

Práctica 1

Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

En virtud de que la red social Instagram concentra grupos de individuos centennials y millennials superando el número de usuarios activos en Twitter, resulta un medio de estudio de datos importante a ser considerado en varios ámbitos como en la política. Es así que, Instagram, con su enfoque en el intercambio de imágenes, funciona para los actores políticos como una herramienta de bajo costo, de fácil difusión, interactiva, fluida y espontánea que facilita la comunicación de mensajes poderosos, personales y potencialmente decisivos en elecciones presidenciales llegando a los jóvenes e impulsando sus objetivos de maximizar los votos en forma de me gusta, comentarios y acciones con hashtags o difusión para cubrir áreas de influencia a su favor.

Con el scrapeo de las cuentas de usuario de Instagram, basado en los hashtags más utilizados para los candidatos presidenciales principales en la segunda vuelta electoral 2021 en Ecuador, podemos obtener sus publicaciones y si las mismas inciden en la actividad de otros usuarios dentro la red social desde cualquier lugar y mayoritariamente desde un dispositivo móvil.

Es de mencionar que, los datos extraídos corresponden a información visible para el público y empleamos la búsqueda de hashtags porque la preferencia a un candidato se difunde más a través de etiquetas clave de lo que un candidato alcanzaría con sus propias publicaciones.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Influencia de elecciones presidenciales Ecuador 2021 en usuarios de Instagram con perfil público.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El dataset contiene datos de cuentas de usuarios con perfil público que han registrado palabras clave semilla (hashtag) en relación a los candidatos a la presidencia en Ecuador durante la campaña de la segunda vuelta de las elecciones 2021-2025.

Los candidatos participantes en la segunda vuelta electoral son: Guillermo Lasso y Andrés Arauz. Entonces, mediante una cuenta privada se consulta los siguientes hashtags:

- Para el candidato Guillermo Lasso

Hashtags: #guillermolasso, #andresnomientasotravez, #lassopresidente2021, #encontrémonosparalograrlo

- Para el candidato Andrés Arauz

Hashtags: #andresarauz, #lassoesmoreno, #arauzpresidente2021, #binomiodelaesperanza

Algunos de nuestros datos podrían estar sesgados por publicaciones que no están destinadas a las elecciones; sin embargo, esto es parte de la limpieza y análisis de datos.

Por cada resultado encontrado, el cual corresponde a un registro del conjunto de datos, se tienen las siguientes características:

- **hashtag**: Expresión clave/palabra semilla empleada por el usuario y mediante el cual se puede medir el alcance que ha logrado su creación y uso en tiempos de campaña.
- **sys_time**: Fecha en que se genera el raspado. Formato ISO 8601.
- **user**: Nombre del perfil o cuenta de quien realiza la publicación. Con este atributo se puede conocer el nivel de evangelización, dicho de otra forma, la cantidad de usuarios que incentiva la campaña a un partido político. Además, podemos medir el alcance de un hashtag tomando en cuenta los usuarios únicos con su número de seguidores que a su vez pueden considerarse influencers.
- **location**: Ubicación añadida en la foto o video publicado pudiendo determinar por región (para Ecuador: Costa, Sierra, Oriente) dónde se encuentra localizada la gente que ha subido publicaciones.

- **image_description:** Metadatos de la foto o video publicado.
- **image_url:** Dirección web del recurso publicado en fotografía.
- **image_type:** Tipo del archivo para diferenciar un video o una fotografía.
- **video_url:** Dirección web del recurso publicado en video.

Dispondremos de las URL de imágenes y videos para posteriormente, almacenarlas de manera local en una carpeta.

- **verified:** Marca de verificación de cuenta confiable. Este atributo confirma la autenticidad, singularidad, integridad y notoriedad de la cuenta.
- **publications:** Es el número de posts o noticias colgadas con la cuenta del usuario. Con este atributo podemos medir el rendimiento de las publicaciones en hashtag durante la campaña, es decir, verificamos si se están creando con éxito seguidores y votos.
- **followers:** Es el número de usuarios que sigue la cuenta. Podríamos decir que, a mayor número de seguidores, más posibilidades de uso de hashtag.
- **followed:** Es el número de usuarios a los que sigue la cuenta.
- **comment:** Opinión o apreciación del dueño de la publicación. Como sabemos, los usuarios pueden pronunciarse de manera positiva, negativa y neutral siendo posible realizar un análisis de sentimientos con su contenido.
- **date_time:** Fecha de publicación en la red social.

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

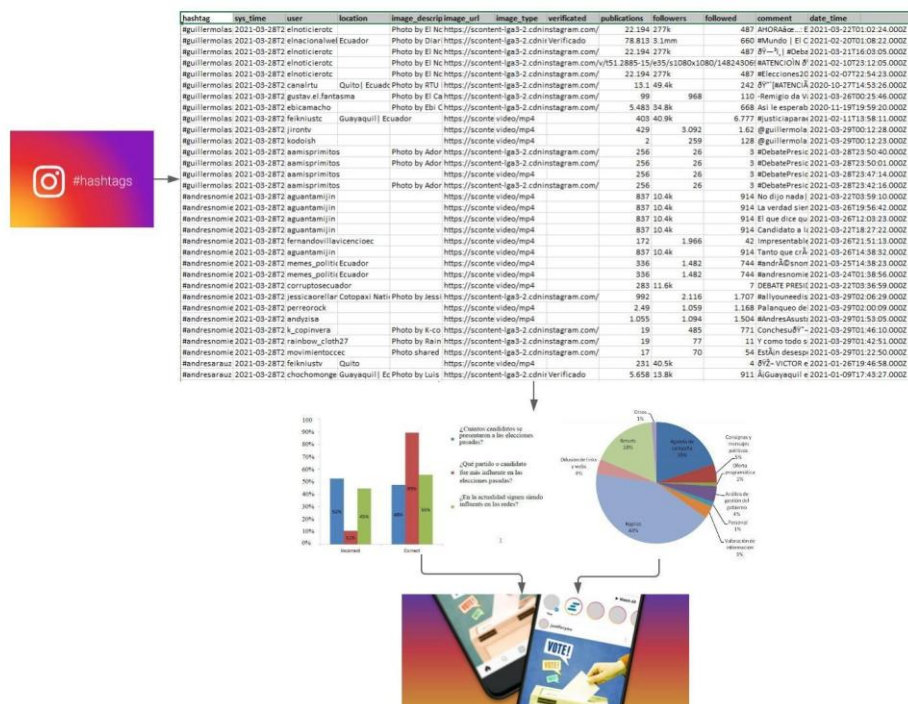


Figura 1. Diagrama del proyecto

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Las variables se encuentran en inglés. Y son:

Campo	Descripción
hashtag	Expresión clave/palabra semilla empleada por el usuario.
sys_time	Fecha en que se genera el raspado. Formato ISO 8601.
user	Nombre del perfil o cuenta de quien realiza la publicación.
location	Ubicación añadida en la foto o video publicado.
image_description	Metadatos de la foto o video publicado.
image_url	Dirección web del recurso publicado en fotografía.
image_type	Tipo del archivo para diferenciar un video o una fotografía.
video_url	Dirección web del recurso publicado en video.
verified	Marca de verificación de cuenta confiable.
publications	Es el número de posts o noticias colgadas con la cuenta del usuario.
followers	Es el número de usuarios que sigue la cuenta.
followed	Es el número de usuarios a los que sigue la cuenta.
comment	Opinión o apreciación del dueño de la publicación.
date_time	Fecha de publicación en la red social.

El periodo de tiempo de captura es desde la semana del 21 al 26 de marzo de 2021 luego de haberse efectuado el gran debate político entre los candidatos Guillermo Lasso y Andrés Arauz.

Para la extracción de datos, se utilizó una cuenta privada y por la actividad inusual que realizamos, intentamos acceder por proxy; sin embargo, como Instagram bloqueó nuestro acceso a su plataforma por las políticas estrictas de redes y no siendo administradores de redes sociales o agencia de marketing, decidimos conectarnos a la red social objetivo a través de una VPN con geolocalización distinta a la actual, para el efecto, establecimos geo Estados Unidos (no Ecuador) para evitar rastreo, bloqueo o verificación de nuestra localización de cuenta Instagram a través del código de seguridad enviado al correo electrónico/teléfono.

Se consultaron los siguientes hashtags o términos para la búsqueda de datos de usuarios:

- Candidato Guillermo Lasso

Hashtags: #guillermolasso, #andresnomientasotravez, #lassopresidente2021, #encontrémonosparalogarlo

- Cara candidato Andrés Arauz

Hashtags: #andresarauz, #lassoesmoreno, #arauzpresidente2021, #binomiodelaesperanza

Recopilamos los datos desde la página web <https://www.instagram.com/?hl=es-la> con la cuenta de usuario *mayuqui5*. Filtramos la búsqueda de cada hashtag para acceder a las cuentas públicas, obtuvimos los datos y se guardaron en un archivo plano consolidado (csv).

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

El propietario de los datos son las cuentas de usuario con perfil público y la aplicación Instagram. Agradecemos a los usuarios que han dado respuesta a diferentes eventos expuestos en la campaña electoral de los candidatos Guillermo Lasso y Andrés Arauz.

Existen análisis similares como el proyecto efectuado en las elecciones de 2009 en Alemania (Tumasjan, Sprenger, Sandner/ & Welpe, 2010) quienes formularon una teoría que muestra que el número de usuarios de Twitter y sus mensajes podían explicar la tendencia de voto. Con un software de análisis de texto que utiliza un diccionario para calcular la categoría de sentimiento a la cual pertenecen las palabras de un texto, computaron 140 mil tweets y fueron comparados con los resultados de las elecciones dando un error en promedio por partido del 1.65%, tomando como posible votación la participación del tráfico de tweets con respecto a algún partido específico, es decir, un tweet mencionando un partido político puede reflejar un voto en potencia.

También encontramos un artículo del año 2016: [Political Data Science: Analyzing Trump, Clinton, and Sanders Tweets and Sentiment](#), el cual comparte el resultado del análisis de texto político (tweets) efectuado en Twitter en relación a las primarias presidenciales de Estados Unidos entre dos partidos para las elecciones de mencionado año. En este análisis se determina si Donald Trump recibe la mayor atención de los medios y si los tweets de Donald Trump son insultantes.

Adscrito al tema están las olas de desinformación electoral transmitidas en redes sociales. Para Barrett P., la red Instagram pudo ser un importante vehículo de desinformación en las elecciones 2020 para Estados Unidos. Permitiendo que sea más práctico avivar discordias y creencias con contenido fácil y económico de producir. Lección aprendida por el sitio web que desde 2016 entró en juego para construir defensas que eviten interferencias de la gente y en lo posible tener elecciones justas que dependan de decisiones basadas en hechos. Por esta razón Instagram, implementó una herramienta de detección automática de comentarios ofensivos como otras redes sociales, siendo así que, en enero 2021, el presidente saliente de Estados Unidos, Donald Trump, fue suspendido en las redes sociales

Twitter, Facebook e Instagram tras publicar mensajes dirigidos a sus partidarios que irrumpieron en el Capitolio en Washington haciendo afirmaciones falsas sobre fraude electoral en las elecciones realizadas el 3 de noviembre 2020.

Otras investigaciones enfocan su trabajo hacia el comportamiento de los usuarios durante las campañas electorales. Por ejemplo, Conover (Conover D., 2010) analizó las elecciones del US del 2010 para detectar polaridad política estudiando el uso de hashtags y la topología de red que forman los retweets entre usuarios. Más adelante utilizó un clasificador (Conover D., 2011) para detectar la polaridad política de los usuarios obteniendo una precisión del 90,8% con hashtags y un 94,9% analizando la red de RTs.

Por lo antes expuesto, de momento no se encuentra un análisis de datos orientado con precisión a campañas electorales en Instagram, red social creada el 6 de octubre de 2010, pero se evidencia que, su uso en marketing político va cada día en ascenso si se compara con otras poderosas redes sociales.

7. **Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.**

Nuestro estudio está basado en el uso de Instagram por parte de los usuarios frente a las elecciones de Ecuador 2021 para conocer el candidato que tiene mayor captación de votos.

Por un lado, es importante para los ecuatorianos conocer el panorama electoral y el posicionamiento ideológico que se propende en el territorio ecuatoriano. Por otro lado, resulta imprescindible para los dirigentes políticos obtener una posición atractiva a la gran masa de votantes registrada en Instagram. En Ecuador, el 92% de los usuarios que utilizan sus teléfonos móviles acceden a las plataformas digitales, pero el acceso no sería constante ya que la gran mayoría de líneas celulares (11,5 millones) son prepago, es decir, no tienen plan de datos fijo y pese a ello 4 millones de perfiles de Instagram ocupa el segundo lugar.

Las preguntas que pretendemos responder son las siguientes:

1. ¿Quiénes son usuarios influyentes (influencers) durante la campaña electoral?
2. Por medio de las menciones de usuarios por candidato o partido dentro de los comentarios publicados podemos obtener una comparativa de inclinación de popularidad del candidato.
3. Conocer cuantitativamente los hashtags más usados en tiempo de elecciones.
4. ¿Cuántas publicaciones promedio de un hashtag se realizan por día?

5. ¿Cuál es el resultado de polaridad política según los comentarios?
6. ¿Cuáles palabras han influido para determinar los sentimientos en campaña electoral?

Adicionalmente, resulta sugerente analizar el contenido audiovisual obtenido para determinar la estrategia electoral, esto es, si son meramente persuasivos al voto con propuestas de gobierno o a acusaciones contrarias al partido u otros tópicos. Como sabemos, una foto o video expresa este mensaje político ante la audiencia.

La fotografía manifiesta propuestas, problemáticas o posturas políticas. En cambio, el video en un tiempo limitado tiene la misión de anunciar un mensaje sencillo y concreto atrayente al público. En ambos casos, esta manera de expresión se ha convertido en medio popular de atención, presentando a los candidatos en campaña para colocar el mensaje en la mente del elector.

8. **Licencia.** Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

Para la publicación de nuestro dataset hemos seleccionado la licencia Creative Commons Reconocimiento-NoComercial-CompartirIgual (CC BY-NC-SA) 4.0 por los siguientes motivos:

- Libertad al compartir/copiar y redistribuir el material en cualquier medio o formato.
- Uso del conjunto de datos sin fines comerciales; sin embargo, obliga a que éste se mantenga en las condiciones en que defina el propietario de los Derechos de Copyright.
- Se otorga el crédito o atribución correspondiente proporcionando un enlace a la licencia e indicando si existen cambios. Puede hacerlo de cualquier manera razonable, pero de ninguna manera sugiere que el licenciante lo respalda a usted o su uso.
- Compartición igual: Si mezcla, transforma o construye sobre este material, debe distribuir sus contribuciones bajo la misma licencia que el original.
- Sin restricciones adicionales: No puede aplicar términos legales o medidas tecnológicas que limiten legalmente a otros de hacer cualquier cosa que permita la licencia.

9. **Código.** Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Consideraciones previas

El código está desarrollado y probado sobre entornos Windows 7 y 10, característica a tener en cuenta previamente a la ejecución del programa para la obtención del conjunto de datos.

El lenguaje del código que ha generado el dataset es Python v3.6, por lo que se necesita tenerlo instalado. El instalador puede descargarse desde [aquí](#).

Luego de su instalación verifique la versión de Python con el siguiente comando en DOS:

```
python --version
Python 3.6.7
```

El navegador web usado para la prueba es Chrome v89.0 y se debe asegurar tenerlo instalado en la máquina. El instalador del navegador puede descargarse desde [aquí](#).

La navegación web automática se realiza a través de la herramienta Selenium y es de incluir como dependencia del código en Python. La instalación de la biblioteca se realiza con el siguiente comando:

```
pip install selenium
```

Se debe descargar el controlador web correcto para el navegador y su versión elegida v89.0 desde [aquí](#). El controlador es un ejecutable que no requiere instalación, pero es de pensar en un directorio donde ubicarlo.

Para desarrollar el código Python utilizamos como editor Visual Studio Code que está disponible [aquí](#).

El acceso a esta red social requiere de un perfil registrado en Facebook o crear una cuenta previamente. Para crear la cuenta es necesario registrar un nombre de usuario, contraseña, móvil o correo electrónico en el sitio oficial <https://www.instagram.com/accounts/emailsignup/>.

En ocasiones es posible descargar la fotografía y video de una manera más eficiente sin entrar al perfil o hacer capturas, ni cosa ilegal; y esto es desde la red oficial de distribución de contenido en el dominio <https://scontent.cdninstagram.com>.

Acerca del desarrollo del código

El código contiene las instrucciones de ingreso, búsqueda, navegación y descarga de lo publicado por usuarios de la red social Instagram a partir de una palabra clave semilla (hashtag). Está disponible en los siguientes repositorios:

<https://github.com/Mayuqui/web-scraping-instagram-presidential-candidates.git>

https://github.com/jayuquina/tipologia_datos_practica1.git

Es indispensable, previamente revisar lo que ofrece la red social, el funcionamiento del sitio web, plataforma, para ver variaciones y limitaciones que tendría que enfrentar el robot codificado, por mencionar a continuación algunas cuestiones:

- La web principal de la red social está en <https://www.instagram.com/?hl=es-la>.
- Admite la función de compartir fotografías, videos y comentarios entre usuarios a través de interacción manual con ayuda del navegador web.
- Según el archivo [robots.txt](#) colgado en el sitio web, están rechazadas todas las operaciones de cualquier agente de internet, exceptuando a los permitidos: Applebot, DuckDuckBot, Googlebot, Yeti, entre otros.
- El sitio no indica mucho de su tecnología. Luego de instalar la herramienta y ejecutar builtwith puede analizarse que utiliza bibliotecas de JavaScript.

```
pip install builtwith  
  
builtwith('http://www.instagram.com')  
  
{'javascript-graphics': ['D3']}
```
- El sitio permite descargar desde la API pública datos del hashtag propuesto a buscar haciendo uso de parámetros de consulta. Para esto no se requiere autenticación ni token y el resultado está en formato estructurado JSON como lo muestra el link https://www.instagram.com/explore/tags/lassoesmoreno/?_a=1

- La plataforma cuenta también con la Instagram Basic Display API e Instagram Graph API para integración de aplicaciones. Tiene métodos de lectura y escritura establecidos por el propietario en formato JSON que requieren de autenticación y autorización de usuario para ejecutarlos. Sin ellos un error HTTP 400 (Bad request) no procesará la solicitud en el servidor.

https://graph.facebook.com/ig_hashtag_search?user_id=17841405309211844&q=coke

```
{
  "error": {
    "message": "A user access token is required to request this resource.",
    "type": "OAuthException",
    "code": 102,
    "fbtrace_id": "AbnJl9x4gRkuMuVKb3OxgLp"
  }
}
```

- Las APIs ofrecidas demuestran tener ciertas ventajas, pero también limitaciones en la exploración de hashtag desde su integración dentro de Facebook Graph API. En ambos casos, el propietario en términos de mejorar la seguridad y privacidad de datos desde 2018 ha cambiado características para los socios, que va desde la reducción del límite de solicitudes por hora, así como eliminar datos que relacionen usuarios y las publicaciones.

En concreto, estas limitaciones reducen efectivamente la capacidad para recopilar cantidades de datos y no permitirá que aplicaciones de terceros, como este robot, permitan ver quién sigue a quién o qué publicaciones les han gustado a los usuarios.

Puesto que esta información se maneja en la página de publicación y el flujo de trabajo no está afectado en la aplicación de Instagram, es determinante acceder a los datos a través de la estructura del mismo sitio web.

El ingreso al perfil de usuario recopila datos de sesión como: dirección IP, ubicación geográfica e idioma, que la plataforma luego usa como variables de configuración y detección en los procesos siguientes de ingreso al sitio.

El tiempo de duración de las peticiones en el navegador es variable según determinados factores. Así, por ejemplo, una conexión al sitio con velocidad de 20Mbps registra un tiempo de descarga de datos de 3 segundos. Además, la página tiene un control de tiempo entre cada interacción con el puntero de 1-2 segundos.

El intento sospechoso de ingreso al perfil es detectado en el inicio de la plataforma. Es posible que se solicite la intervención del usuario con ingreso de un código de seguridad enviado al móvil o correo electrónico.

Para evadir la restricción por IP se ensaya con conexiones a la lista de proxy del generador instalado para Python con el comando respectivo.

```
pip install http-request-randomizer
```

También se realiza otro intento con conexiones a la red del proyecto Tor sobre VPN (mayor información [aquí](#)) para que habilite el proxy <http://127.0.0.1:9125>. Sin embargo, la web de Instagram mantiene activado el control HTTP 429 (Too many requests) sobre la mayoría de direcciones IP proxy públicas encontradas en internet, siendo necesario mantener fuera de foco la ubicación para evitar exponer la IP de navegación real al sitio y luego permanezca bloqueada se opta por utilizar ProtonVPN en la conexión hacia <https://www.instagram.com/?hl=es-la>.

ProtonVPN ofrece el servicio gratuito de VPN y puede descargarse el archivo de instalación [aquí](#) registrando el plan de cuenta [aquí](#).

La descarga de fotografía y video no presenta restricción con IP proxy, por lo que sí es posible utilizar la lista de proxy del paquete de Python o el proxy de Tor + ProtonVPN en la conexión hacia <https://scontent.cdninstagram>.

El fichero `scraper_ins.py` contiene el código del programa para generar el conjunto de datos propuesto desde el sitio web denominado `influencia_elecciones_presidenciales_ec_2021.csv`.

El fichero `scraper_ins_file.py` contiene el código del programa para descargar fotografía y video del conjunto de datos generado previamente con `scraper_ins.py`.

Inicialmente se importa los módulos a necesitar dentro del código, se resalta entre estos el del paquete Selenium preinstalado anteriormente.

```
from selenium import webdriver
```

Se gestiona las propiedades con que el código va a raspar a través de variables, como: usuario, contraseña, idioma, hashtag, deslizamiento y publicaciones.

```
USERNAME = ''
PASSWORD = ''
LANGUAGE = 'spanish'
KEYWORD = ''
NSCROLL = 2
NARTICLES = 3
```

El programa respeta la privacidad de quien lo usa y no recolecta esta información. El valor de estas variables es de ingreso manual mediante definición de una rutina de pantalla principal que contiene campos relacionados a cada propiedad.

```
def main_screen(root):
```

Campo pantalla	Variable	Tipo
Teléfono, usuario o correo electrónico	USERNAME	Texto
Contraseña	PASSWORD	Texto
Idioma del Sistema Operativo	LANGUAGE	Texto
Texto a buscar	KEYWORD	Texto
Número desplazamientos después de buscar	NSCROLL	Número
Número publicaciones o artículos a buscar	NARTICLES	Número

El programa debe apuntar al lugar donde descargó el controlador web. En este ejemplo, se utiliza el controlador de Chrome almacenado en la carpeta de programas.

```
DRIVER_PATH = 'C:/Programas/chromedriver_win32/chromedriver.exe'
```

Dentro de Selenium es necesario definir el navegador web usando la siguiente línea de código.

```
driver = webdriver.Chrome(executable_path=DRIVER_PATH, options=options)
```

A las solicitudes realizadas, el navegador añade la cabecera user-agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.90 Safari/537.36 para que el sitio revise. Así mismo, se deshabilita la propiedad de software automático en Chrome con esta línea.

```
options = webdriver.ChromeOptions()
options.add_experimental_option('excludeSwitches', ['enable-
automation'])
options.add_experimental_option('useAutomationExtension', False)
```

A continuación, se quiere solicitar la URL de la página principal del sitio en la ventana del navegador Chrome con este código pasando como parámetro el idioma.

```
if LANGUAGE=='english' :
    driver.get("https://www.instagram.com/?hl=en-us")
else:
    driver.get("https://www.instagram.com/?hl=es-la")
```

La ubicación de los datos en el sitio web se realiza con Selenium y el método XPath para ingresar, extraer y guardarlos. Se define dos funciones específicas para encontrar el dato como elemento o atributo de la referencia que le contiene.

```
def find_text_by_xpath(xpath, field):
def find_attribute_by_xpath(xpath, attribute, field):
```

Es posible que la plataforma realice cambios en el sitio web y probablemente el código necesite revisiones o ya no funcione. Por eso, a través de excepciones en las funciones se advierte en la ejecución de errores en la localización de los datos en los elementos HTML, por ejemplo.

```
location <class 'selenium.common.exceptions.NoSuchElementException'>
image_description <class 'selenium.common.exceptions.NoSuchElementException'>
image_url <class 'selenium.common.exceptions.NoSuchElementException'>
```

Hay un diccionario con etiquetas por idioma que se configure al inicio del código: inglés o español. Pues los elementos HTML en su mayoría están sin atributo id y class variable, lo que implica buscarlos por estructura y etiquetas. Las etiquetas varían según el idioma determinado por el navegador.

```
dictionary = (('Not Now', 'Ahora no'), ('Search', 'Busca'),
('Verified', 'Verificado'), ('publications', 'publicaciones'),
('followers', 'seguidores'), ('followed', 'seguidos'))
```

El resultado de las consultas se muestra progresivamente desplazando hacia abajo la página. Esta interacción se ejecuta con JavaScript en la ventana mostrada.

```
n_scrolls = NSCROLL
for j in range(0, n_scrolls):
```

```
driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
time.sleep(5)
```

El tiempo de espera entre una solicitud y la extracción de datos se define con esta instrucción y un valor en rango. Como puede observar, se limita la velocidad de scrapeo con una pausa entre 5 y 10 segundos.

```
time.sleep(5)
```

Los datos ocultos en el sitio que aparecen cuando se coloca el puntero sobre el nombre del usuario se ejecuta mediante el siguiente JavaScript que dispara el evento.

```
move_to_user_el_script = """if(document.createEvent){
var event=document.createEvent('MouseEvents');
event.initMouseEvent('mouseover', true, false);
arguments[0].dispatchEvent(event);
} else if(document.createEventObject){
arguments[0].fireEvent('onmouseover');}"""

driver.execute_script(move_to_user_el_script, user_el)
```

Los datos extraídos de la página son recuperados en la clase Article con sus respectivos atributos. Tiene el método que pone los atributos en una línea separados por “,” y guarda en un archivo csv.

```
class Article:
    def __init__(self, user, location, image_description, image_url, image_type, video_url, verificated, publications, followers, followed, comment, date_time):
        self.user = user
        self.location = location
        self.image_description = image_description
        self.image_url = image_url
        self.image_type = image_type
        self.video_url = video_url
        self.verificated = verificated
        self.publications = publications
        self.followers = followers
        self.followed = followed
        self.comment = comment
        self.date_time = date_time

    #format to csv text
    def format_to_csv_text(self, text, sep=';'):
```

```

        format_text = text
        format_text = format_text.replace(sep, '|')
        format_text = format_text.replace("'", '"')
        format_text = format_text.replace('\n', '\\n')
        format_text = format_text.replace('\r', '\\r')
        format_text = "'" + format_text + "'"
        return format_text

#save data as csv
def save_as_txt(self, name):
    sep = ';'
    sys_time = datetime.datetime.now().isoformat()
    self.location = self.format_to_csv_text(self.location, sep)
    self.image_description = self.format_to_csv_text(self.image_desc
ription, sep)
    self.publications = self.format_to_csv_text(self.publications, s
ep)
    self.followers = self.format_to_csv_text(self.followers, sep)
    self.followed = self.format_to_csv_text(self.followed, sep)
    self.comment = self.format_to_csv_text(self.comment, sep)
    line = name+sep+sys_time+sep+self.user+sep+self.location+sep+self
f.image_description+sep+self.image_url+sep+self.image_type+sep+self.vide
o_url+sep+self.verificated+sep+self.publications+sep+self.followers+sep+
self.followed+sep+self.comment+sep+self.date_time+'\n'
    fname = 'influencia_elecciones_presidenciales_ec_2021.csv'
    empty = os.path.exists(fname)==False or os.stat(fname).st_size==
0

    fwrite = open(fname, 'a+', encoding='utf-8')

    if empty:
        header = 'hashtag'+sep+'sys_time'+sep+'user'+sep+'location'+
sep+'image_description'+sep+'image_url'+sep+'image_type'+sep+'video_url'
+sep+'verificated'+sep+'publications'+sep+'followers'+sep+'followed'+sep
+'comment'+sep+'date_time'+'\n'
        fwrite.write(header)

    fwrite.write(line)
    fwrite.close()

```

Para la solicitud de descarga de fotografía e imagen se importa los paquetes de las bibliotecas `http_request_randomizer` y `requests`.

```

from http_request_randomizer.requests.proxy.requestProxy import RequestP
roxy
import requests

```


La propiedad de conexión presenta una alternativa de funcionamiento con el servidor Tor en proxy <http://127.0.0.1:9125> y otra con la lista de servidores de la biblioteca de Python. Por defecto el robot está configurado para conectarse mediante la red Tor debido a que muestra mejor eficiencia en funcionamiento.

```
USE_TOR = True

if USE_TOR==False:
    request_proxy = RequestProxy()
    proxy_list = request_proxy.get_proxy_list()
```

Se espacia las peticiones HTTP con intervalo de 5 segundos y tiempo de espera hasta 10 segundos para no saturar el servidor de contenido.

```
time.sleep(5)
response = requests.get(url=url, stream=True, proxies=proxy_temp, timeout=10)
```

Los recursos extraídos de la página son recuperados con el método que pone los ficheros en el directorio `images`.

```
def download_and_save_file(url):
```

El control de errores temporales es manejado a través de excepciones. Un ejemplo de error temporal que puede producirse es el código 403 Forbidden, tras el que el programa reintenta otra solicitud cambiando la IP del proxy.

```
def download_and_save_file(url):
    ret = False
    address = '0.0.0.0:0'

    print(url)

    try:
        proxy_temp = None

        if USE_TOR==False:
            proxy = None

            if proxy_pos>=0:
                proxy = proxy_list[proxy_pos]

            address = proxy.get_address()
```

```

        proxy_temp = {
            "http": address,
            "https": address
        }
    else:
        address = 'socks5://localhost:9150'

        proxy_temp = {
            "http": address
        }

    response = requests.get(url=url, stream=True, proxies=proxy_temp
, timeout=10)

    print('response.status_code', response.status_code)

    if str(response.status_code)=='200':
        base_dir = os.path.dirname(__file__)
        file_name = os.path.basename(url)
        file_name = urlparse(file_name).path
        file_dir = os.path.join(base_dir, 'images')

        print('file_dir=', file_dir)

        os.makedirs(file_dir, exist_ok=True)

        file_path = os.path.join(file_dir, file_name)

        print('file_path=', file_path)

        fwrite = open(file_path, 'wb')

        for chunk in response.iter_content():
            fwrite.write(chunk)

        fwrite.close()

        ret = True
    else:
        if str(response.status_code)=='403':
            ret = True
except:
    e = sys.exc_info()[0]
    print('Error download_and_save_file address', address, str(e))
return ret

```

Finalmente se ejecuta el script desde la línea de comandos o desde Visual Studio Code.

```
python "scraper_ins.py"
python "scraper_ins_file.py"
```

Guía de funcionamiento

Paso 1. Se aconseja al usuario revisar las [condiciones](#) del servicio de Instagram sobre su política de datos y plataforma para asegurar el buen uso, así como, la regulación aplicable para responder en caso de infracciones. El usuario es consciente que al proporcionar su credencial concede acceso a los programas para realizar funciones de scrapeo en el sitio web.

Paso 2. Abrir el proveedor de servicio de red privada ProtonVPN. Esa aplicación debe pedir el usuario y contraseña de la cuenta en <http://protonvpn.com> para encontrar la lista de países y servidores VPN. Luego dar clic en “Quick Connect” para conectarse al servidor más cercano.



Figura 2. Aplicación de ProtonVPN

Paso 3. Abrir la pantalla principal del programa ejecutando la línea de comando `scraper_ins.py`. Esto muestra la interfaz de la imagen siguiente con los campos de entrada para hacer manejable el uso de la aplicación.

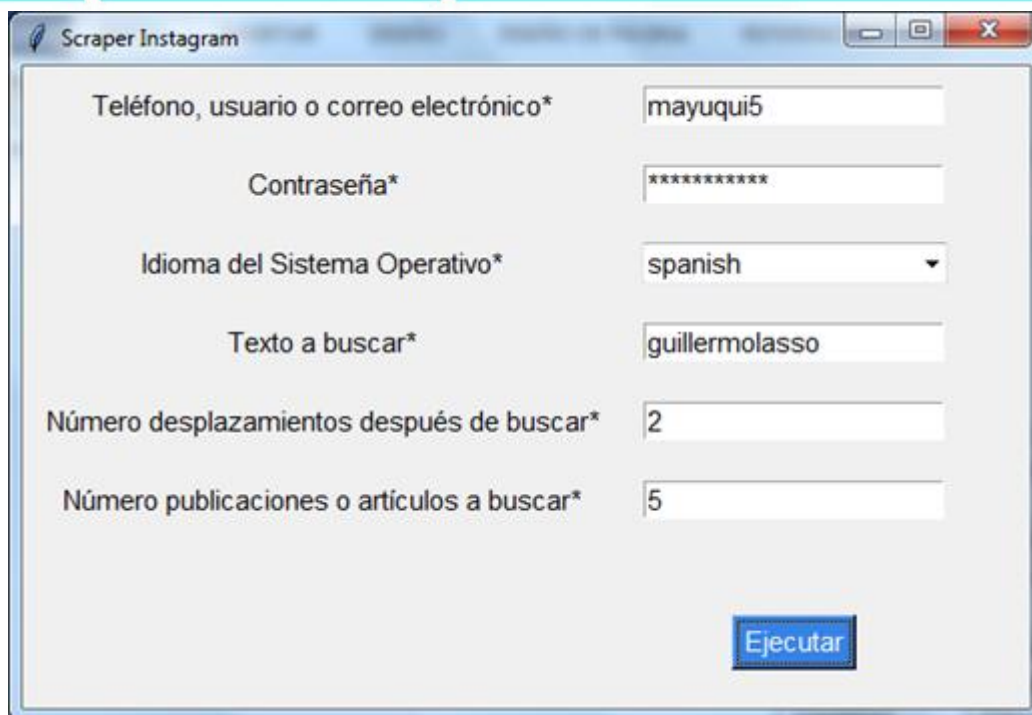



Figura 3. Pantalla Input para raspado

Paso 4. En la pantalla se debe ingresar todos los campos necesarios para efectuar el scrapeo los cuales son los siguientes:

- **Teléfono, usuario o correo electrónico:** Nombre del perfil de usuario, correo o móvil registrado en el sitio web.
- **Contraseña:** Palabra secreta o clave de ingreso al perfil.
- **Idioma del Sistema Operativo:** Es el idioma del Sistema Operativo normalmente el mismo especificado por defecto en Chrome. Puede ser español o inglés.
- **Texto a buscar:** Expresión clave/palabra semilla a consultar en el sitio web, preferiblemente en minúscula.
- **Número desplazamientos después de buscar:** El número de deslizamientos en la ventana para aumentar el resultado de la consulta.
- **Número publicaciones o artículos a buscar:** El número de artículos a recorrer de los que se espera recopilar datos.

Paso 5. Dar clic en el botón  luego de tener los campos llenos, caso contrario aparecerá el mensaje siguiente.

(*) Los campos marcados son obligatorios

Paso 6. Esperar que el proceso concluya sin salir de la pantalla. Mientras en segundo plano el programa debe efectuar la rutina a continuación sino se presentan errores.

Abrir automáticamente el navegador Chrome en el sitio web de la red <https://www.instagram.com>.

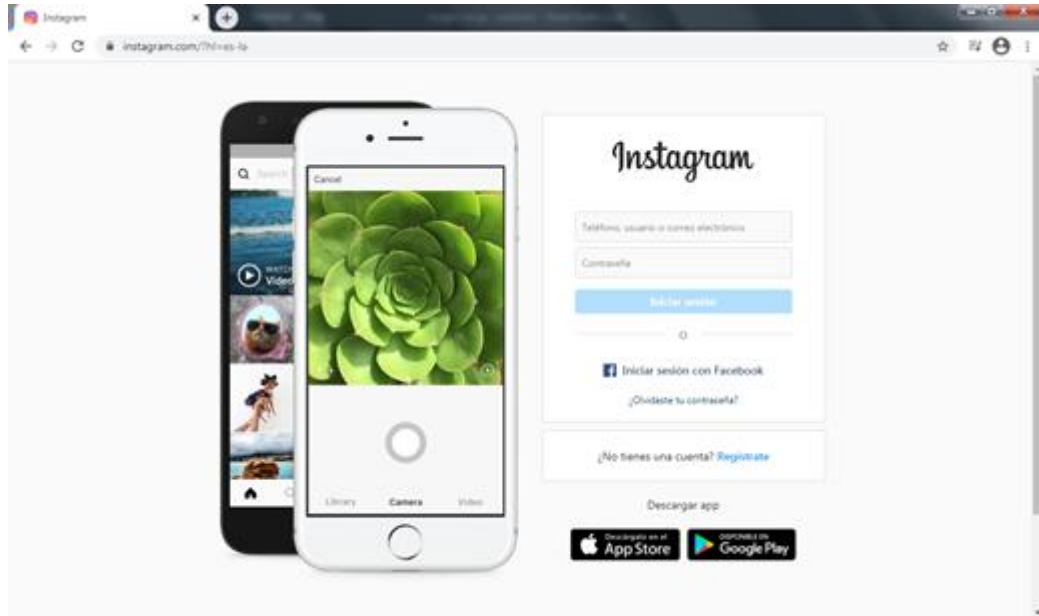


Figura 4. Sitio web Instagram

Pasar los datos de la pantalla principal Teléfono, usuario o correo electrónico y la Contraseña. El programa debe dar clic al botón “Log In”.



Figura 5. Ingreso del programa a una cuenta Instagram

El sitio ha de hacer login y posteriormente confirmar la pregunta ¿Guardar tu información de inicio de sesión? El programa debe dar clic al botón “Ahora no”.

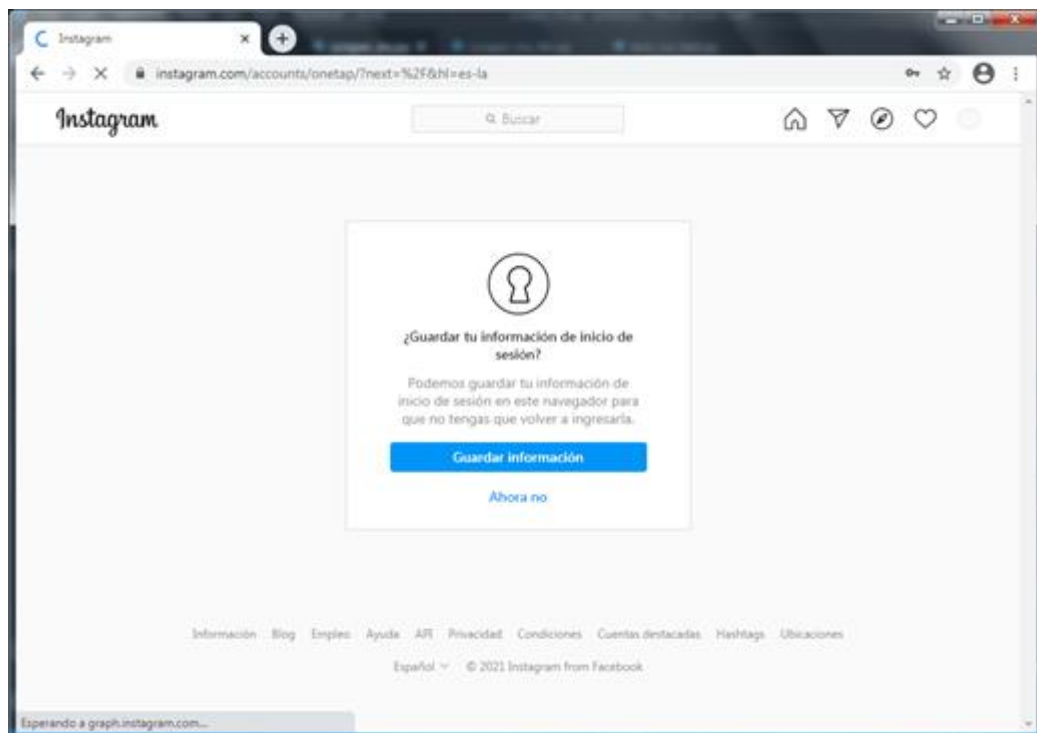


Figura 6. Confirmación de pregunta 1

El sitio nuevamente confirmará la pregunta ¿Activar notificaciones? El programa debe dar clic al botón “Ahora no”.

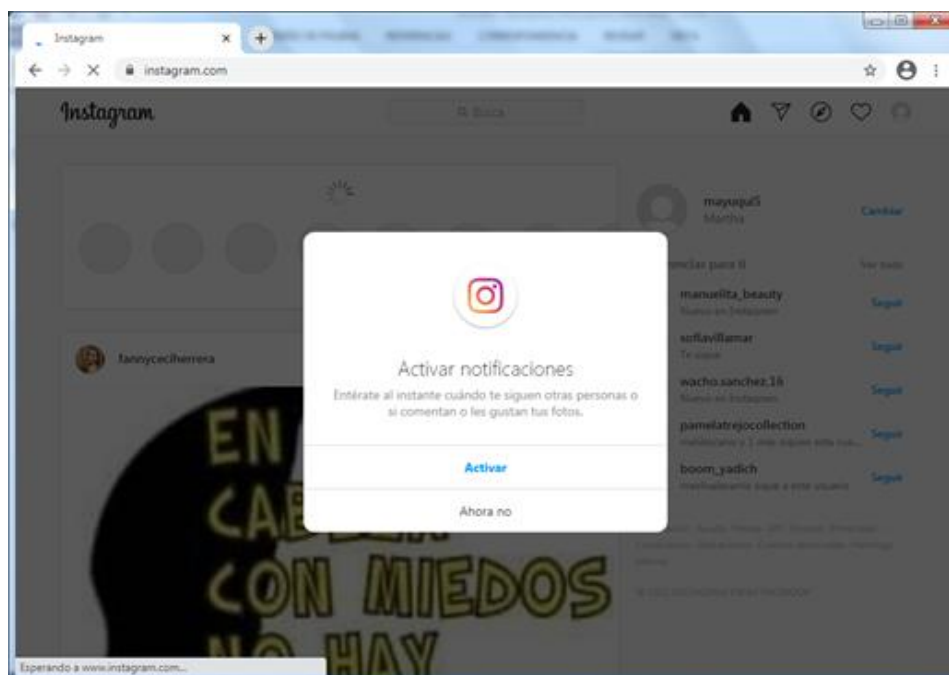


Figura 7. Confirmación de pregunta 2

Debe pasar los datos de la pantalla principal Idioma del Sistema Operativo, Texto a buscar, Número desplazamientos después de buscar y Número publicaciones o artículos a buscar.

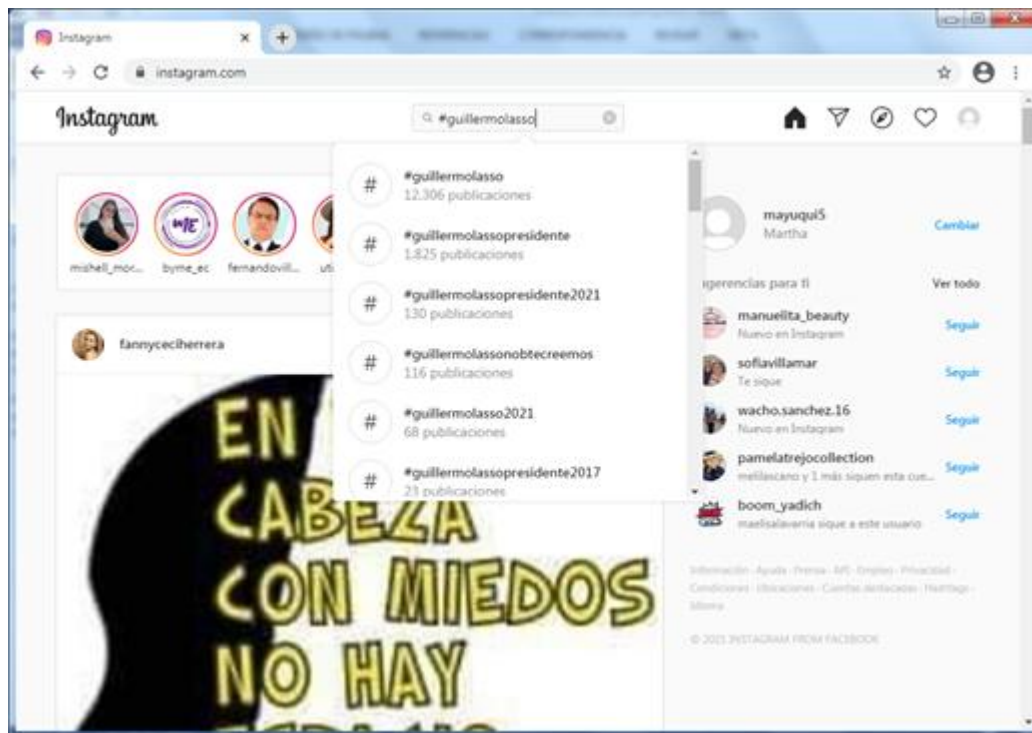


Figura 8. Pantalla principal de una cuneta Instagram

Recopilar y añadir los datos del resultado dentro del fichero `influencia_elecciones_presidenciales_ec_2021.csv`. Sino existe, el archivo se creará en el mismo directorio donde está ubicado el programa.

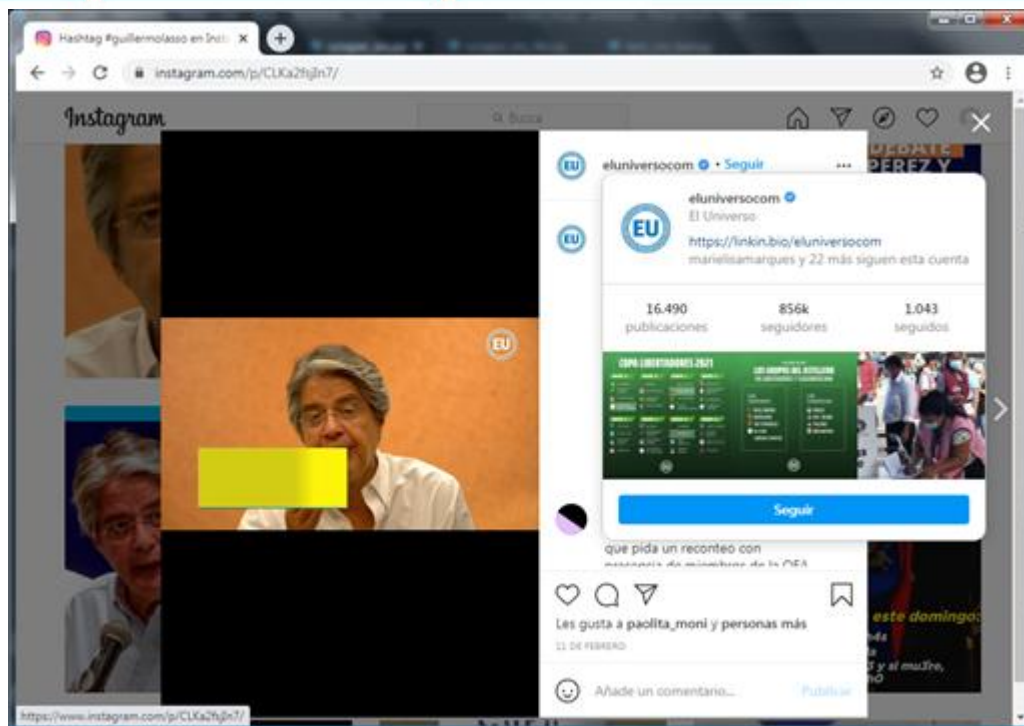



Figura 9. Resultado de la búsqueda por hashtag

 influencia_elecciones_presidenciales_ec_2021.csv

Archivo de valores...

Paso 7. Abrir el navegador Tor para la carga de estado de red y habilitación del proxy local <http://127.0.0.1:9125>.

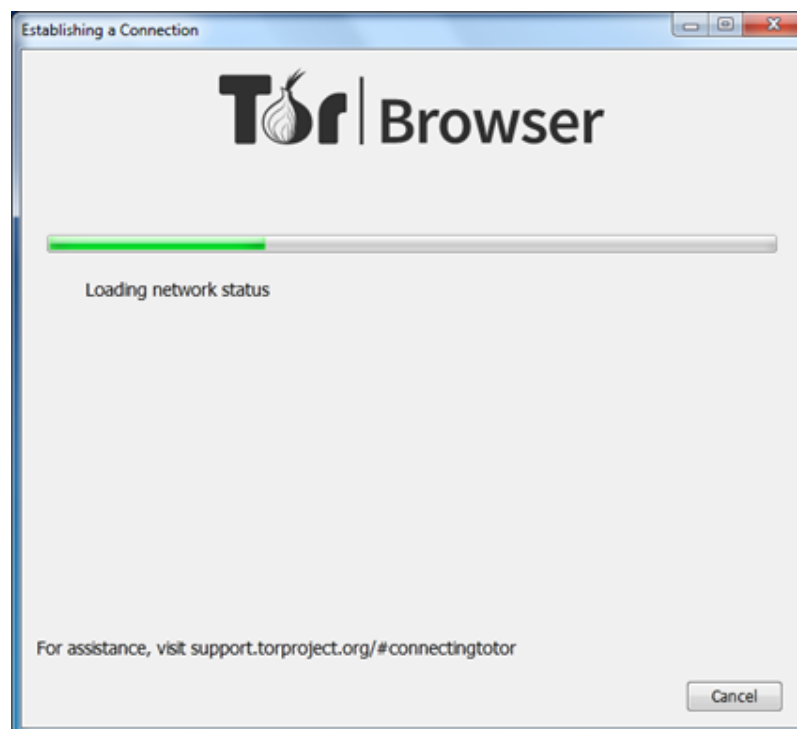


Figura 10. Aplicación Tor

Paso 8. A partir del fichero generado `influencia_elecciones_presidenciales_ec_2021.csv` se ejecuta la línea de comando `scraper_ins_file.py`.

`influencia_elecciones_presidenciales_ec_2021.csv` Archivo de valores...

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	hashtag	sys_time	user	location	image_descr	image_url	image_type	video_url	verificated	publications	followers	followed	comment	date_time
2	#guillermola	2021-04-10T1	santiagolasso			https://scon	video/mp4	https://scon	No verificad	817	13,1k		1.434	¿Por quÃ© 2021-02-03T12
3	#guillermola	2021-04-10T1	eluniversocom			https://scon	video/mp4	https://scon	Verificado	16.491	856k		1.043	Lasso reaccic 2021-02-11T18
4	#guillermola	2021-04-10T1	canalrtu	Quito, Ecuad	Photo by RTU	https://scon	jpg		No verificad	13.303	50,2k		242	ðŸ™” [HATENC 2020-10-27T14
5	#guillermola	2021-04-10T1	eluniversocom			https://scon	video/mp4	https://scon	Verificado	16.491	856k		1.043	CNE decidiÃ© 2021-02-13T15
6	#guillermola	2021-04-10T1	ebicamacho		Photo by Ebi	https://scon	jpg		No verificad	5.563	35k		672	Hasta el Voci 2020-09-11T13
7	#guillermola	2021-04-10T1	eluniversocc	Ecuador		https://scon	video/mp4	blob:https://	Verificado	16.491	856k		1.043	Desde las 11 2021-02-12T20
8	#guillermola	2021-04-10T1	ecuador_tierrita_mia		Photo by Ecu	https://scon	jpg		No verificad	282	3.436		77	PensarÃ©n bi 2021-02-04T03
9	#guillermola	2021-04-10T1	elnoticierotc		Photo by El n	https://scon	jpg		No verificad	22.285	280k		487	#ATENCIOIN 2021-02-12T16
10	#guillermola	2021-04-10T1	eluniversocom			https://scon	video/mp4	https://scon	Verificado	16.491	856k		1.043	La disputa er 2021-02-19T01
11	#guillermola	2021-04-10T1	eldiarioec		Photo by El l	https://scon	jpg		No verificad	6.198	78,8k		43	#ACTUALIDA 2021-02-22T17
12	#guillermola	2021-04-10T1	carlo_celi		Photo by Car	https://scon	jpg		No verificad	573	2.971		1.031	Estrategias e 2020-11-11T15
13	#guillermola	2021-04-10T1	ecuadormemezonico		Photo by Ecu	https://scon	jpg		No verificad	498	7.247		94	RESULTADOS 2021-02-07T22

Figura 11. Dataset

En referencia a la calidad y robustez del contenido del archivo, se observa:

- ✓ **Exactitud:** al filtrar por usuario, en distintas publicaciones de un mismo intervalo de tiempo, el número de seguidores y seguidos es igual.
- ✓ **Compleitud:** 100% de usuarios tienen datos obligatorios completos, incluso si sus datos opcionales están en blanco.

La proporción de datos completos en las columnas del fichero frente al potencial 100% se resume en el siguiente estadístico.

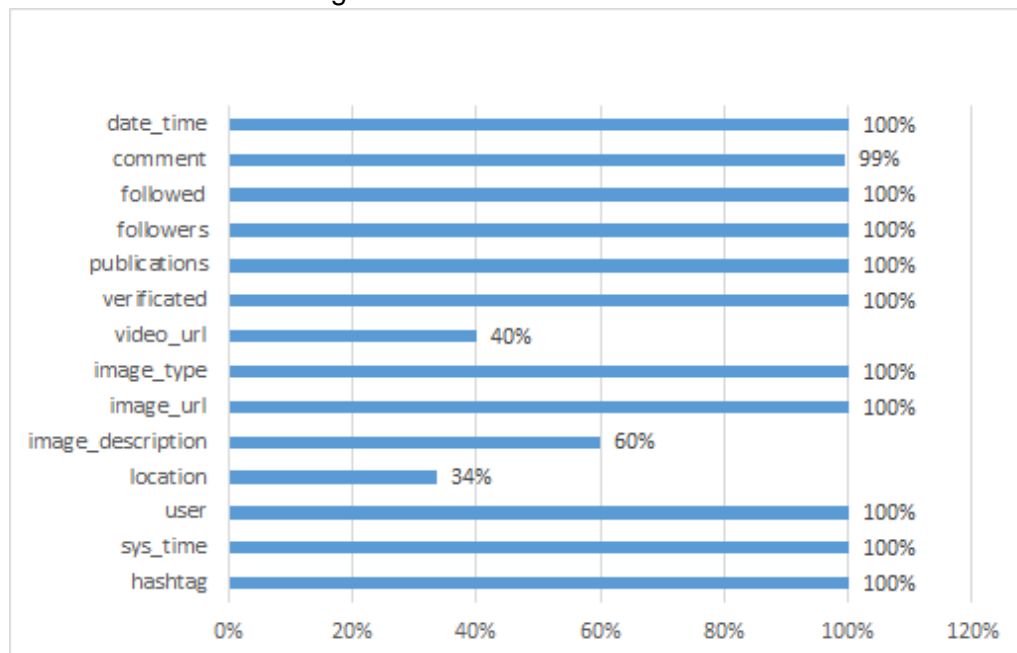


Figura 12. Resultado de Compleitud de datos

- ✓ **Consistencia:** El valor del atributo hashtag por cada registro se encuentra en el contenido del atributo comment.
La columna image_type diferencia el recurso publicado de una fotografía o video. Cuando parte la información sobre la columna image_url está disponible, este campo tiene el valor “jpg”. Cuando la información de la columna video_url está disponible, este campo tiene valor “video/mp4”.
- ✓ **Puntualidad:** Los datos corresponden a publicaciones realizadas en el año actual 2021.
- ✓ **Unicidad:** Las URLs de imágenes y videos son únicas, no se repiten por hashtag.
- ✓ **Validez:** Los atributos obligatorios siguen un patrón. Ejemplo, el atributo sys_time tiene formato ISO 8601

Paso 9. Esperar que el proceso siga su curso en segundo plano. Si no presenta errores, el proceso debe descargar la fotografía y video según admita la red de contenidos <https://scontent.cdninstagram> dentro del directorio images.





Nombre	Tipo	Tamaño
 145495809_2743592632524092_4508328467877522432_n.jpg	Imagen JPEG	43 KB
 149424599_3688759051159654_5111831991726094181_n.jpg	Imagen JPEG	41 KB
 150148804_698270880855069_2466646696111543484_n.jpg	Imagen JPEG	180 KB
 10000000_789497051660284_8450755311660379212_n.mp4	Vídeo MP4	13,077 KB
 10000000_1155544728220957_3699550065625454012_n.mp4	Vídeo MP4	328 KB
 148983140_493942915118869_7557984281308869592_n.mp4	Vídeo MP4	6,656 KB

Figura 13. Fotos y videos descargadas

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Influencia_elecciones_presidenciales_ec_2021.csv			
Influencia de elecciones presidenciales Ecuador 2021 en usuarios de Instagram con perfil público.			
Columna (inglés)	Columna (español)	Descripción	Ejemplo
hashtag	Etiqueta	Expresión clave usada por el usuario. Ej.:	#guillermolasso
sys_time	Fecha_raspado	Es la fecha en que se genera el raspado en formato ISO 8601.	2021-03-28T21:09:24.293887
user	Usuario	Nombre del perfil o cuenta que hace la publicación.	elnacionalweb

location	Ubicación	Ubicación añadida en la foto o video.	Ecuador
image_description	Imagen_descripcion	Metadatos de la foto o video.	Photo by Diario El Nacional in Ecuador.
image_url	Imagen_url	Dirección web del recurso publicado.	https://scontent-lga3-2.cdninstagram.com/v/t51.2885-15/e35/151829529_838814143342197_1230412458173389856_n.jpg?tp=1&nc_ht=scontent-lga3-2.cdninstagram.com&nc_cat=109&nc_ohc=VWtUou2E4MEAX9Vunip&ccb=7-4&oh=bcb432bb0c279ebc098007fa556e9fa6&oe=608A6298&nc_sid=4f375e
image_type	Imagen_tipo	Tipo del archivo para diferenciar un video o una foto.	jpg
video_url	Video_url	Dirección web del recurso publicado.	--
verified	Verificación	Marca de verificación de cuenta confiable.	Verificado
publications	Publicaciones	Es el número de posts o noticias colgadas con la cuenta.	78.813
followers	Seguidores	Es el número de usuarios que sigue la cuenta.	3,1mm
followed	Seguidos	Es el número de usuarios a los que sigue la cuenta.	660
comment	Comentario	La opinión o apreciación del dueño de la publicación.	#Mundo El Consejo Nacional Electoral (CNE) de Ecuador anunció este viernes que será el candidato conservador por la coalición Creo-psc Guillermo Lasso quien competirá con Andrés Arauz de Unión por la Esperanza (UNES)

date_time	Fecha_hora	Es la fecha en que se realizó la publicación en formato ISO 8601.	2021-02-20T01:08:22.000Z
-----------	------------	---	--------------------------

Puede descargar el dataset desde el enlace (obtención del DOI), adjuntamos el ID y la URL para que pueda descargar el conjunto de datos acerca de las publicaciones en Instagram por campaña electoral en las elecciones presidenciales de Ecuador en 2021.

DOI

10.5281/zenodo.4679841

Target URL

<https://doi.org/10.5281/zenodo.4679841>

Tabla de contribuciones al trabajo

Contribuciones	Firma
Investigación previa	J.A., M.A
Redacción de las respuestas	J.A., M.A
Desarrollo código	J.A., M.A

Recursos

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Zambrano, R. (2020) .Hasta en TikTok los políticos buscarán los votos en Ecuador; COVID-19 cambia la estrategia electoral para elecciones de 2021.

<https://www.eluniverso.com/noticias/2020/05/24/nota/7849353/elecciones-presidenciales-2021-ecuador-redes-sociales/>

- Dugué, C. (2018). Predicting the number of likes on Instagram.

<https://towardsdatascience.com/predict-the-number-of-likes-on-instagram-a7ec5c020203>

- Barret P. (2020). Why Instagram could be a major site for disinformation in the 2020 US election.

<https://www.theguardian.com/commentisfree/2019/sep/12/why-instagram-could-be-a-major-site-for-disinformation-in-the-2020-election>

- Quentin Simms, ParseHub (2016). Political Data Science: Analyzing Trump, Clinton, and Sanders Tweets and Sentiment

<https://www.theguardian.com/commentisfree/2019/sep/12/why-instagram-could-be-a-major-site-for-disinformation-in-the-2020-election>

- Carrillo A., Urueña J., Forero J., Caicedo L. (2015). Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales.

<https://repository.javeriana.edu.co/bitstream/handle/10554/20516/CaicedoOrtizLuisEduardo2016.pdf?sequence=1&isAllowed=y>

- Conover M., Ratkiewicz J., Francisco M., Gonçalves B., Flammini A., Menczer F. (2010). Political Polarization on Twitter

http://www.cse.cuhk.edu.hk/~cslui/CMSC5734/Conover_PoliticalPolarizationTwitter.pdf