

SISTEMAS WEB

2. Práctica: Descarga de ficheros PDF de eGela

OBJETIVO

Usando el IDE PyCharm, tenéis que programar un cliente web en Python que descargue a una carpeta de ordenador los ficheros PDF que aparecen en la página principal de eGela de esta asignatura. El cliente web deberá cumplir la siguiente condición:

- Para la implantación y mantenimiento de la sesión http con el servidor de eGela, las cabeceras Cookie y Set-Cookie se gestionarán de forma bruta (no se puede utilizar `requests.Session()`)

ENTREGABLES

El resultado de la práctica está compuesto por dos entregables:

- **Burp**: un documento Word con las capturas de pantalla de las solicitudes que hay que realizar al servidor de eGela para establecer la sesión http, se vea el nombre del usuario en la última.
- **Python**: cliente web que descarga a una carpeta del ordenador los ficheros PDF que aparecen en la página principal de eGela de esta asignatura.

Los entregables deben subir a la tarea disponible en eGela al finalizar la práctica para cada uno de los grupos.

PASOS E INSTRUCCIONES PARA LA PRÁCTICA

- 1.- *Identificar las peticiones que debe realizar el cliente web utilizando el monitor de red del navegador.*
- 2.- *Utilizando Burp conocer el proceso de implementación de la sesión HTTP con el servidor de eGela.*
- 3.- *Programar en Python, utilizando la librería requests, una sesión HTTP con el servidor de eGela.*
- 4.- *Localizar los PDF que aparecen en la página principal de eGela de esta asignatura en la estructura del HTML utilizando el bookmarklet "Visual Source Chart".*
- 5.- *Realizar el cliente Python que descarga los PDF.*

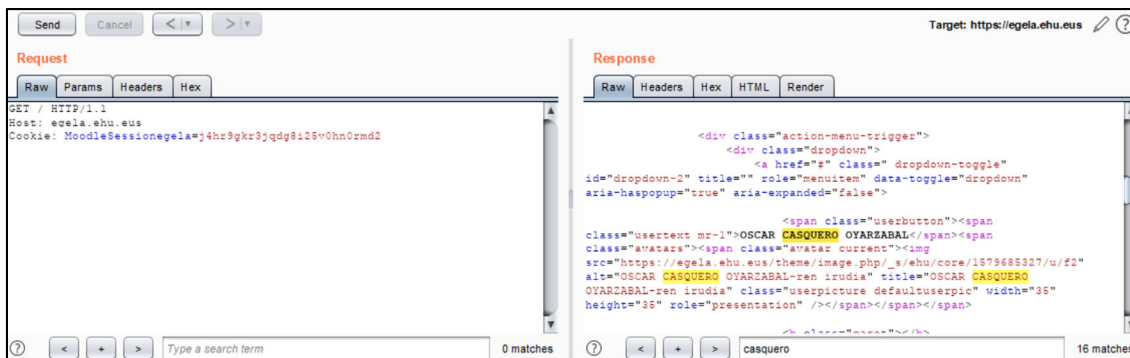
1. Identificar las peticiones que debe realizar el cliente web utilizando el monitor de red del navegador.

Utilizando el navegador, solicitar la página de login de eGela, autenticarte en la misma e identificar las solicitudes que se realizan hasta que aparece la página de este curso. Para ello utiliza el monitor de red de las herramientas de desarrollo del navegador (https://developer.mozilla.org/en-US/docs/Tools/Network_Monitor).

Nota: la solicitud de la página principal de inicio <https://egela.ehu.eus/> no se tendrá en cuenta. El proceso comenzará solicitando la hoja de login <https://egela.ehu.eus/login/index.php>
Pregunta: ¿qué vais a utilizar para hacer las solicitudes? ¿Requests o selenium+geckodriver?

2.- Utilizando Burp conocer el proceso de implementación de la sesión HTTP con el servidor de eGela

Los pasos para llevar a cabo este apartado se dieron en la clase del 17-02-2022 (<https://egela.ehu.eus/mod/resource/view.php?id=5232111>). Para acreditar la correcta ejecución del procedimiento de implementación, en la respuesta que se devuelve al realizar la última solicitud, deben figurar vuestro nombre y apellidos en el contenido (indicativo de que se ha autenticado correctamente).



3.- Programar en Python, utilizando la librería requests, una sesión HTTP con el servidor de eGela.

Basado en el paso anterior, llevaréis a cabo el proceso de implementación en Python. Por lo tanto, tenéis que programar cuatro parejas de solicitudes y respuestas cumpliendo las siguientes condiciones:

- El programa se llamará desde el terminal siguiendo la siguiente estructura:
>python eGela_PDF_downloader.py usuario "NOMBRE APELLIDO". Por ejemplo:

```
>python eGela_PDF_downloader.py bcpalgun "LUZ ALVAREZ"
```

- La contraseña se solicita utilizando la librería *getpass*.
- Introduciréis manualmente el valor del URI de la primera solicitud ("hardcoded"). **Los URIs de las siguientes solicitudes no pueden ser hardcoded, es decir, los valores de**



las URIs de estas solicitudes deben estar parametrizados (salir de la respuesta anterior).

- Extraer el valor del parámetro de **logintoken** parseando el HTML (no se puede introducir manualmente).
- Los valores de las cabeceras de **cookies** deben estar parametrizados en función de las respuestas anteriores (no se pueden introducir manualmente).
- El programa no leerá todo el valor de la cabecera Set-Cookie, sino sólo la parte que contenga el modelo **MoodleSessionegela=n03hvgma567kq290985634afsp**. Esta parte es la que incluiréis en la cabecera de Cookie.
- En relación a cada solicitud-respuesta, **en la terminal de ejecución del programa se imprimirá de forma ordenada la siguiente información:**
 - **En la solicitud**, en la primera línea hay que imprimir el método y el URI (completo). Si la solicitud tiene contenido, imprimir ésta en la siguiente línea.
 - En cuanto a **la respuesta**, en la primera línea hay que imprimir el status y la descripción. En las siguientes las cabeceras location y set-cookie.
- En la cuarta solicitud, que devuelven la lista de asignaturas de cada usuario, se realizará la comprobación de la autenticación buscando el nombre y apellido del usuario en HTML (esta búsqueda se puede realizar directamente en la cadena de texto HTML; es decir, no se debe utilizar BeautifulSoup).
 - Si la autenticación no es correcta, el programa termina y sale.
 - Si la autenticación se realiza correctamente, el mensaje que lo indica se imprime y el programa se detiene hasta que el usuario pulsa cualquier tecla.

4. Localizar los PDF que aparecen en la página principal de eGela de esta asignatura en la estructura del HTML utilizando el bookmarklet Visual Source Chart.







A continuación, se realizará una quinta solicitud de la página de eGela de esta asignatura.

Para ello, parseando el HTML que vuelve de la anterior solicitud (4ª solicitud que devuelve la lista de asignaturas), se buscará el URI del enlace de esta asignatura.







- Acceder al aula virtual como usuario. ¿Cuál es el elemento característico de los PDF?
- Localiza uno de estos elementos en el HTML. ¿Cómo puedes acceder desde ahí al enlace del PDF asociado?

Sistemas Web


[Página Principal](#) / [Mis cursos](#) / [Estudios de Grado](#) / [20200_363_GIIGSI30_27702_01](#)

-  Novedades
-  Clases On-Line
-  lunes, 25 enero 09:00 (Duración del curso)
-  Tutorías María Luz Álvarez
-  PRESENTACIÓN
-  PLANIFICACIÓN

Recursos SW

-  2021-01-29 INSTALACIÓN SOFTWARE I
-  pruebaGeckodriver.py
-  2021-01-29 TUTORIAL DE PYTHON Y PYCHARM
-  2021-02-25 INSTALACIÓN SOFTWARE II
-  mysql-connector-java-5.1.38-bin.jar
-  java_mysql_test.java

HTTP

-  2021-02-04 HTTP-SOLICITUD-RESPUESTA (BURP)

4. Realizar el cliente Python que descarga los PDF.

Utilizando la librería BeautifulSoup, examinar el contenido de la última respuesta http del punto 3 (página HTML) para obtener los enlaces del punto 4 y descargar los PDF.

NOTA: recibir un 200 OK no indica que hayas recibido el PDF.