

Recolección de imágenes

En esta actividad se explora la técnica de *web data scraping* para recopilar imágenes de una base de datos en línea.

La herramienta que utilizaremos para el *scraping* es Selenium, la cual es una utilidad que nos permite automatizar el navegador web. El siguiente video presenta una introducción a esta librería:

<https://www.youtube.com/watch?v=mOAXEQevCAE>

La base de datos que usaremos es ImageNet:

<http://image-net.org/about-overview>

Pasos a seguir:

1. Instalar Selenium en su ambiente de desarrollo utilizando el navegador de Anaconda.
2. Seguir los pasos de instalación del driver de su explorador web de Selenium:

<https://www.youtube.com/watch?v=dz59GsdvUF8>

3. Revisar la documentación y guía de Selenium:

<https://selenium-python.readthedocs.io/>

4. Explorar y familiarizarse con la base de datos:

<http://image-net.org/download-API>

5. Investigar cómo descargar imágenes vía URL:

<https://www.pyimagesearch.com/2015/03/02/convert-url-to-image-with-python-and-opencv/>

Entregable

El programa debe solicitar un término de búsqueda al usuario, una vez ingresado el programa accesa a ImageNet y efectúa una consulta por ese término. El programa descarga las imágenes que arrojó la base de datos, el 80% de las imágenes deben ser guardadas en una carpeta llamada `./train/<término_de_búsqueda>`, el 20% restante deben ser guardadas en una carpeta `./test/<término_de_búsqueda>`.

Ejemplo, si el usuario ingresa el término de búsqueda “dog”, se deben descargar el 80% de las imágenes que arrojó ImageNet en `./train/dog` y el otro 20% en `./test/dog`.

NOTA: La base de datos ImageNet es administrada por la universidad de Stanford y Princeton, al ser verano las universidades pueden programar un mantenimiento al portal lo cual lo haría inaccesible. Una alternativa es utilizar <https://www.shutterstock.com> y recolectar sus imágenes de ahí. ImageNet es una base de datos significativamente más simple de efectuar el scraping una vez que se familiaricen con su API y recomendaría la usaran en caso de estar disponible.