Jesús Omar Cuenca Espino A01378844

Estrategia descarga de Imagenes usando Selenium

Para esta tarea estaré usando la página de *unsplash.com* para realizar la estrategia de descarga de imágenes, pues al momento de escribir este documento, la página de *image-net* se encuentra bajo mantenimiento y no tengo acceso a las imágenes almacenadas ahí.

Sucede que como muchas páginas que proveen imágenes, tiene una barra de búsqueda, donde al colocarse algún texto, te trae una sección considerable de la base de datos y lo acomoda en un grid, desplegando las imágenes que el sitio poseé, respecto a la *keyword* que se le entregó.

Esta página tiene un *infinite-scroll*, así que conforme se va bajando en la página se van mostrando progresivamente más imágenes.

Lo primero que se debe de hacer con *selenium* es ubicar la barra de búsqueda de modo que podamos introducir el valor (keyword) que necesitamos de la BD.

Dentro del programa en cuestión vamos a tener una variable que contenga en número mínimo de imágenes que queremos extraer del sitio y que progresivamente estaremos bajando y limpiando los URL's del HTML interior del sitio almacenado.

De modo que la página está diseñada como un *infinite-scroll,* tenemos que adaptarnos y esperar a que las imágenes carguen para posteriormente bajar en la página, por lo que constantemente estaremos pidiéndole a *selenium* que prosiga a bajar en la página de modo que nos escupa más imágenes hasta alcanzar nuestro objetivo.

El programa tiene 3 posibles condiciones de salida para evitar ciclos infinitos o problemas:

1. El *height* del sitio web no ha cambiado, después de ordenarle a *selenium* que baje más.
    1. *El infinite scroll* terminó y no hay más imágenes que descargar.
    2. Es importante considerar que:
        1. Las imágenes tardan en cargar
        2. Si se baja directo al final del sitio, este no carga más imágenes y se congela.
        3. Se tiene que llegar a una cierta altura del html para que este responda y solicite más imágenes a la BD.
2. El target de imágenes se cumplió y no es necesario proseguir.
    1. Hemos obtenido el numero de imágenes deseado.
3. Hemos superado un numero de iteraciones y no hemos alcanzado nuestro objetivo de imágenes.
    1. El *infinite-scroll* de la página prosigue pero el numero de URLs que tenemos no aumenta, por lo que posiblemente se estén repitiendo imágenes, y nunca terminaríamos.

En cuanto tengamos en un arreglo los URL's de todas las imágenes que necesitamos extraer, pasaremos a otro ciclo que progresivamente mandara peticiones *get* a los URLs en

cuestión y pasará a descargar las imágenes de manera local, estas en primera instancia se estarán almacenando dentro de la carpeta *train*.

Dentro de este último ciclo tendremos un contador que al pasar un cierto porcentaje, pasará a dividir el set de imágenes en otro grupo. Es decir en lugar de guardarlo dentro de la carpeta *train*, pasará a guardarlo en la carpeta *test*.

Evidencia instalación y resultados de Selenium

Para probar las funcionalidades de *selenium* y su correcta instalación ya he creado el programa arriba descrito de una manera práctica.
Adjunto una imagen del programa cuando se le ordenó descargar al menos 100 imágenes con la palabra clave “dog”.