uso específico:**
  - Abrir la(s) página(s) web
  - Hacer Web Scraping para obtener el listado de URLs de Youtube

IDE - VSCodium



- **Función**: La alternativa más fácil a VS Code, construido sin las adiciones propietarias de Microsoft.
- **Instalación**: https://vscodium.com/
- **Uso específico:** Editar los scripts de nuestra solución Big Data

Requests: HTTP for Humans

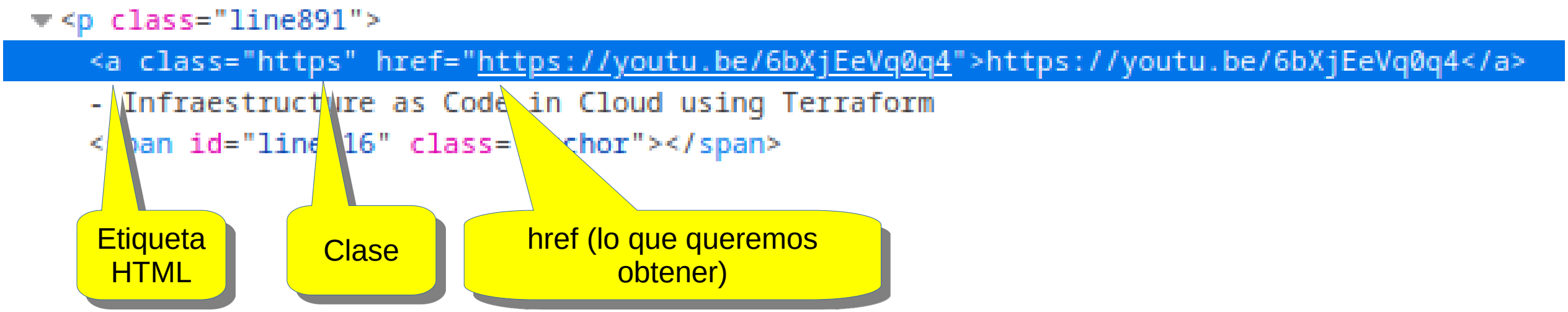
- **Función**: Es una biblioteca HTTP muy simple y elegante, para Python.
- **Instalación**:
  ```
  python -m pip install requests
  ```
- **Uso específico**:
  Abrir nuestra(s) página(s) objetivo para obtener las URLs de Youtube.

Beautiful Soup

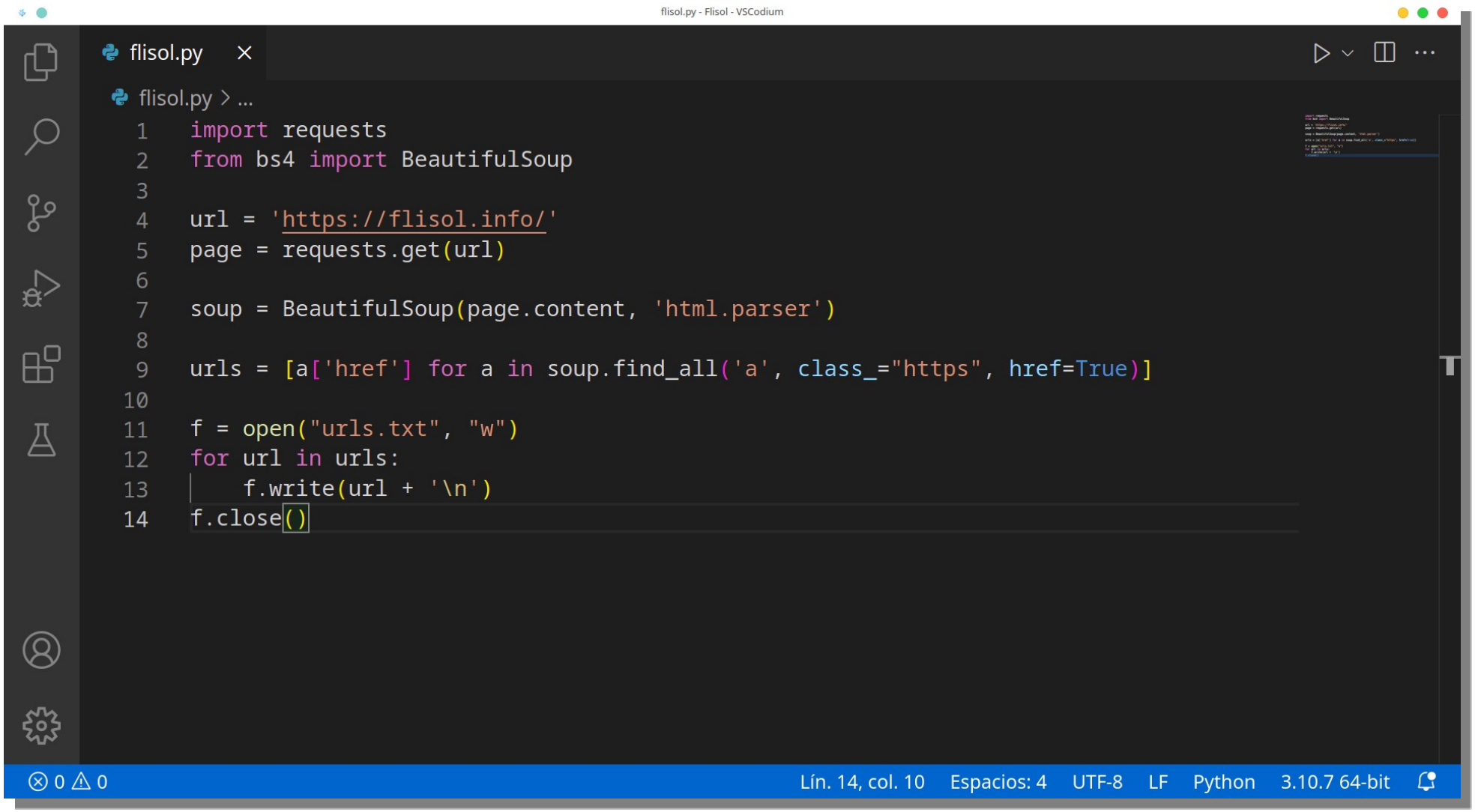
- **Función**: Beautiful Soup es una biblioteca de Python para extraer datos de archivos HTML y XML.
- **Instalación**:
  ```
  pip install beautifulsoup4
  ```
- **Uso específico**:
  Hacer Web Scraping a nuestra(s) página(s) objetivo para obtener las URLs de Youtube.

Herramientas del Desarrollador (F12)

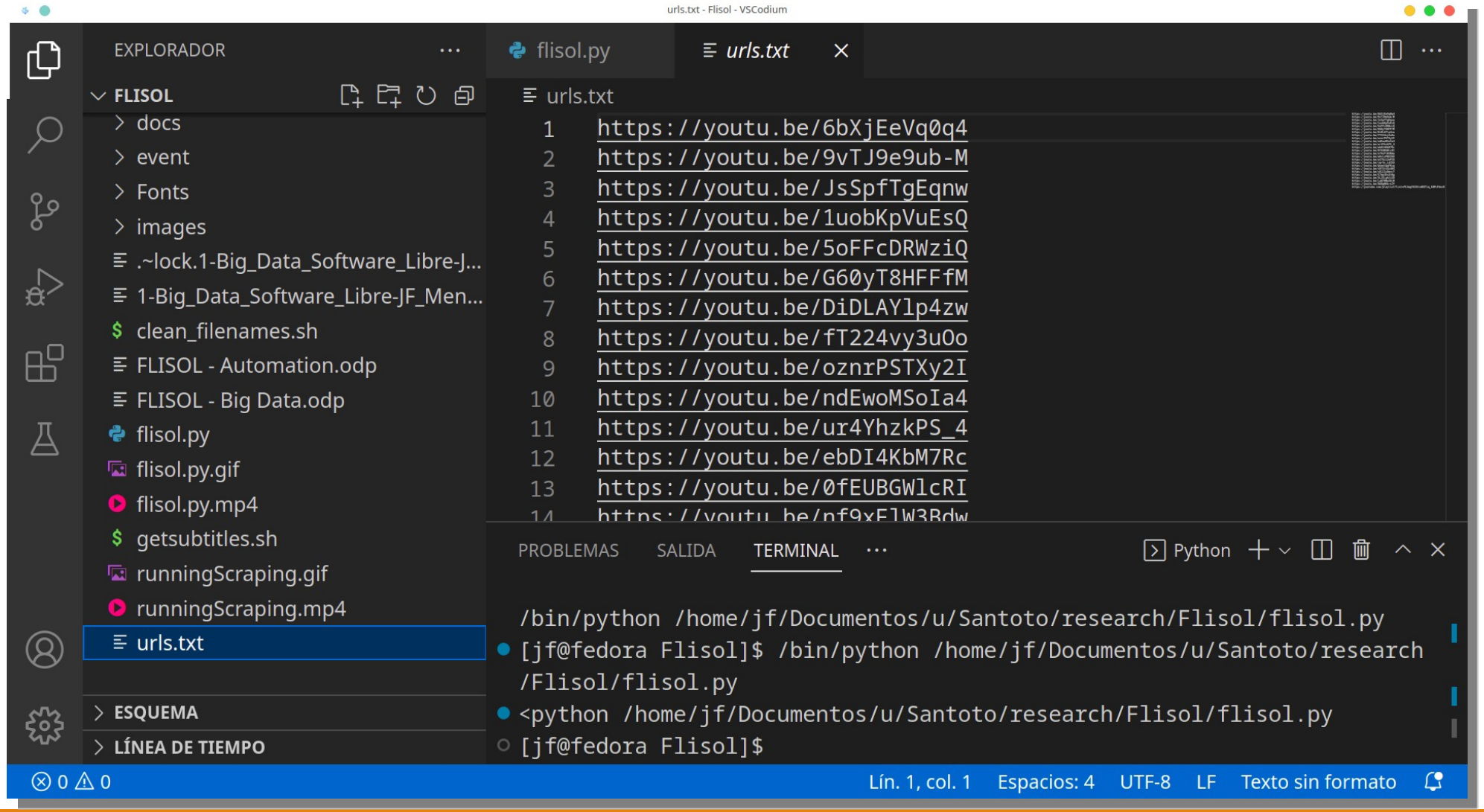
Análisis para obtener la URL



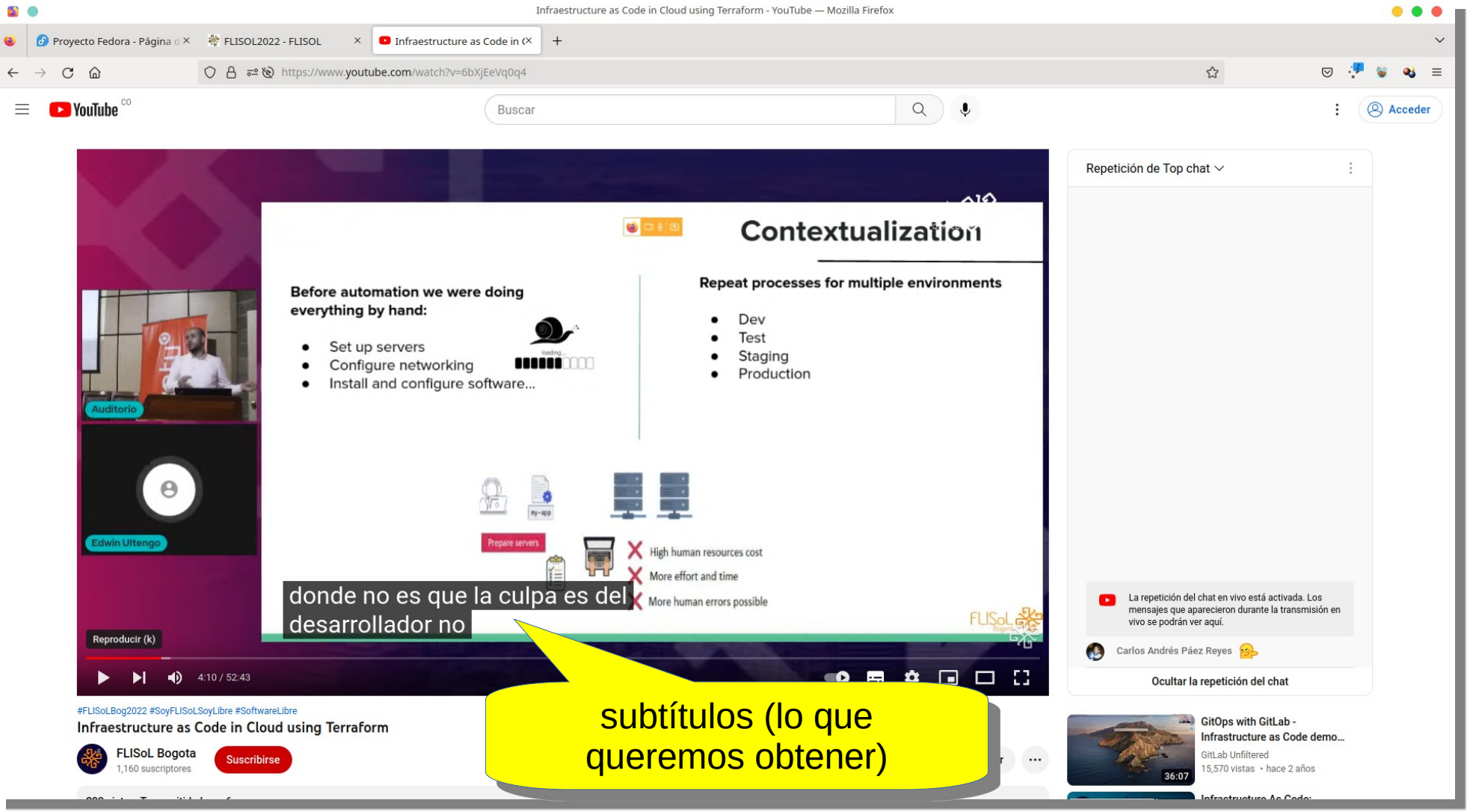
Demo codificación del Web Scraping



Demo ejecución del Web Scraping

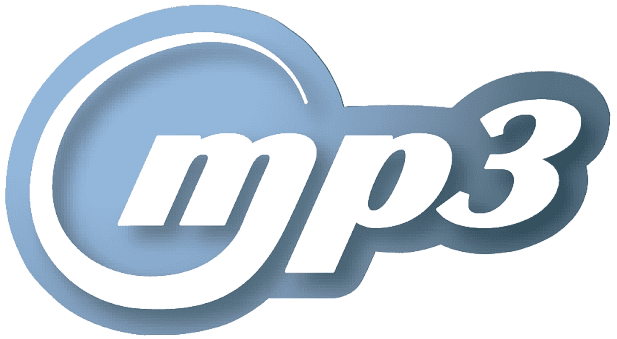


Subtítulos de Youtube

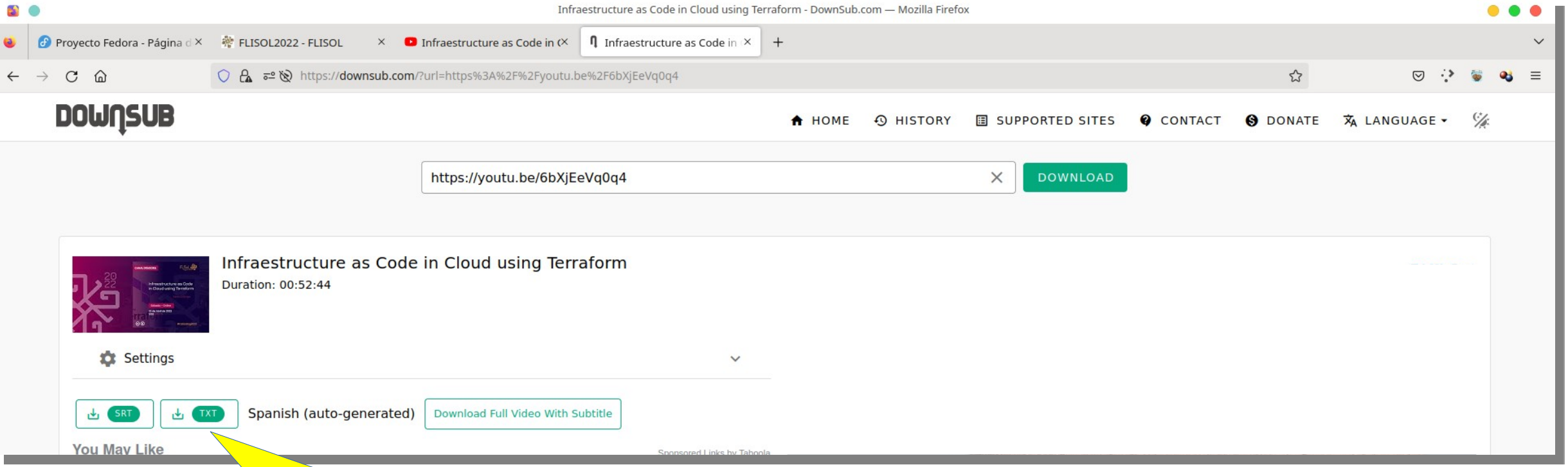


Software de transcripción

- Transcriberry
- Transcriber
- Typingpool
- Transcription Buddy
- Transcription Helper
- VoiceWalker
- Transcribe!
- Express Scribe
- Happy Scribe
- Otranscribe
- Rev
- Express Scribe
- Trint
- Otter
- Temi
- Descript
- Nuance
- Sonix
- Audext
- **Youtube**



Descarga de subtítulos de Youtube



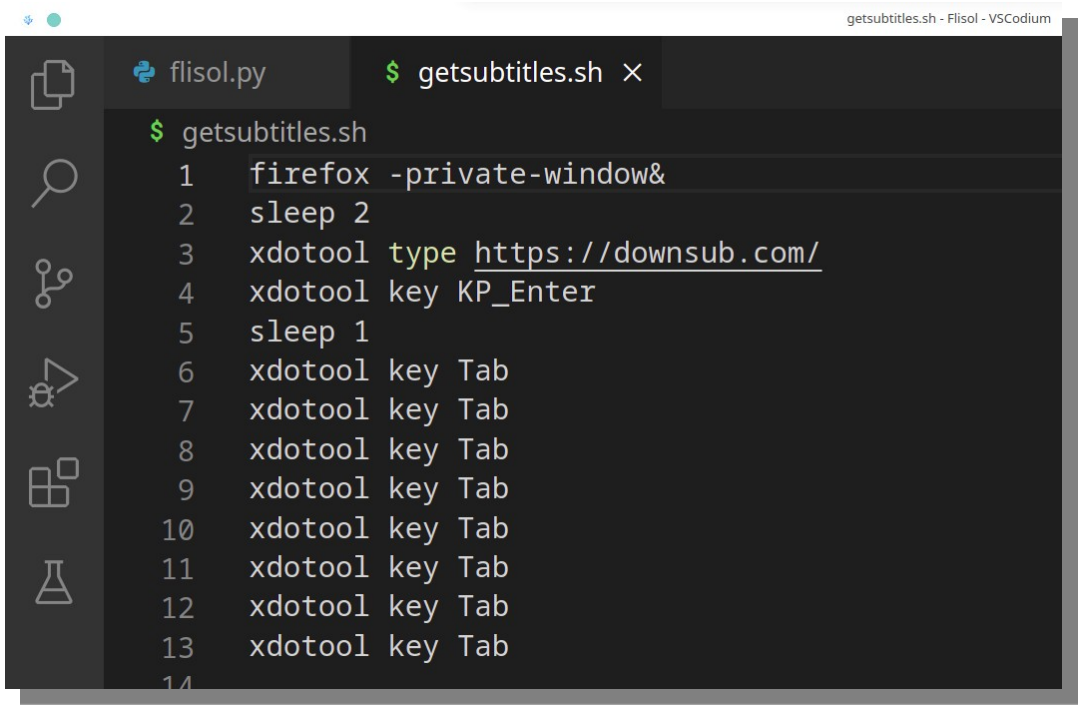
Descargar subtítulos

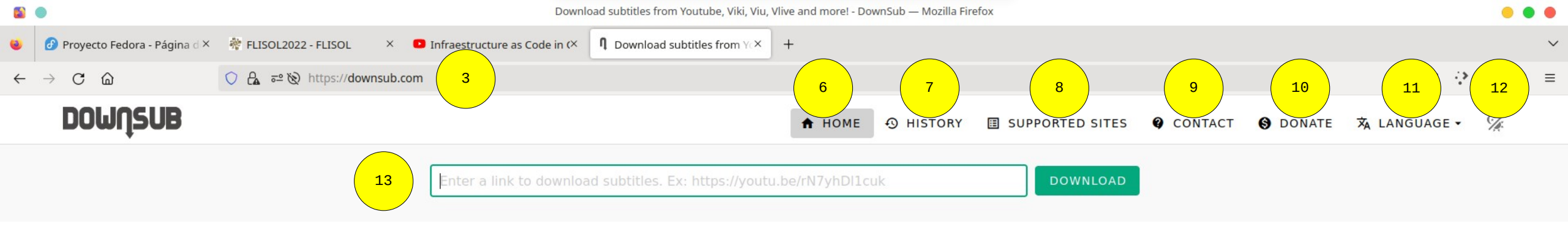
Herramientas automatización

BASH
THE BOURNE-AGAIN SHELL

Shell
Scripting

Codificar automatización descarga



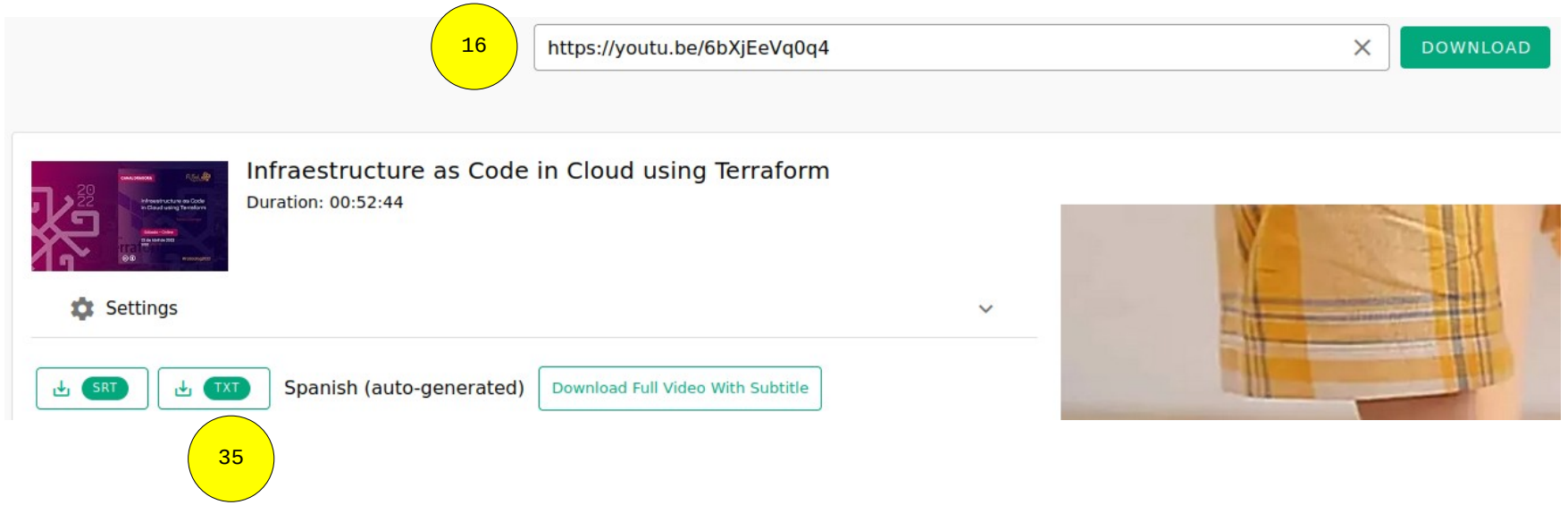


Pegar URLs

flisol.py | $ getsubtitles.sh

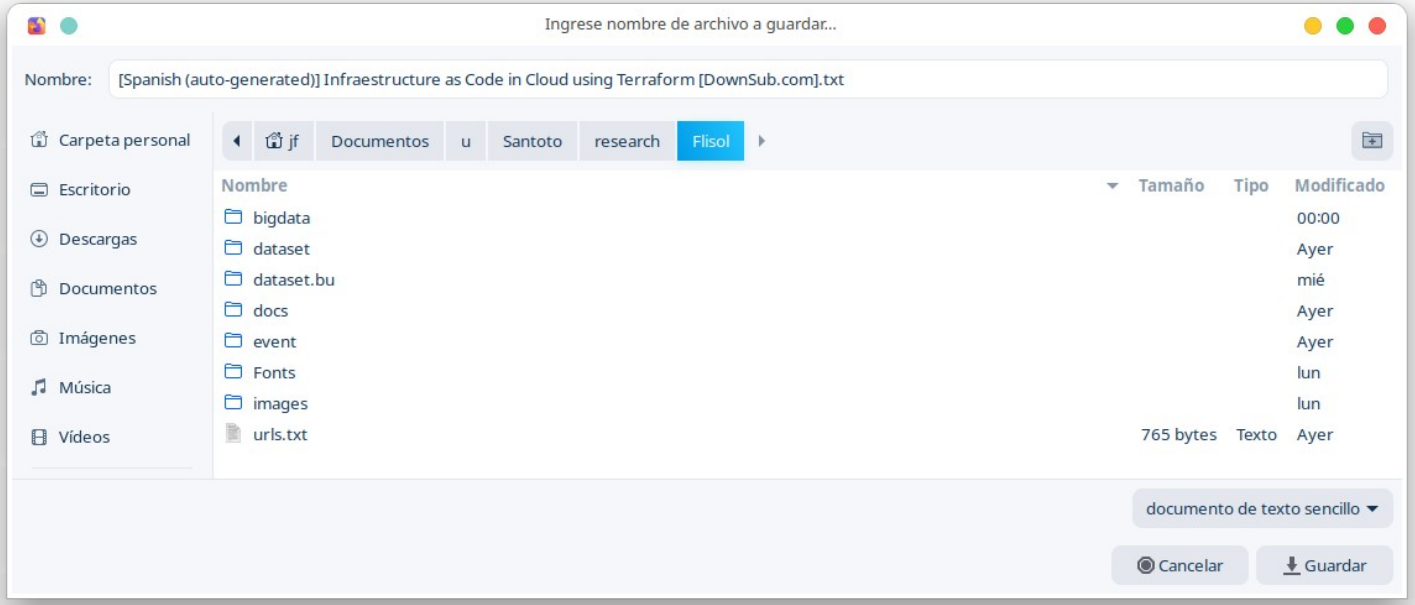
$ getsubtitles.sh

```
15  while read p; do
16      xdotool type "$p"
17      xdotool key KP_Enter
18      sleep 2
19      xdotool key Tab
20      xdotool key Tab
21      xdotool key Tab
22      xdotool key Tab
23      xdotool key Tab
24      xdotool key Tab
25      xdotool key Tab
26      xdotool key Tab
27      xdotool key Tab
28      xdotool key Tab
29      xdotool key Tab
30      xdotool key Tab
31      xdotool key Tab
32      xdotool key Tab
33      xdotool key Tab
34      xdotool key Tab
35      xdotool key KP_Enter
36      sleep 2
```

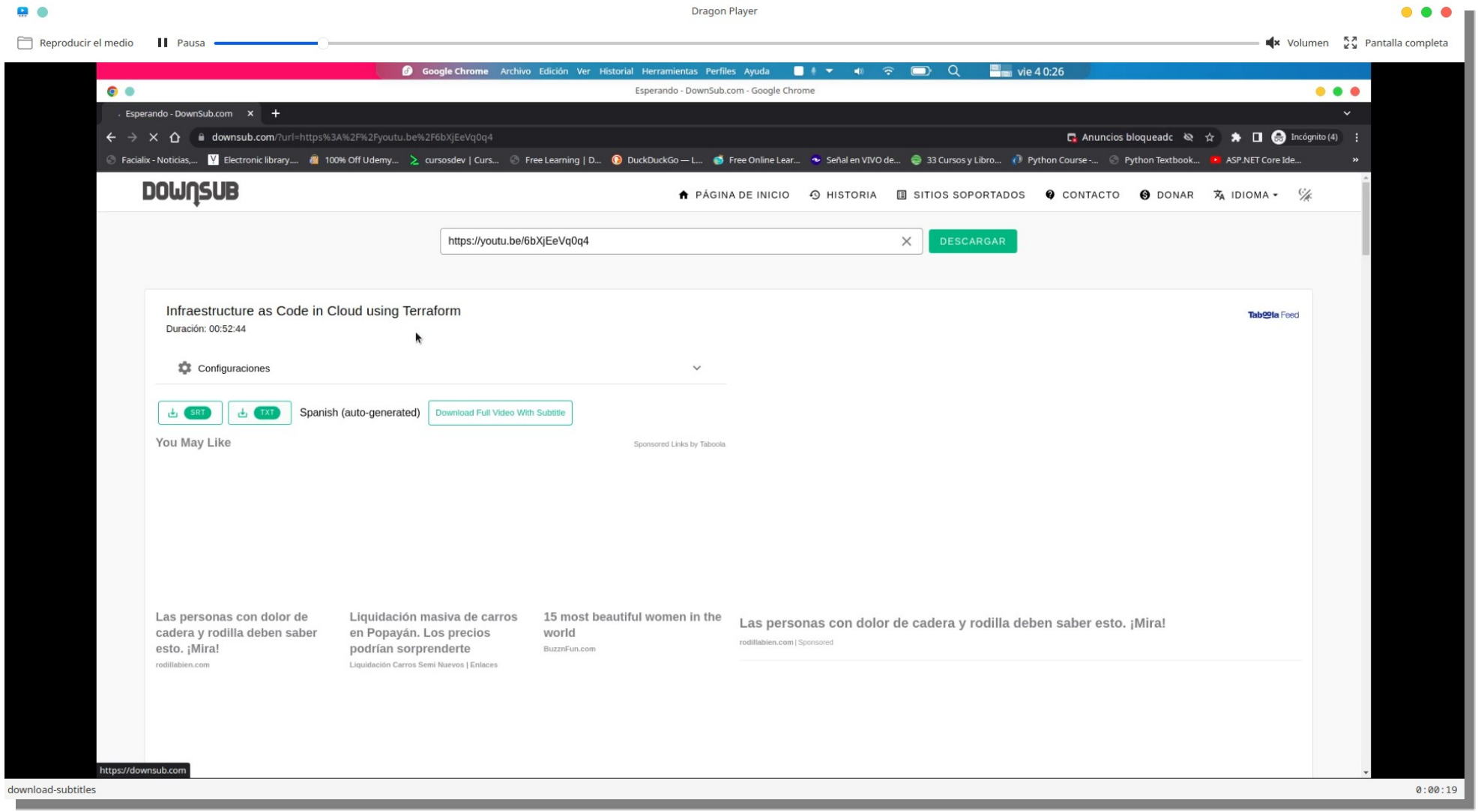


Descargar el archivo de subtítulos

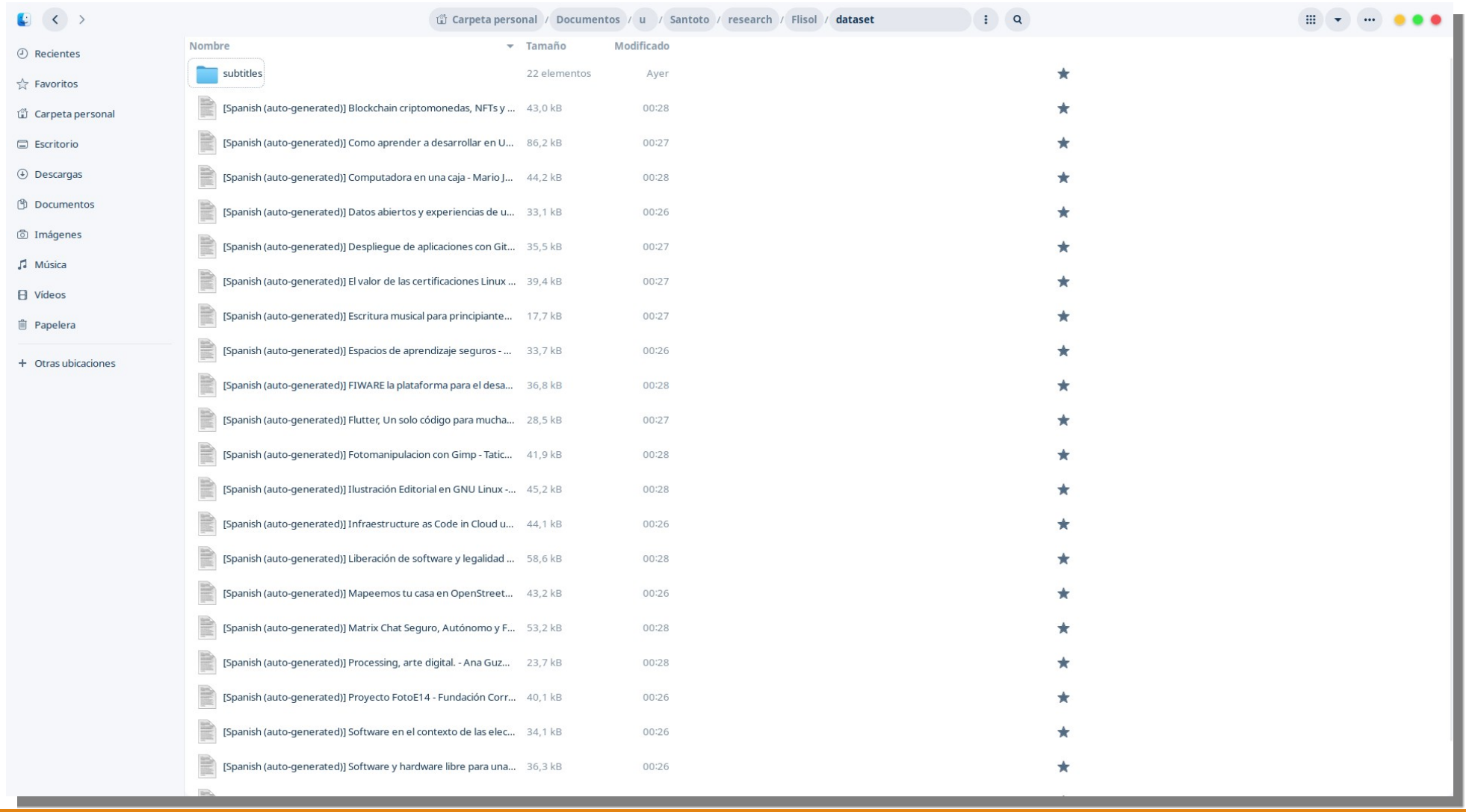
```
$ getsubtitles.sh
37      xdotool key KP_Enter
38      sleep 2
39      xdotool key Shift+Tab
40      xdotool key Shift+Tab
41      xdotool key Shift+Tab
42      xdotool key Shift+Tab
43      xdotool key Shift+Tab
44      xdotool key Shift+Tab
45      xdotool key Shift+Tab
46      xdotool key Shift+Tab
47  done < urls.txt
48  xdotool key Ctrl+w
49
```



Demo Descargas de los subtítulos

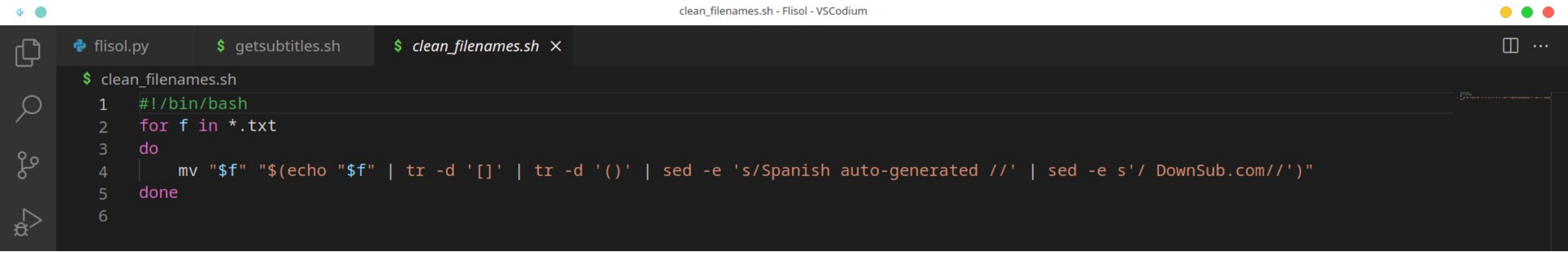


Dataset de subtítulos

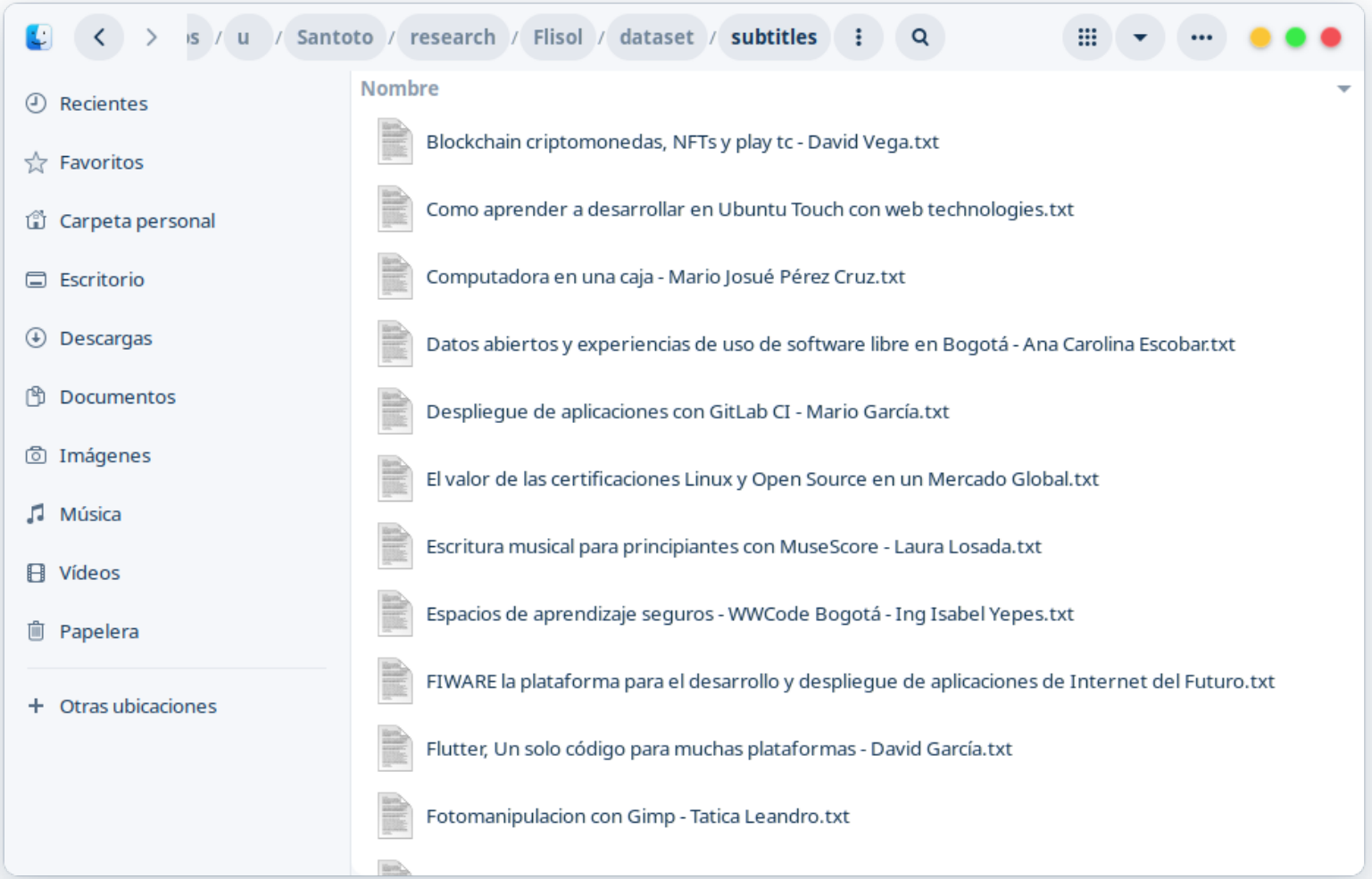


Herramientas reemplazo cadenas caracteres

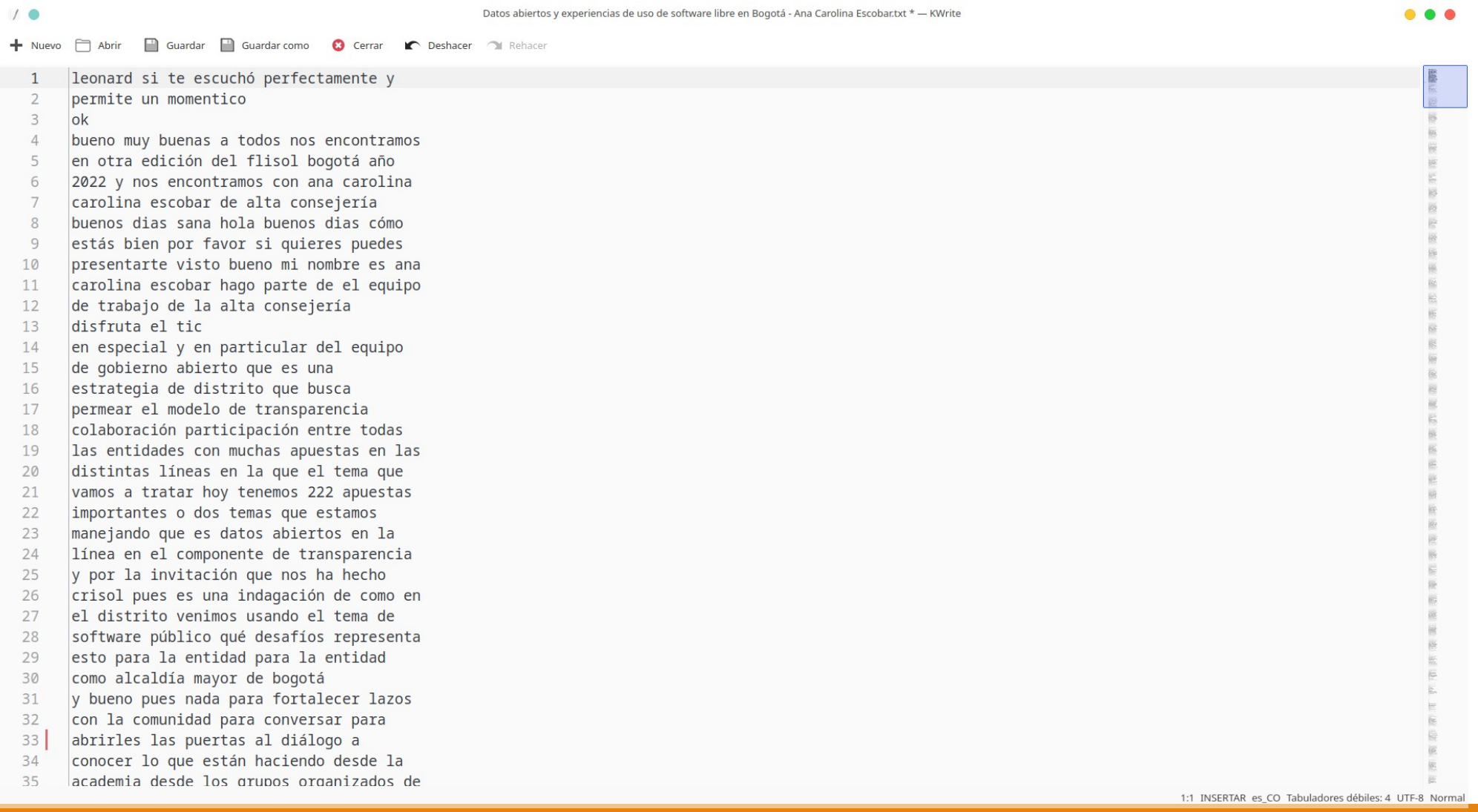
Código depuración de nombres de archivos



Dataset depurado



Ejemplo archivo de subtítulo



Herramientas Populares de Big Data

Este es el inicio…
Pero la historia con Big Data continuará en un próximo episodio...

Juan Francisco Mendoza Moreno

@jfmendozam

jfmendozam

jfmendozamoreno

¿Preguntas?

Gracias

FLISOL 2022, Puro software Libre!

FLISOL |Programa Ingeniería de Sistemas|
FUP | POPAYÁN 2022