Write a design document on how to crawl a website site and save the images and text from the website. Write your thoughts on how you will approach the problem along with components you need.

Before crawling a website, you need to download and install **Octaparse** on your devices, it's open source. It can help you scrape images without coding and other public content from any web page.

Example 1: Fetching Images Directly from Webpage

To demonstrate, we are going to scrape dogs' images from Pixabay.com. To follow along, search for "dogs" on Pixabay.com then you should arrive at this page: https://pixabay.com/images/search/dogs/

Step 1: Enter the URL

Click "+ Task" to start a new task under Advanced Mode. Then, input the URL of the target webpage into the text box and click "Save URL".

Step 2: Select the images you want to crawl

Next, we are going to tell the bot what images to fetch. Click on the first image, the Action Tips panel now reads "Image selected, 100 similar images found". Go on to select "Select all", then "Extract image URL in the loop".

Step 3: Crawl images across pages

Of course, we don't just want the images from page 1, but from all pages (or as many pages as needed). To do this, scroll down to the bottom of the current page, spot the "next page" button, and click on it.

We obviously want to click the "next page" button many times, so it makes sense to select "Loop click the selected link" from the Action Tips panel.

Now, just to confirm if everything was set up properly. Toggle the workflow switch in the upper right corner. Also, check the data panel and make sure we have the desired data extracted correctly.

Step 4: Crawl with Auto-scrolling settings

There's just one more thing to tweak before running the crawler.

While debugging, I noticed that the HTML source code is being refreshed dynamically as one scrolls down the webpage. In other words, if the webpage is not scrolled down, we will not be able to get the corresponding image URLs from the source code. Luckily, Octoparse does auto-scroll down easily.

We will need to add auto-scroll both when the website loads for the first time as well as when it paginates.

Click on "Go to Webpage" from the workflow. On the right side of the workflow, spot "Advanced options", check "Scroll down to the bottom of the page when finish loading".

Then, decide how many times to scroll as well as at what pace. Here I set scroll times = 40, interval=1 second, and scroll way = scroll down for one screen. This basically means Octoparse will scroll down one screen for 40 times with 1 second between each scroll.

I did not come up with this setting randomly but did a bit of fine-tunning to make sure this setting works. I also noticed that it was essential to use "Scroll down for one screen" as opposed to "scroll down to the bottom of the page". Mainly because the image URLs we need only get refreshed to the source code gradually.

Apply the same setting to the pagination step. Click on "Click to paginate" on the workflow, use the exact same setting for auto-scroll.

Step 5: Start your crawler

That's it. Let's run the crawler and see if it works. Click "Start Extraction" from the upper left-hand corner. Pick "local extraction". It basically means you'll be running the crawler on your own computer instead of the Cloud server.