Universidad Central del Ecuador

Facultad de Ingeniería y Ciencias Aplicadas

Sistemas de Información

Sexto Semestre

Desarrollo de Sistemas de Información

Grupo 1

Mateo Cobo

Frixon Luna

Luis Paspuezan

José Soto

Web Scraping y Ley de Benford

1. **Selección del usuario**

Para aplicar la técnica de Web Scraping y posteriormente analizar los datos mediante la Ley de Benford, se seleccionó la cuenta de Instagram @ecuador_fotografia. Esta cuenta fue elegida por contar con una cantidad considerable de seguidores (534) y seguidos (349), lo que permite obtener un conjunto de datos adecuado para realizar el análisis.

2. **Búsqueda de repositorio o código base**

Para realizar la extracción de datos se utilizó el repositorio de GitHub:

https://github.com/tonoli/instagram-followers-scraper

Este proyecto emplea Python y Selenium para obtener la lista de seguidores y seguidos de una cuenta de Instagram, además de comparar los resultados con listas previas.

Aunque el código tenía 5 años de antigüedad, fue actualizado y adaptado para asegurar su correcto funcionamiento con la interfaz actual de Instagram y para que cumpliera con los objetivos del análisis propuesto.

3. **Revisión de tecnologías y librerías**

- **Selenium:** se utiliza para automatizar el navegador web (Google Chrome), iniciar sesión en Instagram, abrir el perfil del usuario, hacer scroll y extraer la lista completa de seguidores y seguidos.
- **Matplotlib**: se usa para generar las gráficas comparativas entre las frecuencias observadas y las teóricas según la Ley de Benford.
- **Pandas:** permite organizar los datos extraídos en tablas, calcular las frecuencias de los primeros dígitos y facilitar el análisis numérico.
- **Math:** se emplea para calcular los valores teóricos de la Ley de Benford mediante el uso de funciones logarítmicas.
- **Pickle:** se utiliza para guardar y cargar los datos obtenidos del scraping en archivos locales (.pkl), permitiendo conservar los resultados y compararlos luego.
- **Glob:** se usa para buscar los archivos de datos guardados y localizar las versiones más recientes para su lectura o análisis.
- **OS:** permite manejar rutas del sistema y crear carpetas donde se almacenan los resultados obtenidos del scraping.
- **Datetime:** se emplea para registrar la fecha y hora en que se generan los archivos con los datos extraídos.
- **Time:** se utiliza para controlar los tiempos de espera entre acciones dentro del scraping y permitir que las páginas carguen correctamente.
- **Getpass:** se usa para solicitar contraseñas de manera segura sin mostrarlas en pantalla durante la autenticación.

4. Ejecución del proyecto

4.1. Proceso de instalación

Para ejecutar correctamente la versión actualizada del proyecto, es necesario realizar los siguientes pasos de instalación y configuración antes de su ejecución:

4.1.1. Entorno de desarrollo

El proyecto fue desarrollado y probado utilizando Visual Studio Code (VS Code) con la versión Python 3.13.9. Se recomienda usar este mismo entorno o uno equivalente que soporte la ejecución de scripts Python y permita la instalación de dependencias mediante pip.

Asimismo, es necesario contar con Google Chrome instalado, ya que el proyecto utiliza ChromeDriver para controlar el navegador durante el proceso de scraping.

4.1.2. Descarga del proyecto

Descargar el archivo comprimido (.zip) que contiene la versión final del proyecto y descomprimirlo en una carpeta local.

4.1.3. Configuración de las cookies del usuario

En el archivo “cookies.json”, se deben colocar las cookies del usuario de Instagram que se utilizará para el proceso de scraping.

Estas cookies deben corresponder al mismo perfil configurado en el ChromeDriver, garantizando así el acceso correcto a la cuenta desde la automatización del navegador.

4.1.4. Configuración de rutas en el archivo principal

En el archivo main.py, es necesario modificar las líneas 63 y 64, estableciendo las rutas correctas del ejecutable de ChromeDriver y del archivo de cookies.

```
63    chromedriver_path = r"C:\Users\USUARIO\Desktop\instagram-followers-scraper-master\drivers\chromedriver.exe"
64    cookies_path = r"cookies.json"
```

4.1.5. Instalación de dependencias

Antes de ejecutar el proyecto, se deben instalar todas las librerías necesarias. Para ello, abrir una terminal en la carpeta principal del proyecto y ejecutar el siguiente comando:

```
pip install -r requirements.txt
```

Este comando descargará e instalará automáticamente todas las dependencias indicadas en el archivo requirements.txt, garantizando que el entorno esté completo para su correcta ejecución.

4.2. Proceso de ejecución

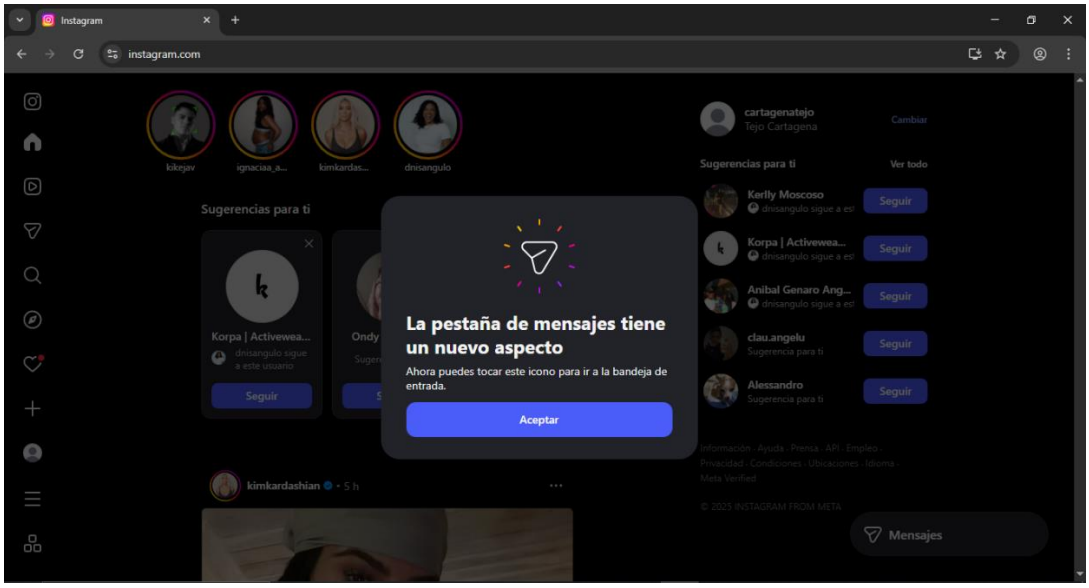
Una vez finalizada la instalación y configuración del proyecto, la ejecución se realiza de la siguiente manera:

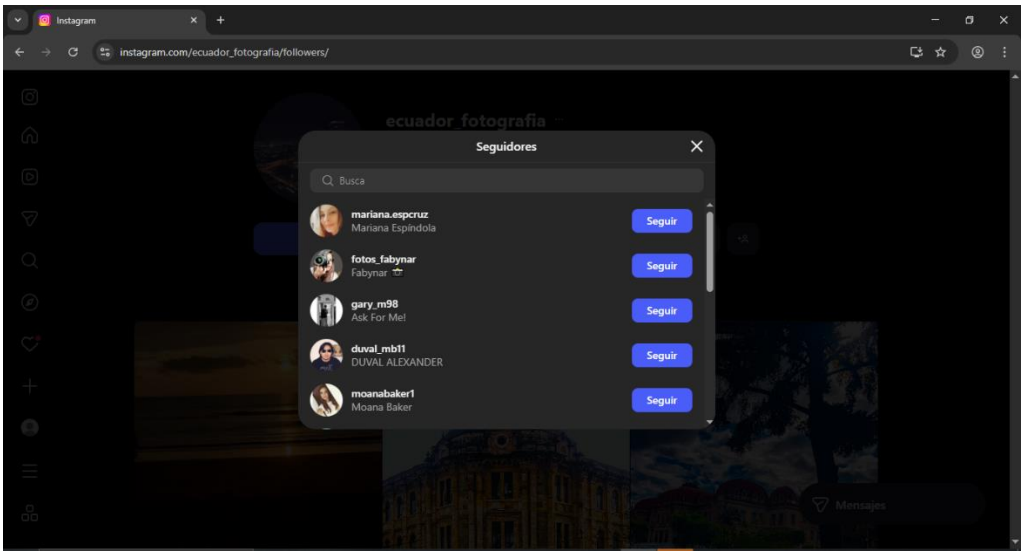
4.2.1. **Inicio del programa:** Se ejecuta el archivo main.py desde el entorno de desarrollo Visual Studio Code.

4.2.2. **Entrada de datos:** El programa solicita el nombre de usuario de Instagram del perfil que se desea analizar. A continuación, pide las credenciales de acceso (usuario y contraseña) para iniciar sesión y realizar el scraping.

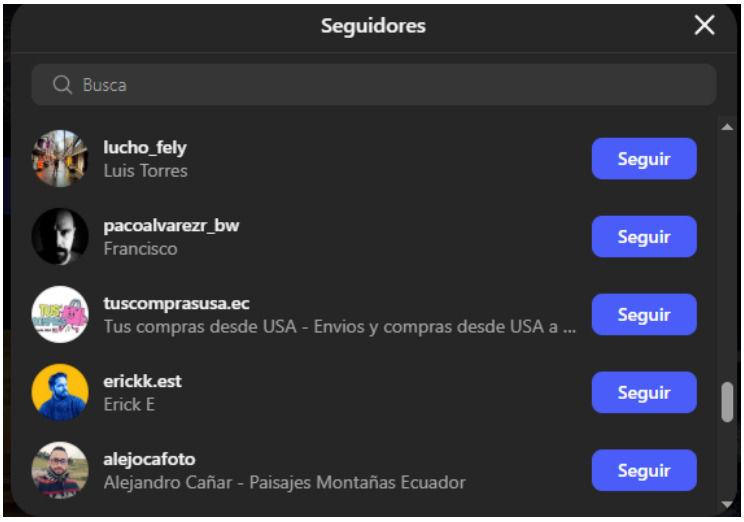
```
Enter the target username: ecuador_fotografia
Choose between
  1 - followers
  2 - following
  3 - both
: 1

Enter your Instagram credentials
Username: cartagenatejo
Password:
Intentando cargar cookies desde cookies.json...
```

4.2.3. **Carga del entorno de scraping:** El sistema carga las cookies del usuario, abre una instancia del navegador Chrome mediante ChromeDriver e inicia sesión en Instagram. Luego, se redirige automáticamente al perfil objetivo y abre la sección de seguidores.



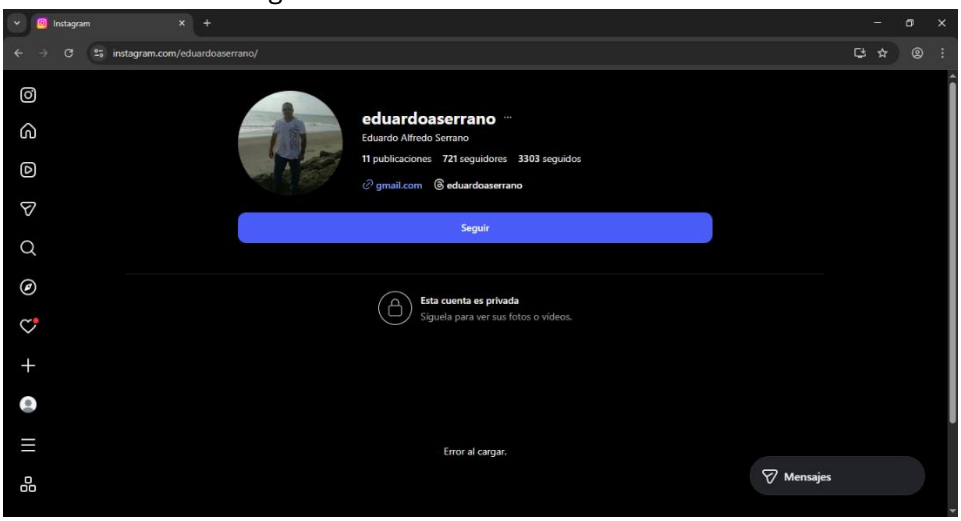


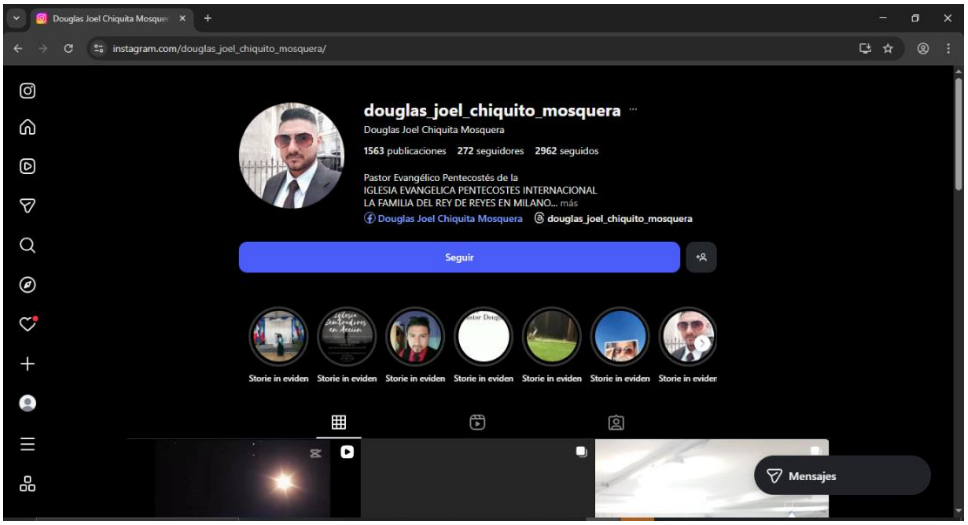
4.2.4. Extracción de información: Se ejecuta un proceso de scroll automático dentro del listado de seguidores, el cual puede observarse tanto visualmente en el navegador como en la consola.




```
dianytacedeno
micheortiz28
rogellourdes
patoguss45
🔹 Scroll #14 – Nuevos: 19
leugimange
fotoideacreativa
jrge04
levantaelvuelo04
negrita268
brayanona19
laninashirleyskitchen
franciscochamputiz
shoniisra
retratos.rotsoa.nature
erik.ramirezvera
legion_estudio
🔹 Scroll #15 – Nuevos: 12
gaby.sosam
```

Cuando finaliza el desplazamiento, el programa accede a cada perfil para registrar el número total de seguidores de cada usuario.






```
vive_conociendo
delicias_y_belleza_de_ecuador
aventurcars13
aventurcars13
lilystoregye
opticasvisualstore.ec
josejiza
pipo_elbebe
hjhj_40_
nats_amdg
🔹 Scroll #42 – Nuevos: 12
🔹 Scroll #43 – Nuevos: 0
No se detectan nuevas peticiones, scroll detenido.

Usuarios obtenidos: 534
Se obtuvieron 534 usuarios.
```

4.2.5. Análisis de datos: Una vez recolectada la información, el programa genera en consola diversas tablas de resultados:

- Lista de usuarios con su número de seguidores y primer dígito.

```
Resultados:

=== TABLA DE RESULTADOS ===
                    Usuario  Número de Seguidores  Primer Dígito
                 alexpalric                   251              2
         bestplaces_ecuador23                   150              1
                miviajedesde0                   228              2
                krliitasuarez                   395              3
                  johaloartes                   696              6
              daniabelmendoza                   655              6
                     eikadr24                   776              7
                  mafferbb89                   271              2
              jacquipaocrespo                  1595              1
                gabrielamoyat                   874              8
                    joha_ccsz                   440              4
          ecuadorenunasolafoto                   174              1
                stalyn_torresh                   237              2
```

- Frecuencia de los primeros dígitos observados.

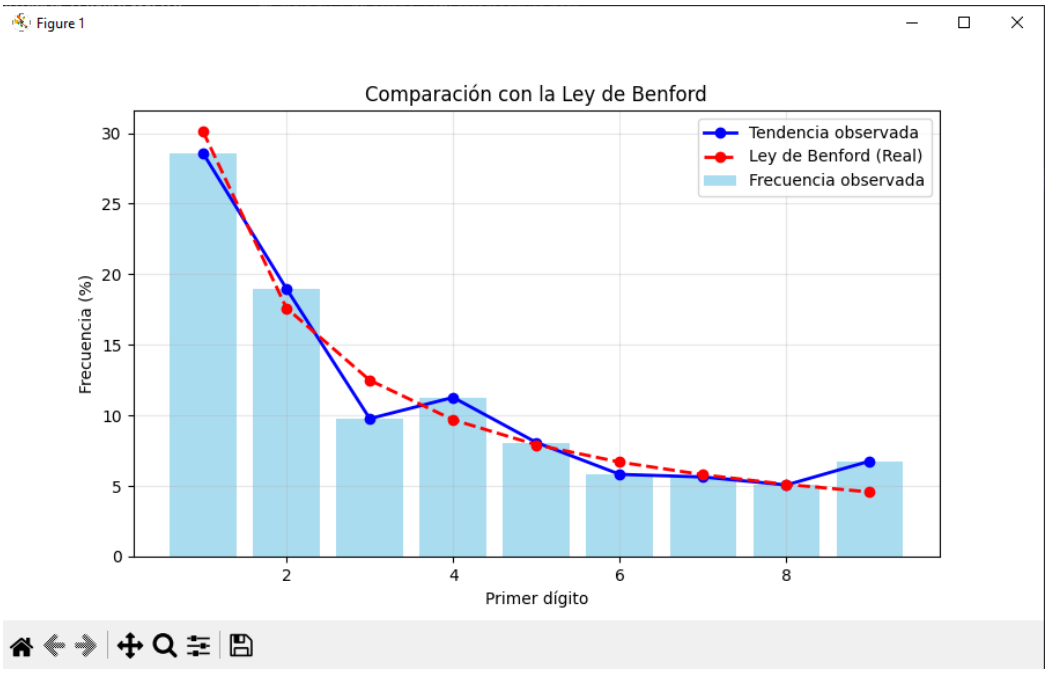
```
=== FRECUENCIA DE PRIMEROS DÍGITOS ===
 Dígito  Frecuencia
      1         152
      2         101
      3          52
      4          60
      5          43
      6          31
      7          30
      8          27
      9          36
```

- Análisis de la Ley de Benford, que incluye el dígito, la frecuencia observada, el valor teórico de Benford y la desviación porcentual. También se imprime la desviación promedio calculada.

```
=== ANÁLISIS LEY DE BENFORD ===
Dígito  Observado (%)  Benford (%) Desviación (%)
1       28.57          30.10       1.53
2       18.98          17.61       1.38
3       9.77           12.49       2.72
4       11.28          9.69        1.59
5       8.08           7.92        0.16
6       5.83           6.69        0.87
7       5.64           5.80        0.16
8       5.08           5.12        0.04
9       6.77           4.58        2.19

Desviación promedio: 1.18%
```

4.2.6. Visualización de resultados: El sistema genera un gráfico comparativo con:

- Frecuencia observada
- Tendencia observada
- Ley de Benford



Finalmente, se muestra un mensaje interpretativo, indicando si existe o no probabilidad de que la cuenta analizada sea real o bot.

```
Conclusión: Cuenta real (desviacion inferior al 12%).
```

5. Conclusiones del análisis

5.1. Evaluación del funcionamiento del código

El código actualizado logró extraer correctamente los datos de los perfiles analizados. Se realizaron pruebas con cuentas que tenían entre 50 y 500 seguidores, obteniendo resultados coherentes y precisos tanto en la recolección como en el análisis estadístico de la información. Esto demuestra que el programa cumple adecuadamente su función principal: recopilar los números de seguidores y aplicar la Ley de Benford para evaluar la autenticidad del comportamiento de las cuentas.

5.2. Limitaciones identificadas

Entre las limitaciones detectadas, se observó que el scroll automático puede fallar o detenerse según el rendimiento de la computadora utilizada, ya que el proceso exige recursos considerables al abrir múltiples perfiles en el navegador.

Además, es necesario iniciar sesión con una cuenta válida de Instagram, pues solo así la plataforma permite acceder a la información de los seguidores.

El proceso de scraping puede volverse lento en cuentas con muchos seguidores, y existe el riesgo de bloqueos temporales o restricciones por parte de Instagram si se realizan demasiadas solicitudes en poco tiempo.

Otra limitación importante es la dependencia del DOM: cualquier cambio en la estructura interna de la página puede hacer que el código deje de funcionar correctamente.

Finalmente, el análisis se limita a cuentas públicas y no considera cuentas privadas.

5.3. Implicaciones éticas y legales

El web scraping en redes sociales debe aplicarse con responsabilidad y respetando los términos de servicio de Instagram. Aunque el propósito del proyecto es puramente académico, es importante preservar la privacidad de los usuarios y no recopilar información sensible ni utilizarla con fines comerciales o de vigilancia.

La práctica ética del scraping implica obtener consentimiento del usuario o utilizar solo datos públicos y visibles sin vulnerar mecanismos de seguridad.

5.4. Propuestas de mejora o alternativas

Entre las mejoras y alternativas posibles se sugiere:

- Implementar la API oficial de Instagram, que permite acceder de forma controlada y conforme a las políticas de la plataforma.
- Sustituir Selenium por herramientas más modernas como Playwright, que ofrece mejor rendimiento y control sobre el navegador.
- Optimizar la lógica de scroll y el manejo de errores para perfiles con un alto número de seguidores.
- Agregar un sistema de límite de solicitudes y caché, que reduzca la probabilidad de bloqueos.
- Incorporar una interfaz gráfica o un módulo de exportación de resultados, para facilitar el uso y la interpretación del análisis.

5.5. Verificación del cumplimiento de la Ley de Benford

Los resultados del análisis estadístico mostraron que la frecuencia de los primeros dígitos del número de seguidores se ajusta adecuadamente a la Ley de Benford, con una desviación promedio baja entre los valores observados y teóricos.

Esto indica que la distribución de los datos sigue un patrón natural y no presenta anomalías propias de cuentas automatizadas.

Por lo tanto, se concluye que la cuenta analizada muestra un comportamiento orgánico y humano, sin indicios de ser un bot.