

Learn > Google Dorking



# Google Dorking

Explaining how Search Engines work and leveraging them into finding hidden content!

 **Easy** ⌚ 0 min

 Start AttackBox



Help



 Save Room

 4802



 Options



Google Dorking | DarkStar • Jan 16, 2021

Source: YouTube

## TryHackMe Google Dorking Official Walkthrough



TryHackMe



Access Machines




1



Room progress ( 6% )



Task 1  Ye Ol' Search Engine



Task 2  Let's Learn About Crawlers



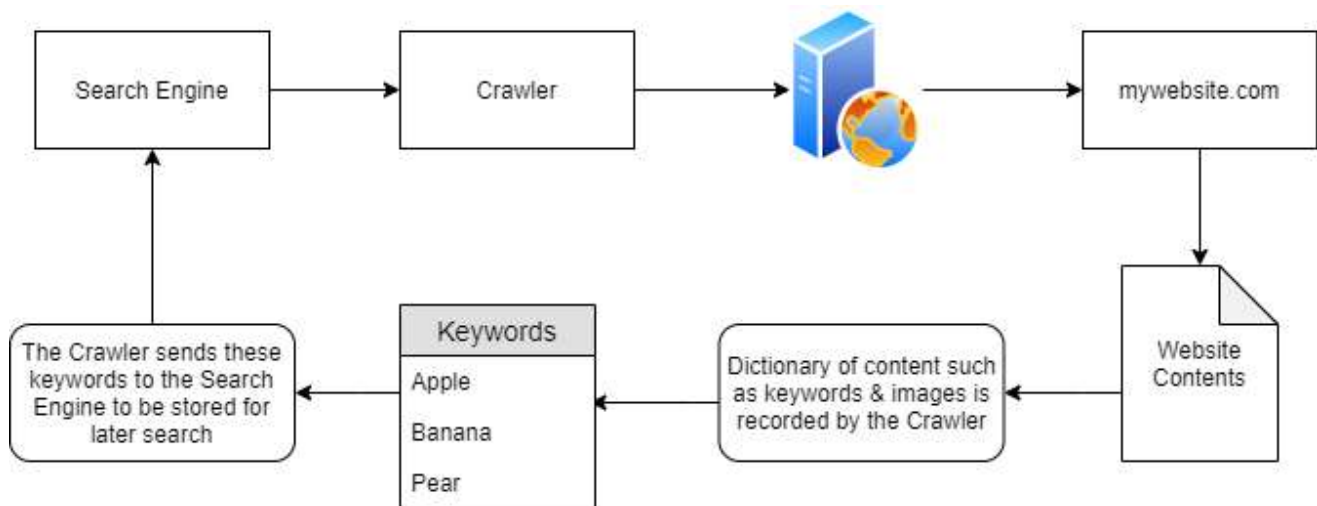
## What are Crawlers and how do They Work?

These crawlers discover content through various means. One being by pure discovery, where a URL is visited by the crawler and information regarding the content type of the website is returned to the search engine. In fact, there are lots of information modern crawlers scrape – but we will

discuss how this is used later. Another method crawlers use to discover content is by following any and all URLs found from previously crawled websites. Much like a virus in the sense that it will want to traverse/spread to everything it can.

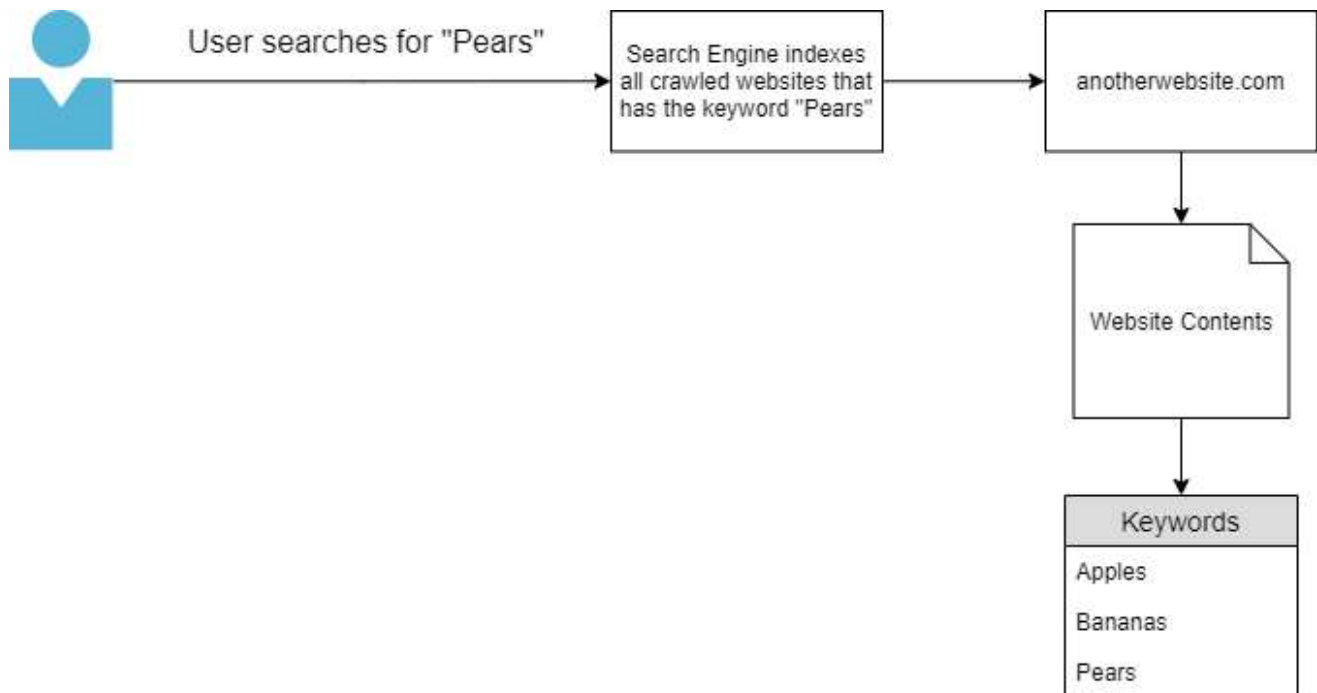
## Let's Visualise Some Things...

The diagram below is a high-level abstraction of how these web crawlers work. Once a web crawler discovers a domain such as **mywebsite.com**, it will index the entire contents of the domain, looking for keywords and other miscellaneous information - but I will discuss this miscellaneous information later.



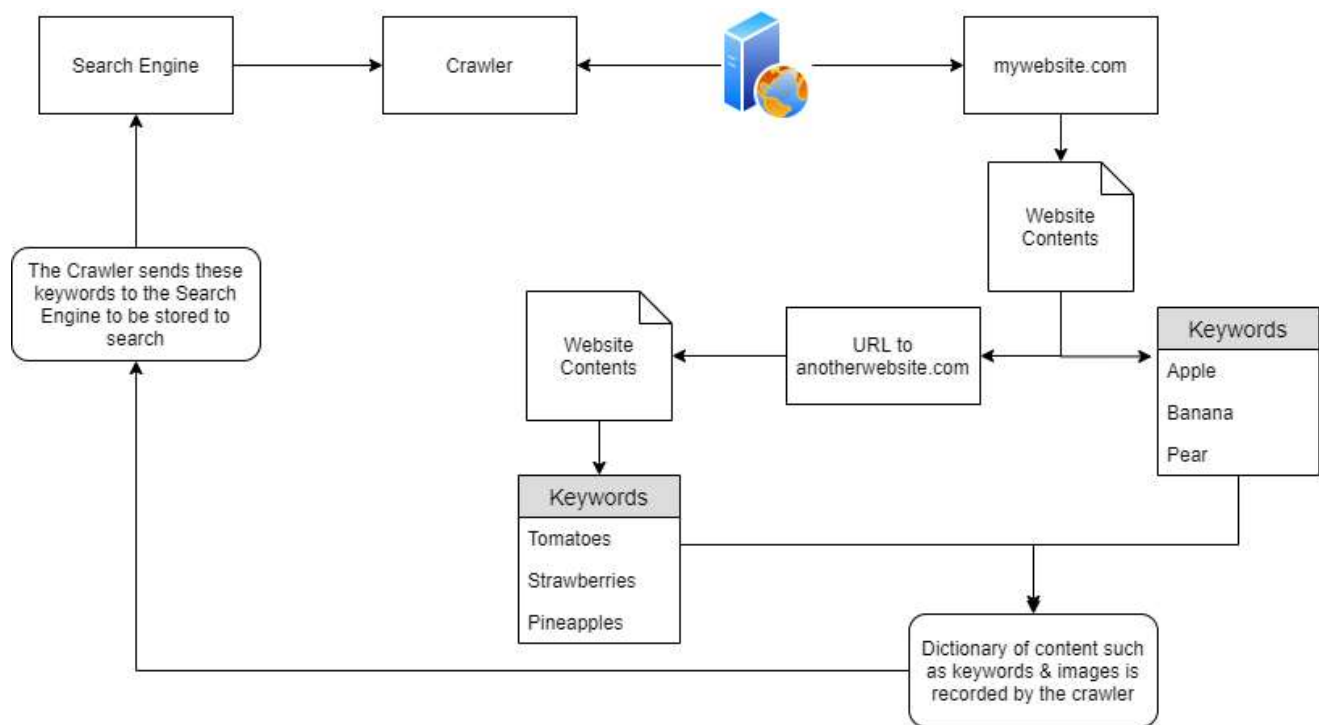
In the diagram above, "**mywebsite.com**" has been scraped as having the keywords as "Apple" "Banana" and "Pear". These keywords are stored in a dictionary by the crawler, who then returns these to the search engine i.e. Google. Because of this persistence, Google now knows that the domain "**mywebsite.com**" has the keywords "Apple", "Banana" and "Pear". As only one website has been crawled, if a user was to search for "Apple"... "**mywebsite.com**" would appear. This would result in the same behaviour if the user was to search for "Banana". As the indexed contents from the crawler report the domain as having "Banana", it will be displayed to the user.

As illustrated below, a user submits a query to the search engine of "Pears". Because the search engine only has the contents of one website that has been crawled with the keyword of "Pears" it will be the only domain that is presented to the user.



However, as we previously mentioned, **crawlers attempt to traverse, termed as crawling, every URL and file that they can find!** Say if “mywebsite.com” had the same keywords as before (“Apple”, “Banana” and “Pear”), but also had a URL to another website “anotherwebsite.com”, the crawler will then attempt to traverse everything on that URL (**anotherwebsite.com**) and retrieve the contents of everything within that domain respectively.

This is illustrated in the diagram below. The crawler initially finds “mywebsite.com”, where it crawls the contents of the website - finding the same keywords (“Apple”, “Banana” and “Pear”) as before, but it has additionally found an external URL. Once the crawler is complete on “mywebsite.com”, it'll proceed to crawl the contents of the website “anotherwebsite.com”, where the keywords (“Tomatoes”, “Strawberries” and “Pineapples”) are found on it. The crawler's dictionary now contains the contents of both “mywebsite.com” and “anotherwebsite.com”, which is then stored and saved within the search engine.



## Recapping

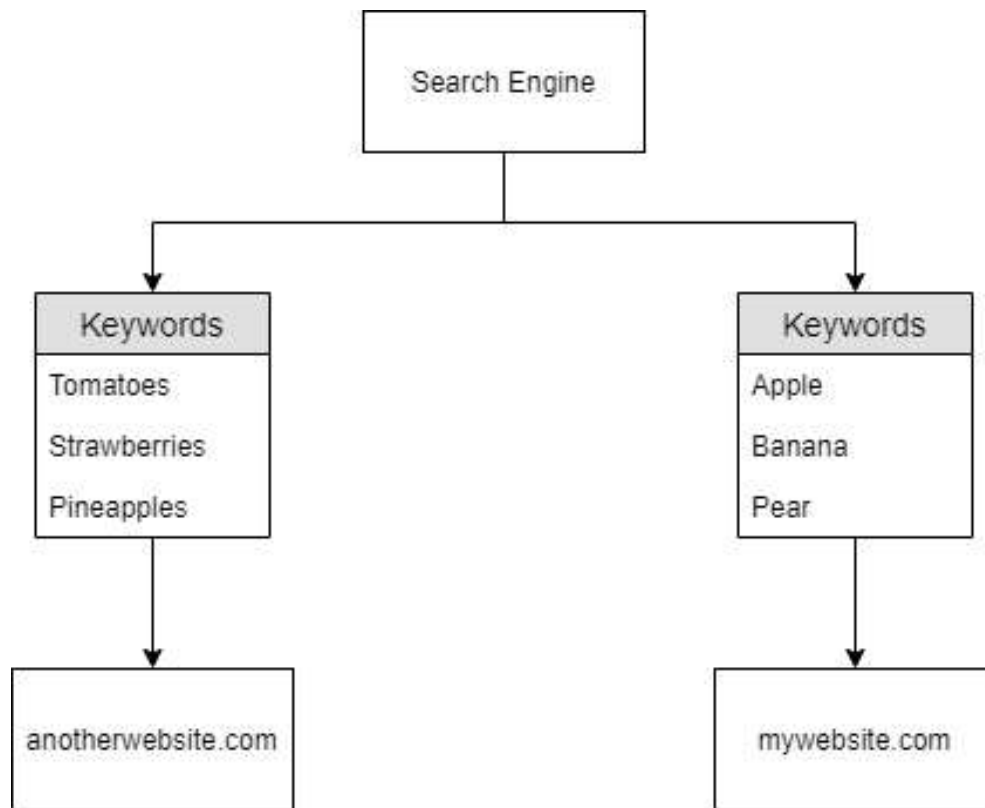
So to recap, the search engine now has knowledge of two domains that have been crawled:

1. mywebsite.com
2. anotherwebsite.com

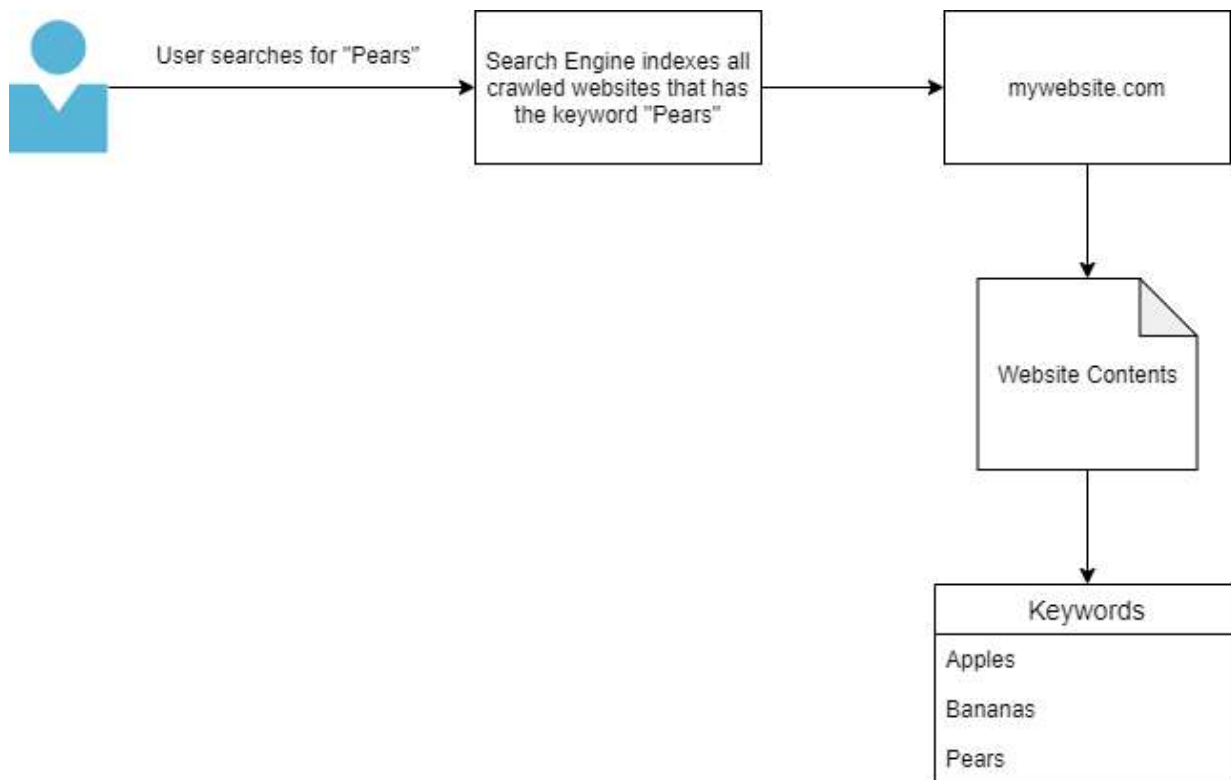
Although note that “anotherwebsite.com” was only crawled because it was referenced by the first domain “mywebsite.com”. Because of this reference, the search engine knows the following about the two domains:

Domain Name	Keyword
mywebsite.com	Apples
mywebsite.com	Bananas
mywebsite.com	Pears
anotherwebsite.com	Tomatoes
anotherwebsite.com	Strawberries
anotherwebsite.com	Pineapples

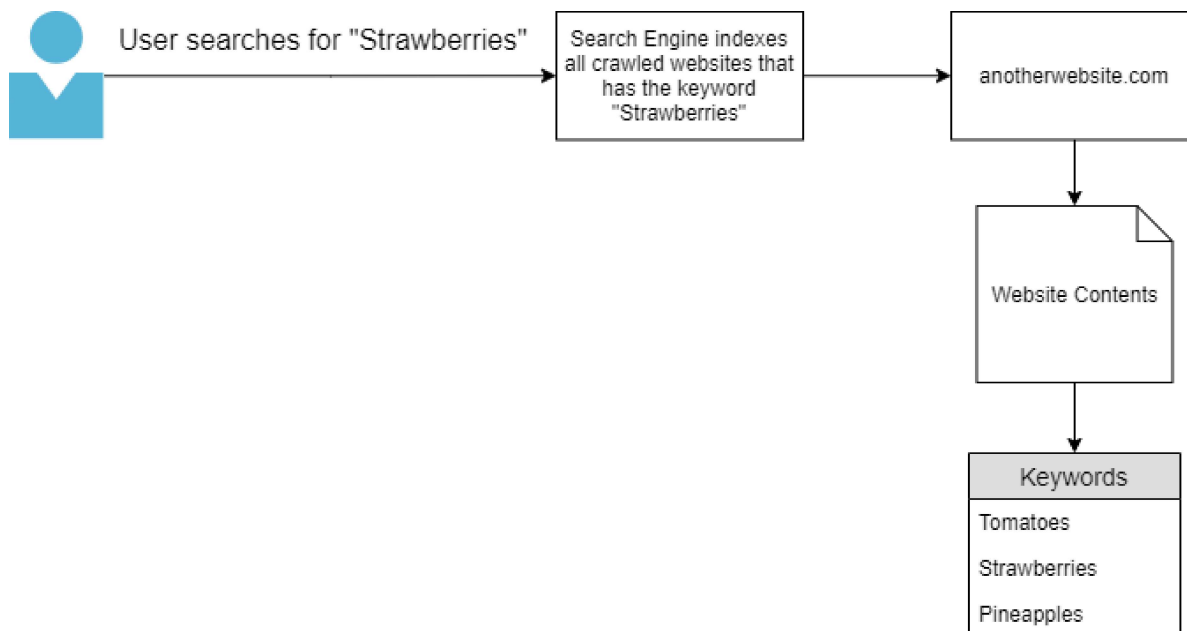
Or as illustrated below:



Now that the search engine has some knowledge about keywords, say if a user was to search for "Pears" the domain "mywebsite.com" will be displayed - as it is the only crawled domain containing "Pears":



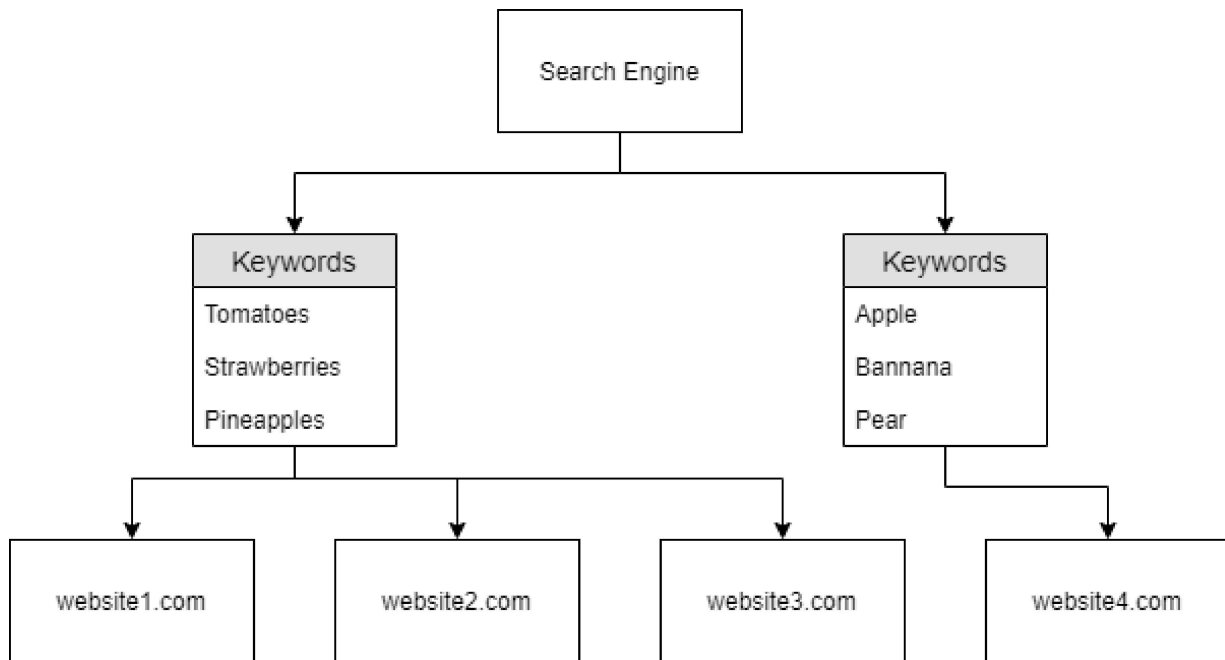
Likewise, say in this case the user now searches for "Strawberries". The domain "**anotherwebsite.com**" will be displayed, as it is the only domain that has been crawled by the search engine that contains the keyword "Strawberries":



This is great...But imagine if a website had multiple external URL's (as they often do!) That'll require a lot of crawling to take place. There's always the chance that another website might have

similar information as of that another website crawled - right? So how does the "Search Engine" decide on the hierarchy of the domains that are displayed to the user?

In the diagram below in this instance, if the user was to search for a keyword such as "Tomatoes" (which websites 1-3 contain) who decides what website gets displayed in what order?



A logical presumption would be that website 1 -> 3 would be displayed...But that's not how real-world domains work and/or are named.

So, who (or what) decides the hierarchy? Well...

Answer the questions below

Name the key term of what a "Crawler" is used to do

Answer format: \*\*\*\*\*

Submit

What is the name of the technique that "Search Engines" use to retrieve this information about websites?

Answer format: \*\*\*\*\*

Submit



What is an example of the type of contents that could be gathered from a website?

Answer format: \*\*\*\*\*

🚩 Submit

Task 3 ☐ Enter: Search Engine Optimisation



Task 4 ☐ Beepboop - Robots.txt



Task 5 ☐ Sitemaps



Task 6 ☐ What is Google Dorking?



**Created by**



cmnatic

**Room Type**

Free Room. Anyone can deploy virtual machines in the room (without being subscribed)!

**Users in Room**

114,695

**Created**

1621 days ago

Copyright TryHackMe 2018-2024

