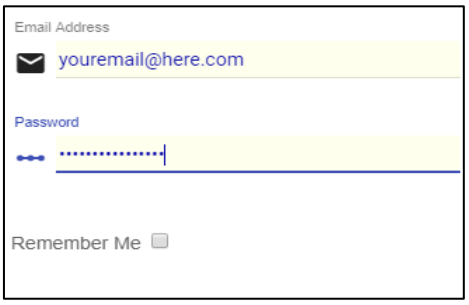
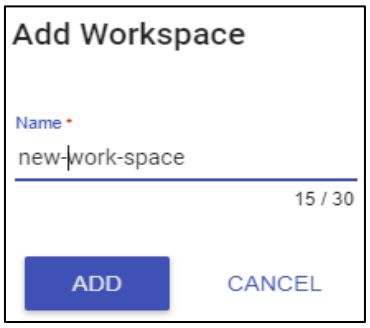
SiteHound Walk-Through Guide

SiteHound is a **Domain Discovery tool** that assists a user with finding websites that are relevant to a particular topic of interest (i.e. a domain). SiteHound performs domain discovery in a fast, automated, and scalable way by combining web crawling and machine learning technologies. The name comes from the idea of a bloodhound, which can take the scent of something and then track it down. In a similar way, our SiteHound can take the "scent" of your domain and track it down across the web! (Also, we just love dogs!)

Here's how you can use it!

1. Login with the provided credentials. If you have an individual account, you may want to change your password.

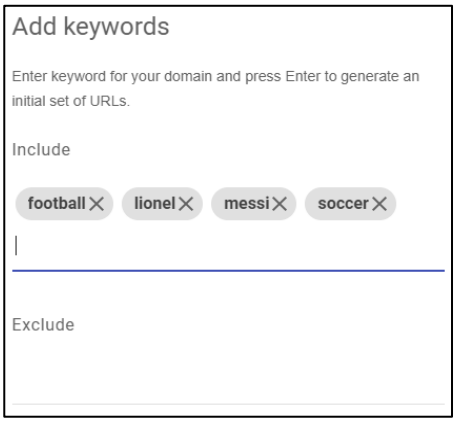




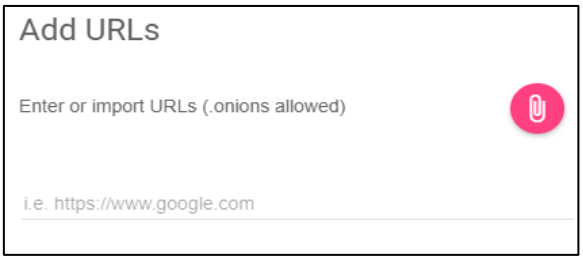
2. The next step is to **create a Workspace**. Your Workspace will contain all the work you've done on a particular topic, or domain. You can name it whatever you'd like, but use hyphens (-) instead of spaces. Enter the name, click the add button, select your new workspace.

3. Your next step is to add keywords that will help the SiteHound find the sites you're looking for. It's a good idea to include between 4-6 keywords to start (you can repeat this step several times with different keywords, so don't worry if you have more keywords you want to try).

The more keywords you add, the broader the results you'll receive. The best keywords are specific to the topic, and not too common or ambiguous (vague keywords will yield vague results). You can also *exclude* keywords that you expect might be associated with your topic, but in the wrong context. For example, if you're looking for "guns" you might want to exclude "toy" and "water".

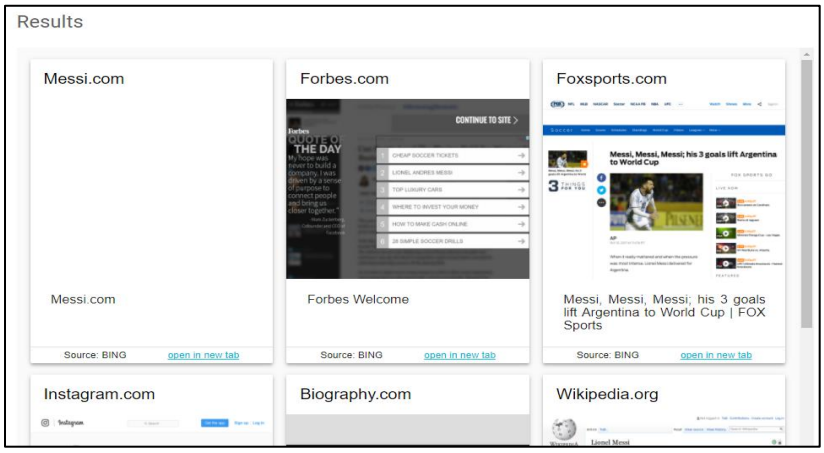


4. Next, click "Fetch Open" to search for these keywords on the open web. If you think your domain might have a presence on the Dark Web (i.e. on the Tor network), you can also click to "Fetch Dark"; if not, you can just ignore this button.

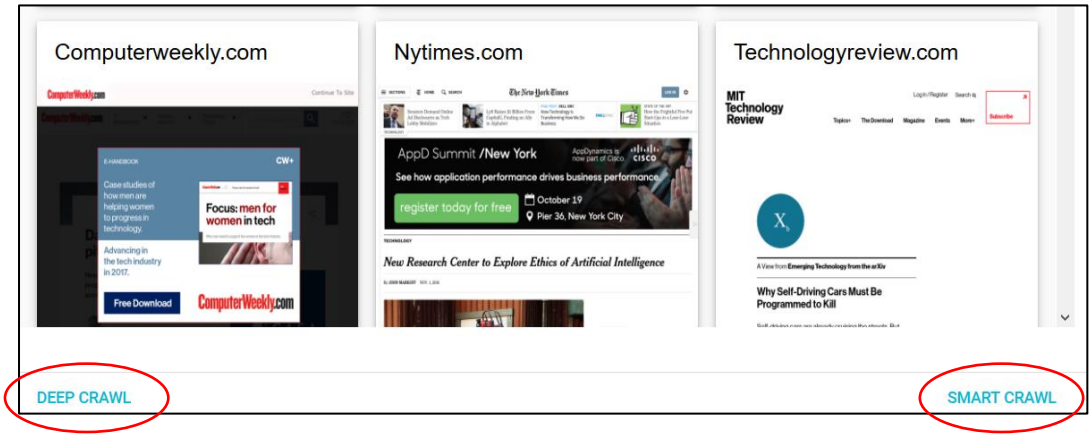


5. You can also provide some seed URLs for the crawler to fetch. Seed URLs are examples for what types of sites you're looking for, such as blog posts, news articles, forum posts, etc. You can provide *only seed URLs*, if you don't have any keywords or if you just want to deep crawl these sites (i.e. fetch as many pages from them as possible, see below), instead of discovering new sites. If you have a big list of sites, you can import them as text file, too.

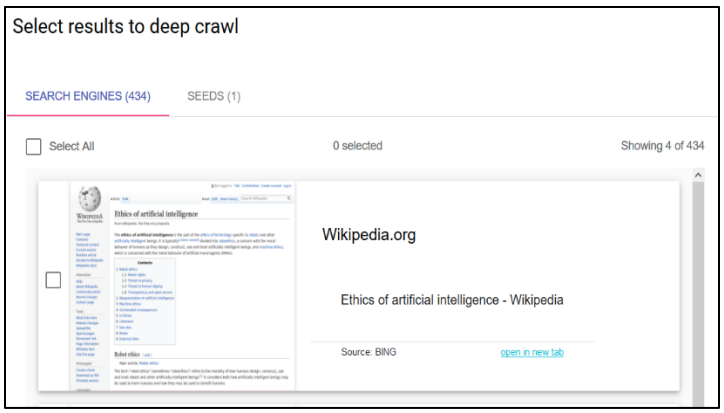
After a few seconds, you will see the results populating the bottom of the page.



Now, from here, there are two different workflows you can choose:
a **Deep Crawl** or a **Smart Crawl**.



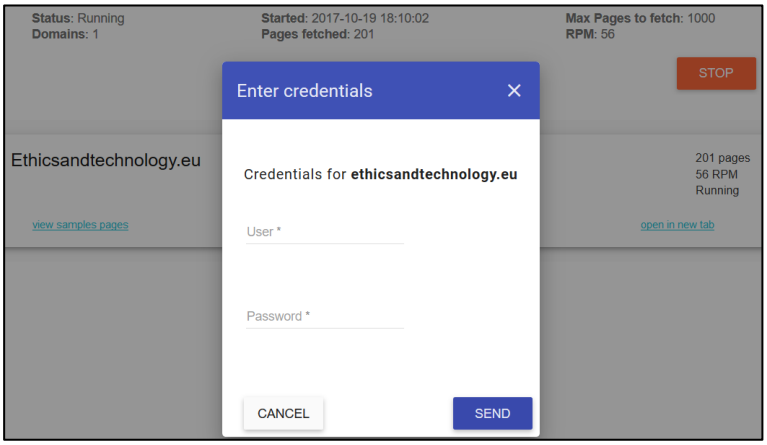
These workflows will each yield different results.

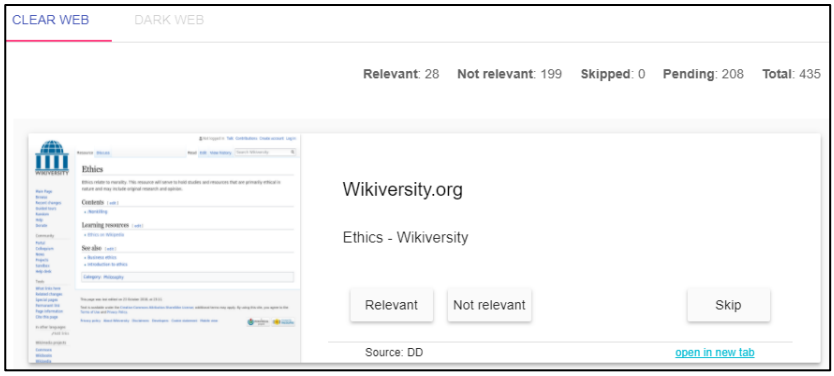


DEEP CRAWL: A Deep Crawl is simple and straightforward. It just collects pages from the sites you have specified; it does not generally go outside of those sites to find new sites. If you want to perform a Deep Crawl, scroll down to the bottom of the page and click the Deep Crawl link.

Then, select the sites you'd like deep crawled, and then click to Start Deep Crawl. You can check the results, and navigate to sample pages of each domain.

SiteHound also gives a user the ability to deep crawl a page that requires credentials to access. If one of the sites you're trying to deep crawl has a login page, you can provide the SiteHound with the username and password, and it will login and perform the deep crawl as the logged in user.

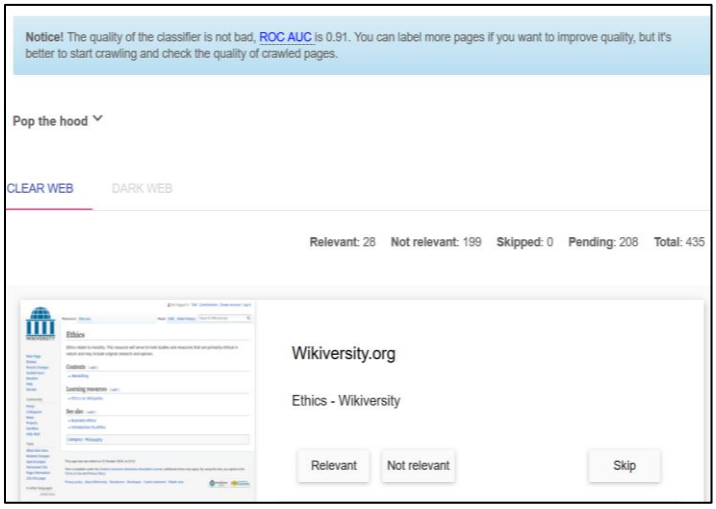


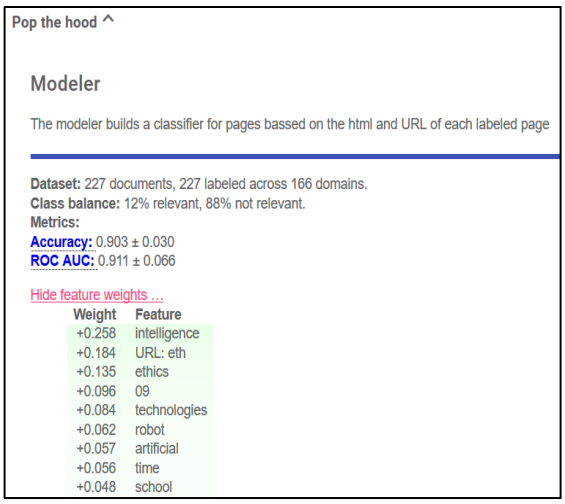


SMART CRAWL: Smart Crawls require a few more steps, but you'll be "teaching" SiteHound to find new sites that are relevant to your domain. Before you start a Smart Crawl, you must take the initial results that SiteHound has returned from your keywords (or seed URLs) and identify which results are relevant (positive) and which are irrelevant (negative).

You'll want plenty of both, so your model will have a good grasp on exactly what you're looking for; check the counts at the top and aim for a total of 100+ sites labeled as either relevant or not relevant. The total number of sites will continue to grow as SiteHound fetches new ones in the background; this is normal and you don't have to label all of them! Just a good sample should do it.

In the background, SiteHound is generating a model of your domain, and it is testing this model by showing you some example pages that it thinks is relevant and seeing if you in fact mark that page as Relevant, and the same for Irrelevant. The more it guesses right, the higher its score. Once the model reaches a threshold score (.85), you will see a blue banner at the top notifying you that SiteHound thinks its model is ready to use in a larger discovery crawl.

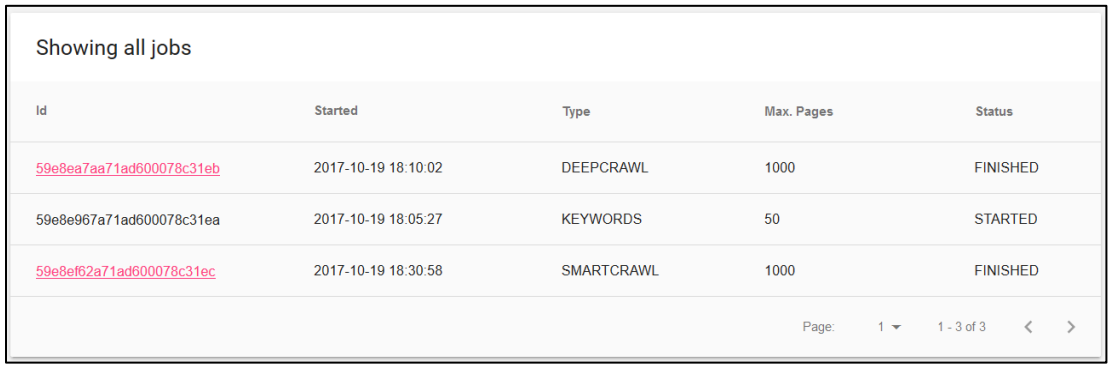




If you want to take a closer look at the model, you can **pop the hood** and inspect the features that the model is using to determine page relevance. These should look relatively intuitive to you; if they don't, or if you can't seem to get your model's score up, you may want to start over and try another set of keywords, seed URLs and/or labeling criteria.

When you're satisfied with your model, you can "turn the SiteHound loose" and start the Smart Crawl, which will spider out to find new pages and sites that are relevant to your domain.

You can check progress and view sample pages from your Smart Crawl at any time by clicking the Jobs tab.



For questions or feedback, please email us at support@hyperiongray.com.

-Cheers!

HYPERION GRAY LLC