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.

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

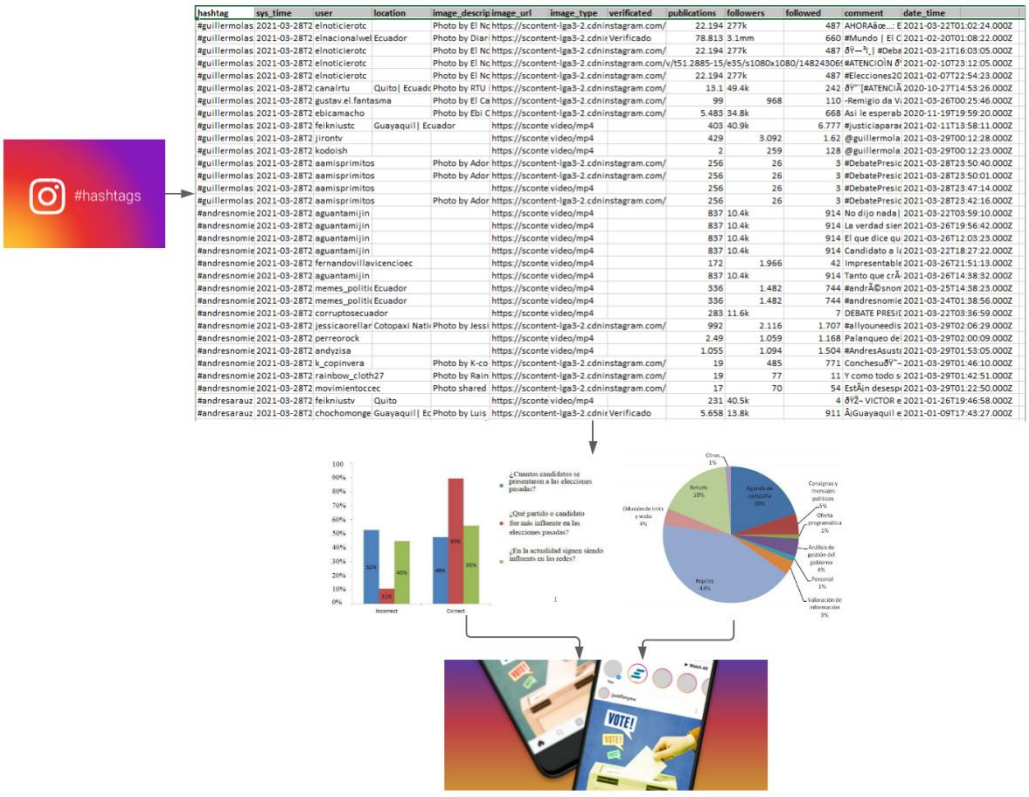
   Influencia de elecciones presidenciales Ecuador 2021 en usuarios con perfil público.

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).

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.

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



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

El conjunto de datos contiene detalle de las cuentas de usuario con perfil público que han empleado hashtags para sus publicaciones en relación a los candidatos presidenciales que pasaron a la segunda vuelta electoral en Ecuador año 2021.

Los campos que se incluyen en el dataset son:

- Hashtag: Expresión clave usada por el usuario.
- Sys_time: Fecha en que se genera el raspado. Formato ISO 8601.
- User: Nombre del perfil o cuenta que hace la publicación.
- Location: Ubicación añadida en la foto o video.
- Image_description: Metadatos de la foto o video.
- Image_url: Dirección web del recurso publicado.
- Image_type: Tipo del archivo para diferenciar un video o una foto.
- Verificated: Marca de verificación de cuenta confiable.
- Publications: Es el número de posts o noticias colgadas con la cuenta.
- 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.

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 conectamos 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

- Cara candidato Andrés Arauz
  Hashtags: #andresarauz, #lassoesmoreno

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.

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. ¿Cuáles son las tendencias políticas según las publicaciones?
2. ¿Quiénes son usuarios influencers en campaña electoral?
3. ¿Cuáles son los términos de hashtags más comunes y más diferentes?
4. ¿Cuántas veces se ha referenciado una tendencia o trending topic?
5. ¿Las publicaciones entre partidos políticos, son ofensivas o inherentes a su propuesta de gestión de gobierno?

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
    - Released Under CC0: Public Domain License
    - Released Under CC BY-NC-SA 4.0 License
    - Released Under CC BY-SA 4.0 License
    - Database released under Open Database License, individual contents under Database Contents License
    - Other (specified above)
    - Unknown License

    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, consideración 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 considerar el directorio donde debe ubicarlo el programa.

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/.

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 el repositorio siguiente.

https://github.com/Mayuqui/web-scraping-instagram-presidential-candidates.git
https://github.com/jayuquina/tipologia_datos_practica1.git

Es indispensable revisar de manera manual el funcionamiento del sitio web para ver sus variaciones y limitaciones que tendría ejecutar el robot codificado, por mencionar a continuación unas cuestiones.

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 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 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.

ProtonVPN ofrece el servicio gratuito de VPN y puede descargarse el archivo de instalación aquí registrando el plan de cuenta aquí.

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

```
from selenium import webdriver
```

Se gestiona las propiedades más importantes que utiliza el código, como idioma, usuario, contraseña con que va a raspar.

```
USERNAME = 'mayuqui5'
PASSWORD = '********'
LANGUAGE ='spanish'
```

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=option
s)
```

Las solicitudes realizadas por el navegador añaden 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. Se deshabilita la detección de software automático de pruebas está controlando Chrome con esta línea.

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

A continuación, se quiere obtener la URL de la página principal en la ventana del navegador Chrome con este código.

```
driver.get("http://www.instagram.com")
```

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):
```

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'),
    ('Verificated', '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 = 2
for j in range(0, n_scrolls):
    driver.execute_script("window.scrollTo(0, document.body.scrollHei
ght);")
    time.sleep(5)
```

El tiempo de espera entre una solicitud y la extracción e datos se define con esta instrucción y un valor de 5 segundos.

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

Los datos que aparecen cuando se coloca el puntero sobre el nombre del usuario se ejecuta con el siguiente JavaScript.

```
    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 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, 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.verificated = verificated
        self.publications = publications
        self.followers = followers
        self.followed = followed
        self.comment = comment
        self.date_time = date_time

    #save data as csv
    def save_as_txt(self, name):
        sep = ';'
        sys_time = datetime.datetime.now().isoformat()
        self.location = self.location.replace(',', '|')
        self.image_description = self.image_description.replace(',',
'|')
        self.publications = self.publications.replace(',', '.')
        self.followers = self.followers.replace(',', '.')
        self.followed = self.followed.replace(',', '.')
        self.comment = self.comment.replace(sep, '|')
        self.comment = self.comment.replace(',', '|')
        self.comment = self.comment.replace('\n', '\\n')
        self.comment = self.comment.replace('\r', '\\n')
        line = name+sep+sys_time+sep+self.user+sep+self.location+sep+
self.image_description+sep+self.image_url+sep+self.image_type+sep+sel
f.verificated+sep+self.publications+sep+self.followers+sep+self.follo
wed+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_siz
e==0

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

```
        if empty:
            header = 'hashtag'+sep+'sys_time'+sep+'user'+sep+'locatio
n'+sep+'image_description'+sep+'image_url'+sep+'image_type'+sep+'veri
ficated'+sep+'publications'+sep+'followers'+sep+'followed'+sep+'comme
nt'+sep+'date_time'+'\n'
            fwrite.write(header)

        fwrite.write(line)
        fwrite.close()
```

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

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 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-

Influencia_elecciones_presidenciales_ec_2021.csv			
Influencia de elecciones presidenciales Ecuador 2021 en usuarios con perfil público.			
Columna (inglés)	**Columna (español)**	**Descripción**	**Ejemplo**
			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.	
Verificated	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

Influencia_elecciones_presidenciales_ec_2021.csv			
Influencia de elecciones presidenciales Ecuador 2021 en usuarios con perfil público.			
Columna (inglés)	**Columna (español)**	**Descripción**	**Ejemplo**
Comment	Comentario	La opinión o apreciación del dueño de la publicación.	#Mundo | El Consejo Nacional Electoral (CNE) de Ecuador anunciÃ³ este viernes que serÃ¡ el candidato conservador por la coaliciÃ³n Creo-psc| Guillermo Lasso| quien competirÃ¡ con AndrÃ©s Arauz| de UniÃ³n 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).

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