

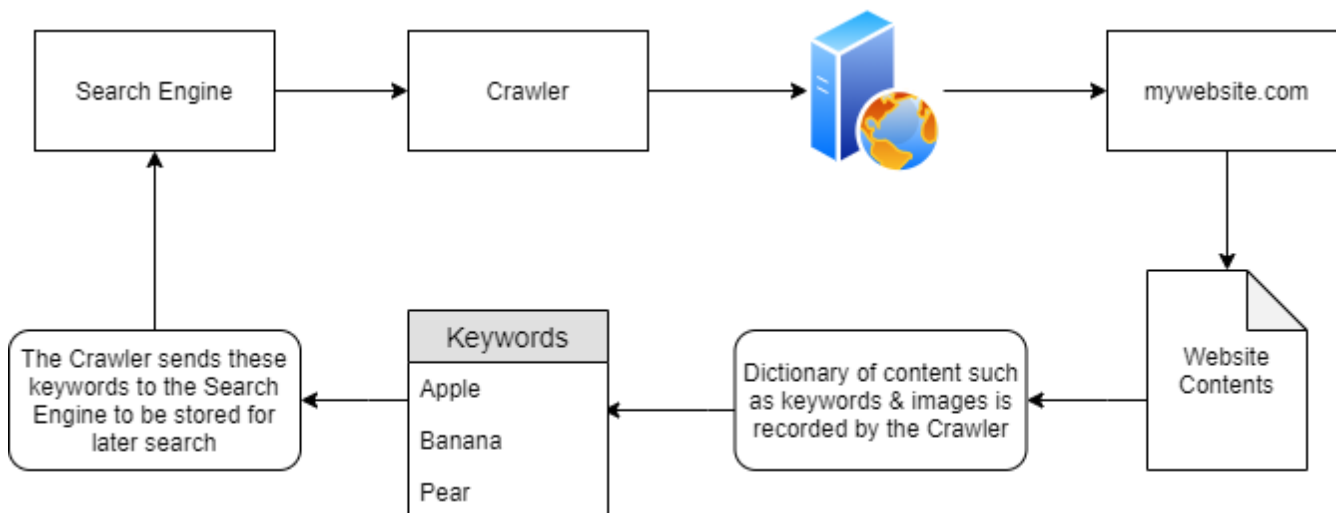
Let's Learn About Crawlers

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...

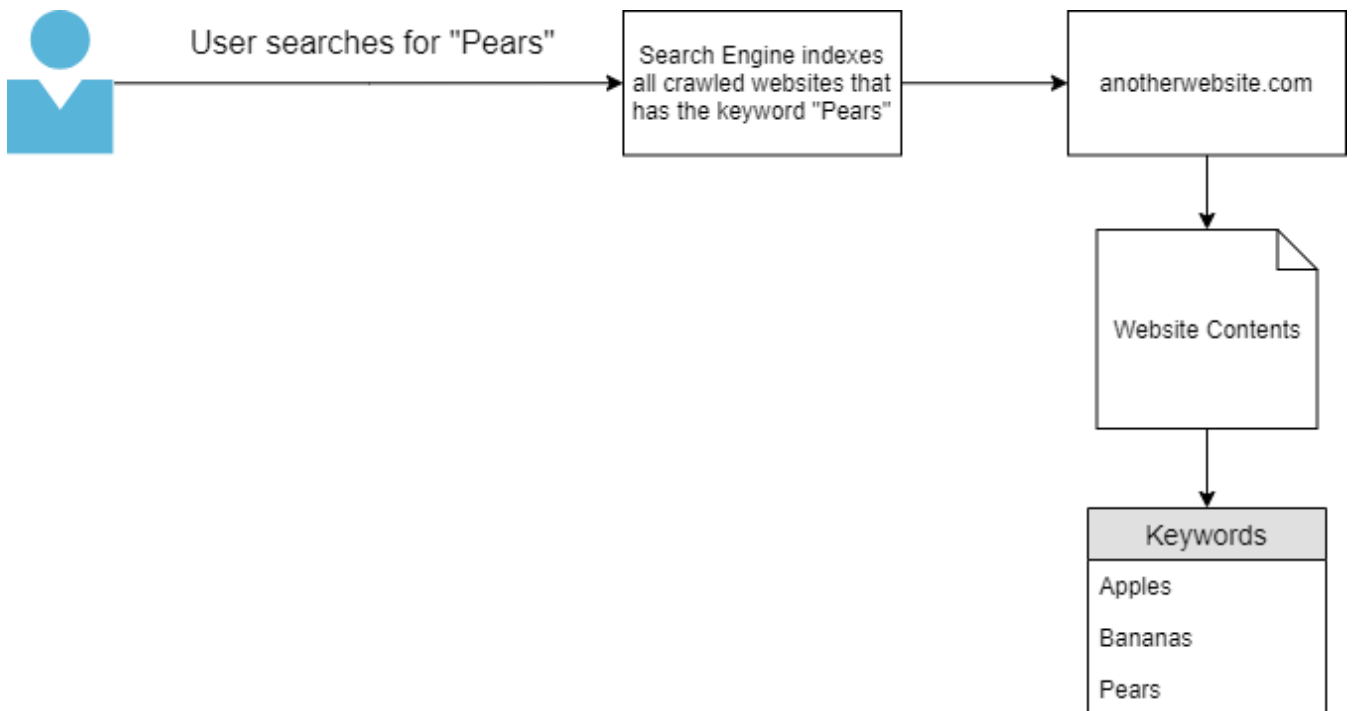
The diagram below is a high-level abstraction of how these web crawlers work. Once a web crawler discovers a domain such as **mywebsite.com**, it will index the entire contents of the domain, looking for keywords and other miscellaneous information - but I will discuss this miscellaneous information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

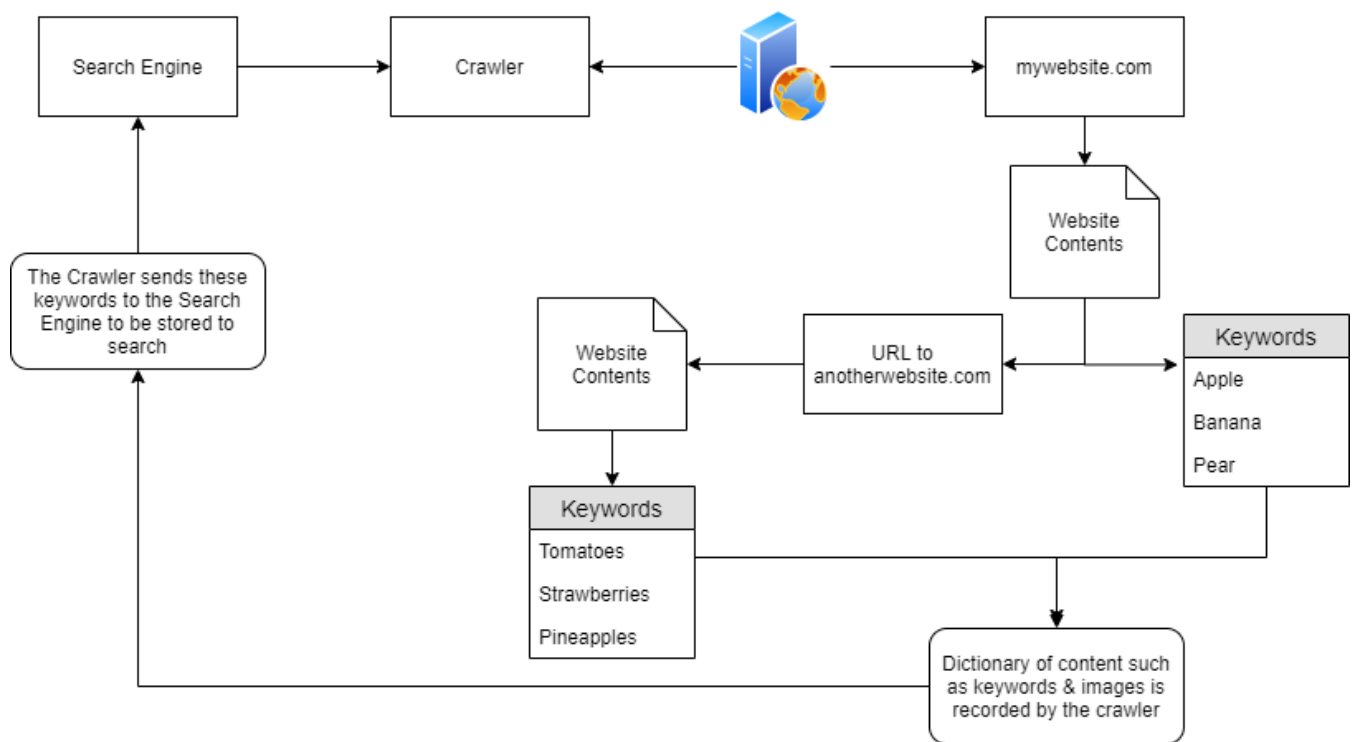
user was to search for “Banana”. As the indexed contents from the crawler report the domain as having “Banana”, it will be displayed to the user.

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

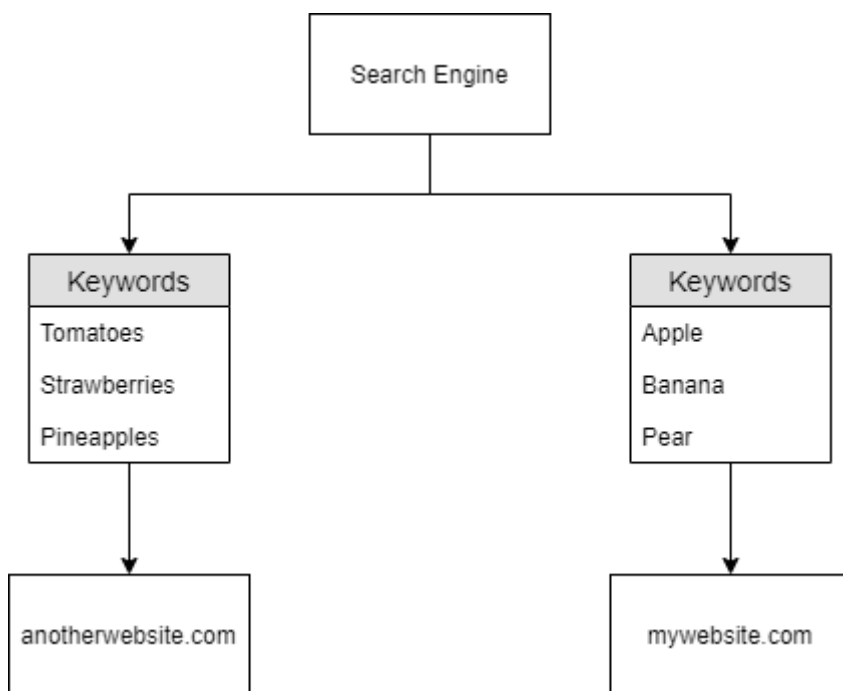
So to recap, the search engine now has knowledge of two domains that have been crawled:

1. mywebsite.com
2. anotherwebsite.com

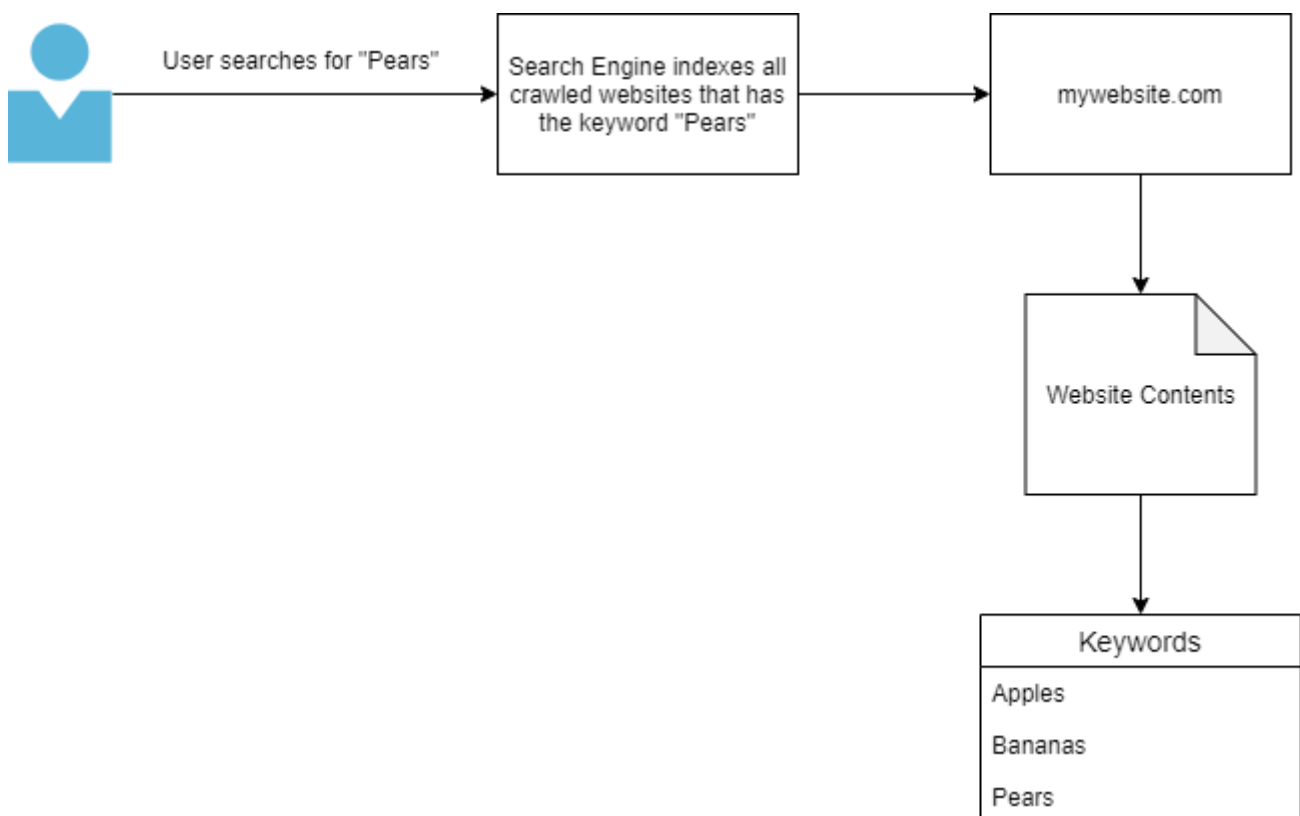
Although note that “**anotherwebsite.com**” was only crawled because it was referenced by the first domain “**mywebsite.com**”. Because of this reference, the search engine knows the following about the two domains:

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

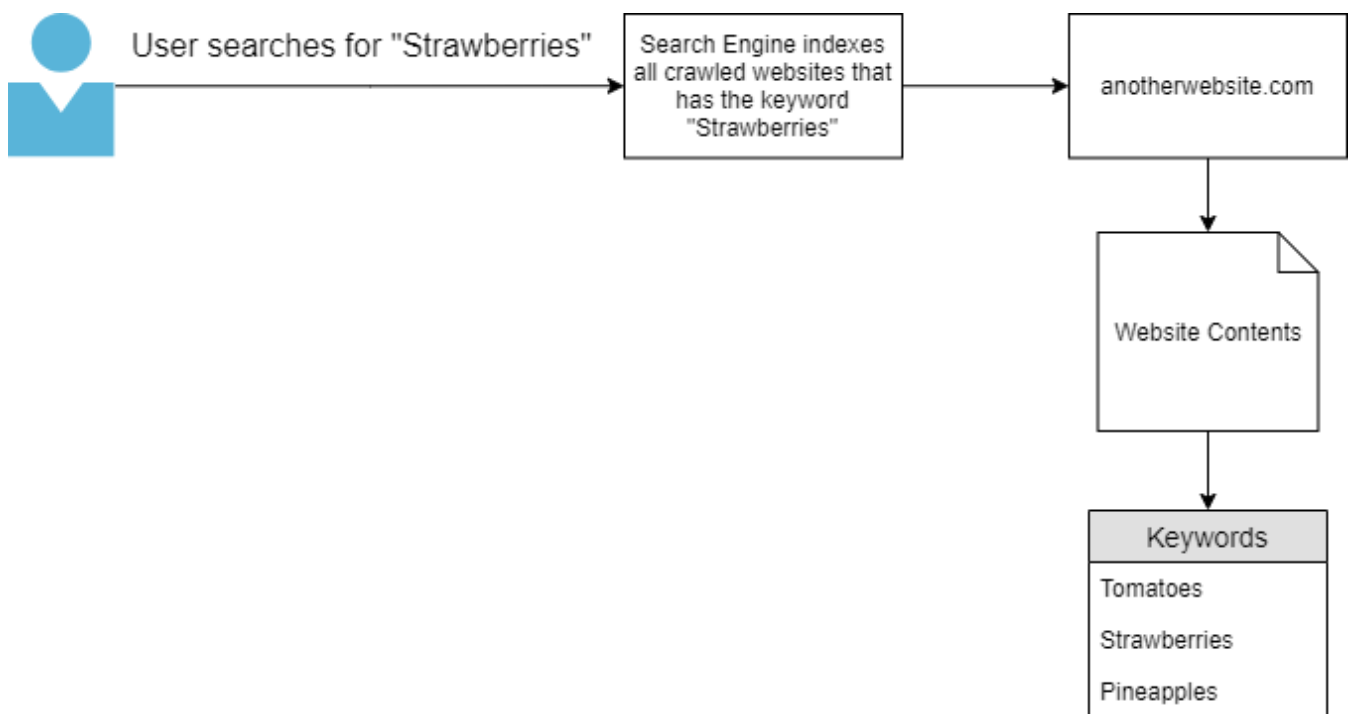
Or as illustrated below:



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":

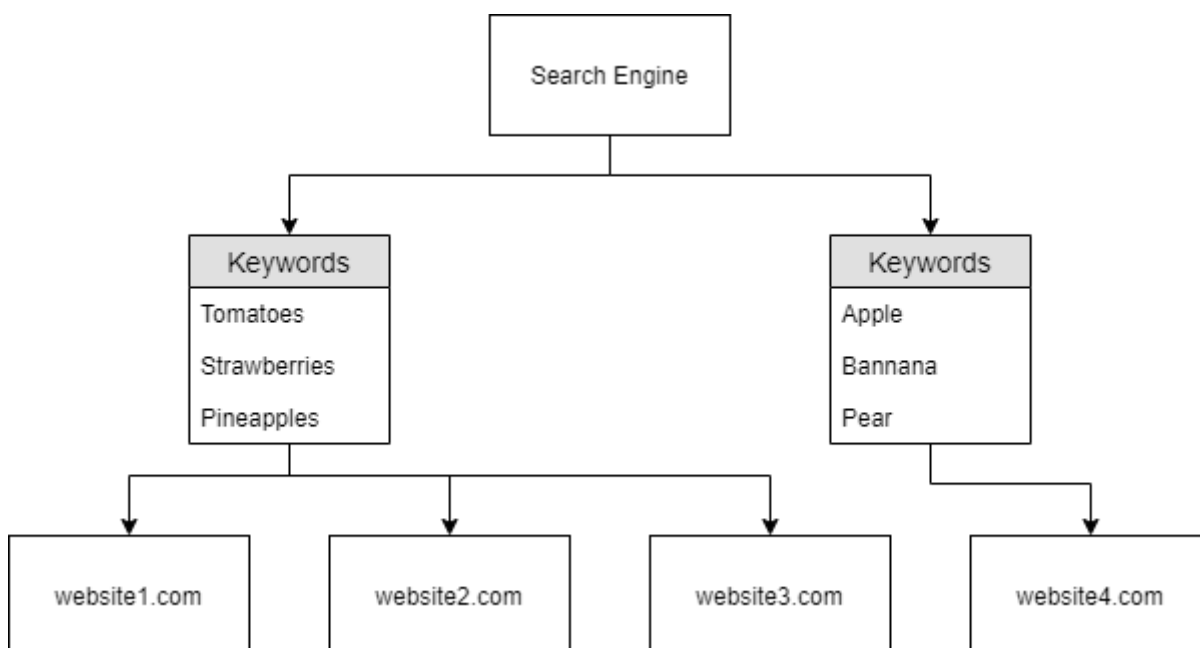


Likewise, say in this case the user now searches for "Strawberries". The domain "**anotherwebsite.com**" will be displayed, as it is the only domain that has been crawled by the search engine that contains the keyword "Strawberries":



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...

Enter: Search Engine Optimisation

Search Engine Optimisation

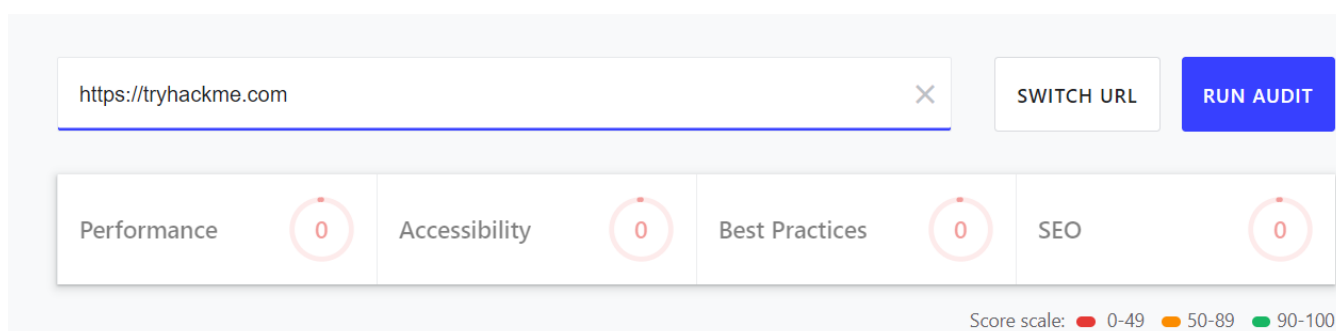
Search Engine Optimisation or SEO is a prevalent and lucrative topic in modern-day search engines. In fact, so much so, that entire businesses capitalise on improving a domains SEO “ranking”. At an abstract view, search engines will “prioritise” those domains that are easier to index. There are many factors in how “optimal” a domain is - resulting in something similar to a point-scoring system.

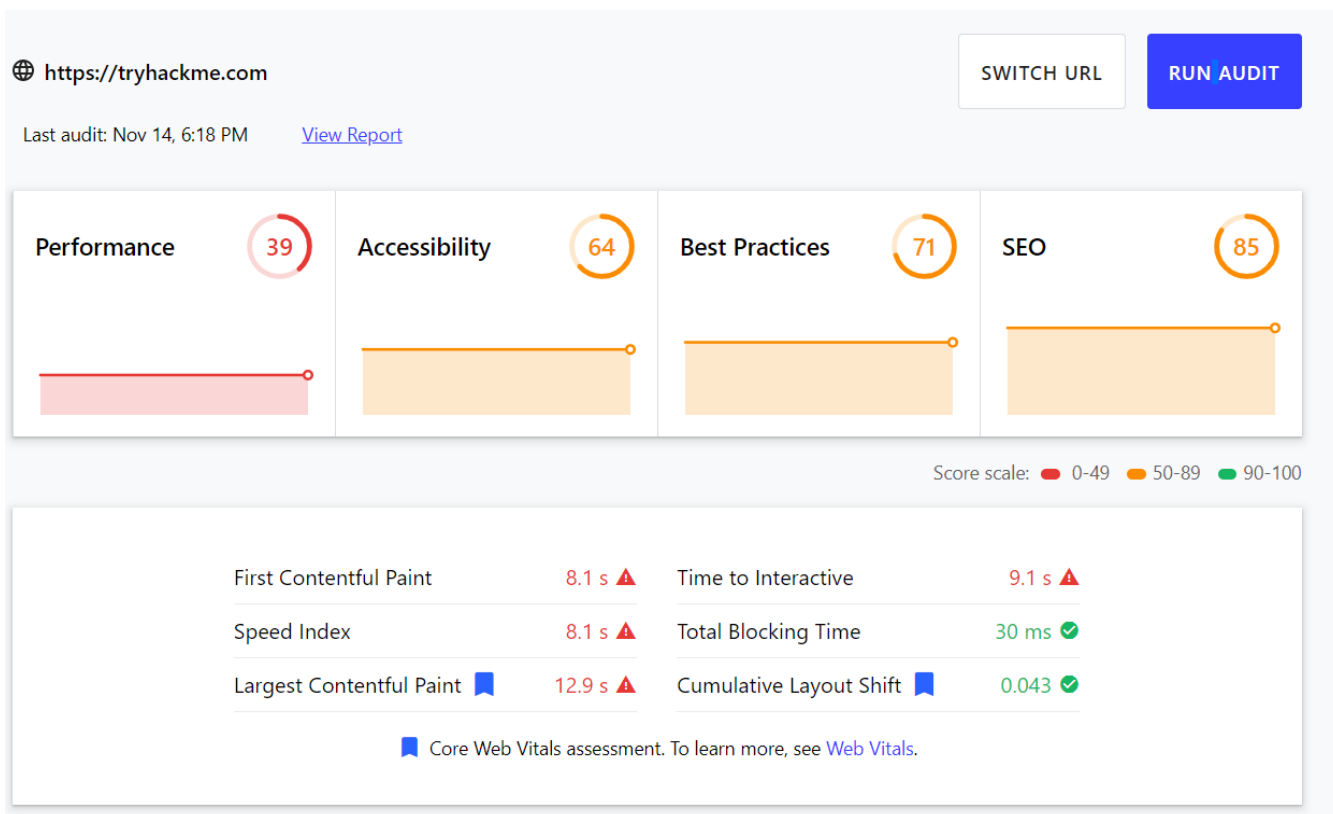
To highlight a few influences on how these points are scored, factors such as:

- How responsive your website is to the different browser types I.e. Google Chrome, Firefox and Internet Explorer - this includes Mobile phones!
- How easy it is to crawl your website (or if crawling is even allowed ...but we'll come to this later) through the use of "Sitemaps"
- What kind of keywords your website has (i.e. In our examples if the user was to search for a query like “Colours” no domain will be returned - as the search engine has not (yet) crawled a domain that has any keywords to do with “Colours”

There is a lot of complexity in how the various search engines individually "point-score" or rank these domains - including vast algorithms. Naturally, the companies running these search engines such as Google don't share exactly how the hierarchic view of domains ultimately ends up. Although, as these are businesses at the end of the day, you can pay to advertise/boost the order of which your domain is displayed.

There are various online tools - sometimes provided by the search engine providers themselves that will show you just how optimised your domain is. For example, let's use [Google's Site Analyser](#) to check the rating of [TryHackMe](#):





According to this tool, TryHackMe has an SEO rating of **85/100** (as of 14/11/2020). That's not too bad and it'll show the justifications as to how this score was calculated below on the page.

But...Who or What Regulates these "Crawlers"?

Aside from the search engines who provide these "Crawlers", website/web-server owners themselves ultimately stipulate what content "Crawlers" can scrape. Search engines will want to retrieve **everything** from a website - but there are a few cases where we wouldn't want **all** of the contents of our website to be indexed! Can you think of any...? How about a secret administrator login page? We don't want **everyone** to be able to find that directory - especially through a google search.

Introducing Robots.txt...

Use the same [SEO checkup tool](#) and other online alternatives to see how their results compare for <https://tryhackme.com> and <http://googledorking.cmnaic.co.uk>

Beepboop - Robots.txt

Robots.txt

Similar to "Sitemaps" which we will later discuss, this file is the first thing indexed by "Crawlers" when visiting a website.

But what is it?

This file must be served at the root directory - specified by the webserver itself. Looking at this file's extension of `.txt`, it's fairly safe to assume that it is a text file.

The text file defines the permissions the "Crawler" has to the website. For example, what type of "Crawler" is allowed (i.e. You only want Google's "Crawler" to index your site and not MSN's). Moreover, Robots.txt can specify what files and directories that we do or don't want to be indexed by the "Crawler".

A very basic markup of a Robots.txt is like the following:

```
1 User-agent: *
2 Allow: /
3
4 Sitemap: http://mywebsite.com/sitemap.xml
5
6
```

Here we have a few keywords...

Keyword	Function
User-agent	Specify the type of "Crawler" that can index your site (the asterisk being a wildcard, allowing all "User-agents")
Allow	Specify the directories or file(s) that the "Crawler" can index
Disallow	Specify the directories or file(s) that the "Crawler" cannot index
Sitemap	Provide a reference to where the sitemap is located (improves SEO as previously discussed, we'll come to sitemaps in the next task)

In this case:

1. Any "Crawler" can index the site
2. The "Crawler" is allowed to index the entire contents of the site
3. The "Sitemap" is located at <http://mywebsite.com/sitemap.xml>

Say we wanted to hide directories or files from a "Crawler"? Robots.txt works on a "blacklisting" basis. Essentially, **unless told otherwise**, the Crawler will index whatever it can find.

```
User-agent: *
Disallow: /super-secret-directory/
Disallow: /not-a-secret/but-this-is/

Sitemap: http://mywebsite.com/sitemap.xml
|
```

In this case:

1. Any "Crawler" can index the site

2. The "Crawler" can index every other content that isn't contained within `"/super-secret-directory/"`.

Crawlers also know the differences between sub-directories, directories and files. Such as in the case of the second "Disallow:" (`"/not-a-secret/but-this-is/"`)

The "Crawler" will index all the contents within `"/not-a-secret/"`, but will not index anything contained within the sub-directory `"/but-this-is/"`.

3. The "Sitemap" is located at <http://mywebsite.com/sitemap.xml>

What if we Only Wanted Certain "Crawlers" to Index our Site?

We can stipulate so, such as in the picture below:

```
1 User-agent: Googlebot
2 Allow: /
3
4 User-agent: msnbot
5 Disallow: /
6
7
```

In this case:

1. The "Crawler" "Googlebot" is allowed to index the entire site (`"Allow: /"`)
2. The "Crawler" "msnbot" is not allowed to index the site (`"Disallow: /"`)

How about Preventing Files From Being Indexed?

Whilst you can make manual entries for every file extension that you don't want to be indexed, you will have to provide the directory it is within, as well as the full filename. Imagine if you had a huge site! What a pain...Here's where we can use a bit of [regexing](#).

```
User-agent: *
Disallow: /*.ini$
Sitemap: http://mywebsite.com/sitemap.xml
```

In this case:

1. Any "Crawler" can index the site
2. However, the "Crawler" cannot index **any** file that has the extension of `.ini` within any directory/sub-directory using (`"$"`) of the site.

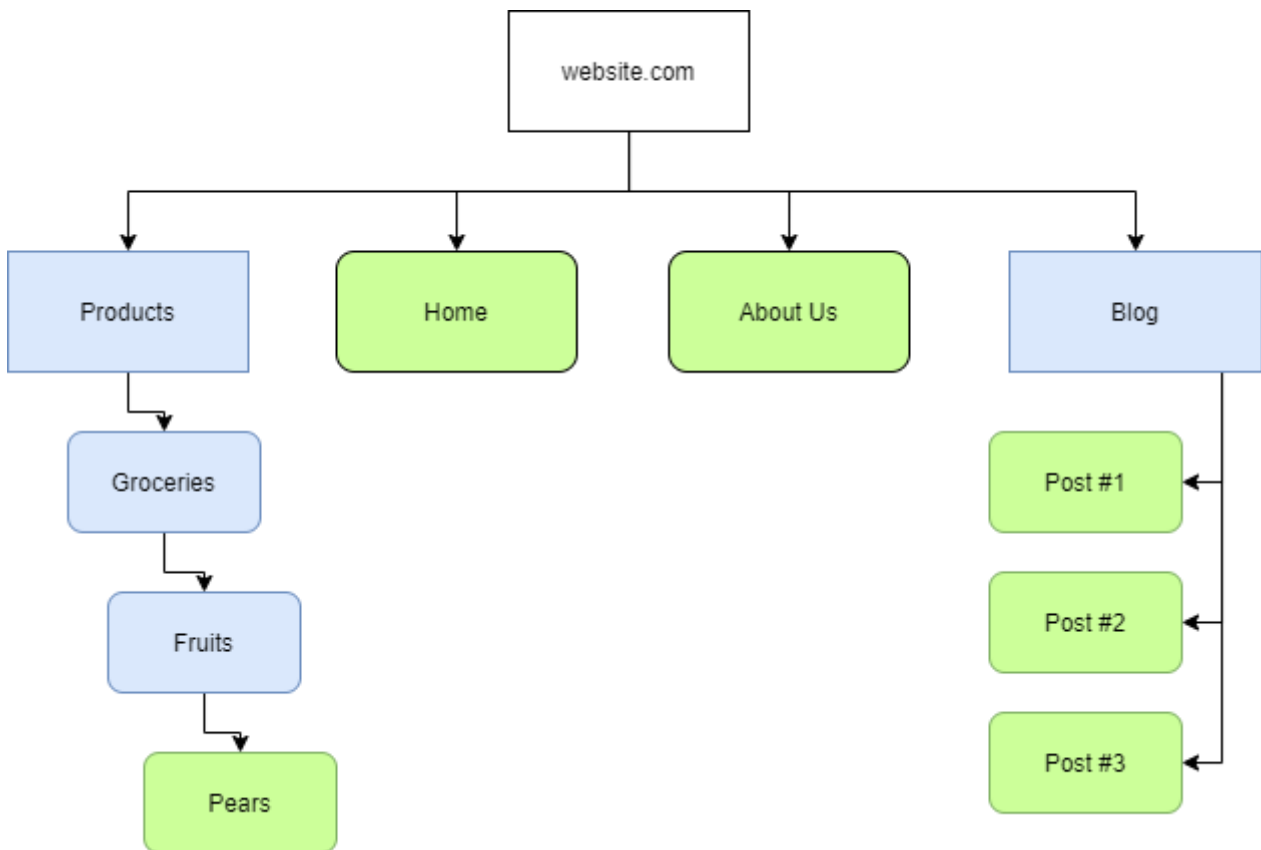
3. The "Sitemap" is located at <http://mywebsite.com/sitemap.xml>

Why would you want to hide a **.ini** file for example? Well, files like this contain sensitive configuration details. Can you think of any other file formats that might contain sensitive information?

Sitemaps

Comparable to geographical maps in real life, "Sitemaps" are just that - but for websites!

"Sitemaps" are indicative resources that are helpful for crawlers, as they specify the necessary routes to find content on the domain. The below illustration is a good example of the structure of a website, and how it may look on a "Sitemap":



The blue rectangles represent the **route** to nested-content, similar to a directory I.e. "Products" for a store. Whereas, the green rounded-rectangles represent an actual page. However, this is for illustration purposes only - "Sitemaps" don't look like this in the real world. They look something much more similar to this:

```

1 <?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="https://blog.cmntatic.co.uk/wp-content/plugins/
2 google-sitemap-generator/sitemap.xsl"?><!-- sitemap-generator-url="http://www.arnebrachhold.de" sitemap-generator-version="4.1.0" -->
3 <sitemapindex xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
4 http://www.sitemaps.org/schemas/sitemap/0.9/siteindex.xsd" xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"> <sitemap>
5   <loc>https://blog.cmntatic.co.uk/sitemap-misc.xml</loc>
6   <lastmod>2020-03-17T02:44:52+00:00</lastmod>
7 </sitemap>
8 <sitemap>
9   <loc>https://blog.cmntatic.co.uk/sitemap-tax-post_tag.xml</loc>
10  <lastmod>2020-03-17T02:44:52+00:00</lastmod>
11 </sitemap>
12 <sitemap>
13   <loc>https://blog.cmntatic.co.uk/sitemap-tax-category.xml</loc>
14   <lastmod>2020-03-17T02:44:52+00:00</lastmod>
15 </sitemap>
16 <sitemap>
17   <loc>https://blog.cmntatic.co.uk/sitemap-pt-post-2020-03.xml</loc>
18   <lastmod>2020-03-17T02:29:13+00:00</lastmod>
19 </sitemap>
20 <sitemap>
21   <loc>https://blog.cmntatic.co.uk/sitemap-pt-post-2020-02.xml</loc>
22   <lastmod>2020-03-16T18:47:14+00:00</lastmod>
23 </sitemap>
24 <sitemap>
25   <loc>https://blog.cmntatic.co.uk/sitemap-pt-page-2020-02.xml</loc>
26   <lastmod>2020-03-01T04:10:14+00:00</lastmod>
27 </sitemap>
28 </sitemapindex><!-- Request ID: 4e2205d5779bd2c538185ee5143bd0da; Queries for sitemap: 7; Total queries: 24; Seconds: 0.01; Memory for sitemap:
    0MB; Total memory: 6MB -->

```

"Sitemaps" are XML formatted. I won't explain the structure of this file-formatting as the room [XXE](#) created by [falconfeast](#) does a mighty fine job of this.

The presence of "Sitemaps" holds a fair amount of weight in influencing the "optimisation" and favorability of a website. As we discussed in the "Search Engine Optimisation" task, these maps make the traversal of content much easier for the crawler!

Why are "Sitemaps" so Favourable for Search Engines?

Search engines are lazy! Well, better yet - search engines have a lot of data to process. The efficiency of how this data is collected is paramount. Resources like "Sitemaps" are extremely helpful for "Crawlers" as the necessary routes to content are already provided! All the crawler has to do is scrape this content - rather than going through the process of manually finding and scraping. Think of it as using a wordlist to find files instead of randomly guessing their names!

The easier a website is to "Crawl", the more optimised it is for the "Search Engine"

What is Google Dorking?

Using Google for Advanced Searching

As we have previously discussed, Google has a lot of websites crawled and indexed. Your average Joe uses Google to look up Cat pictures (I'm more of a Dog person myself...). Whilst Google will have many Cat pictures indexed ready to serve to Joe, this is a rather trivial use of the search engine in comparison to what it can be used for. For example, we can add operators such as that from programming languages to either increase or decrease our search results - or perform actions such as arithmetic!



12 + 1



All

Maps

News

Images

Shopping

More

Settings

Tools

About 25,270,000,000 results (0.55 seconds)



A privacy reminder from Google

[REMIND ME LATER](#)

[REVIEW](#)



12 + 1 =

13

Say if we wanted to narrow down our search query, we can use quotation marks. Google will interpret everything in between these quotation marks as exact and only return the results of the exact phrase provided...Rather useful to filter through the rubbish that we don't need as we have done so below:

"american pscho poster"



All

Images

Shopping

News

Videos

More

Settings

Tools

About 46,100 results (0.44 seconds)

Showing results for "american **psycho** poster"

Search instead for "american pscho poster"

Images for "american psycho poster"



→ More images for "american psycho poster"

Report images

www.redbubble.com > Wall Art > Poster ▾

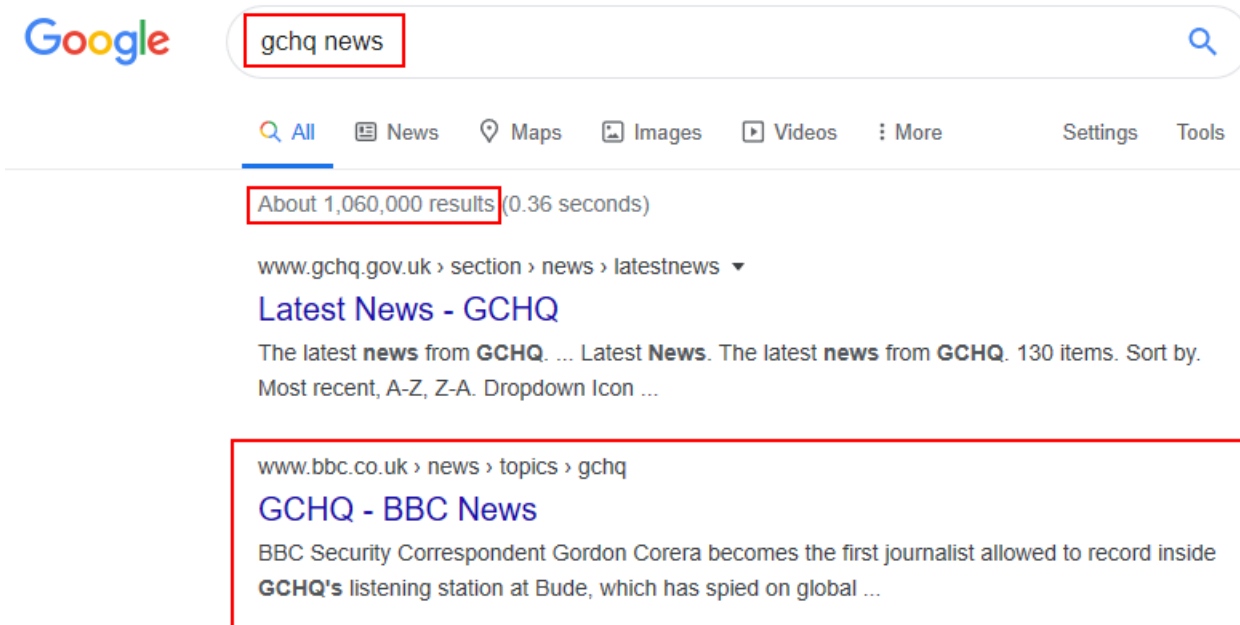
American Psycho Posters | Redbubble

patrick bateman, american psycho, phone, cellphone, cell, christian, bale, christian bale, dubs, checkem. Patrick Bateman on Phone (**American Psycho**) Poster.

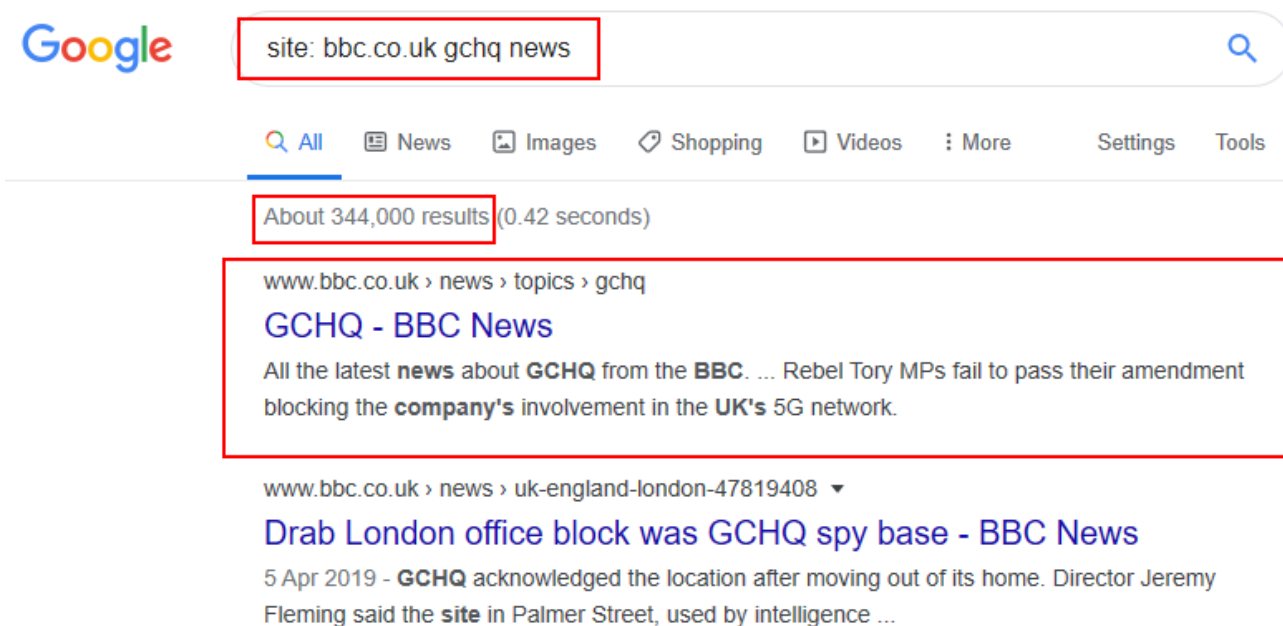
Refining our Queries

We can use terms such as "**site**" (such as bbc.co.uk) and a query (such as "gchq news") to search the specified site for the keyword we have provided to filter out content that may be harder to find otherwise. For example, using the "site" and "query" of "bbc" and "gchq", we have modified the order of which Google returns the results.

In the screenshot below, searching for "gchq news" returns approximately 1,060,000 results from Google. The website that we want is ranked behind GCHQ's actual website:



But we don't want that...We wanted "bbc**.co.**uk" first, so let's refine our search using the "site" term. Notice how in the screenshot below, Google returns with much fewer results? Additionally, the page that we didn't want has disappeared, leaving the site that we did actually want!



Of course, in this case, GCHQ is quite a topic of discussion - so there'll be a load of results regardless.

So What Makes "Google Dorking" so Appealing?

First of all - and the important part - it's legal! It's all indexed, publicly available information. However, what you do with this is where the question of legality comes in to play...

A few common terms we can search and combine include:

Term	Action
filetype:	Search for a file by its extension (e.g. PDF)
cache:	View Google's Cached version of a specified URL
intitle:	The specified phrase MUST appear in the title of the page

For example, let's say we wanted to use Google to search for all PDFs on bbc.co.uk:

site:bbc.co.uk filetype:pdf

Google

site:bbc.co.uk filetype:pdf

Q All Images News Shopping Maps More Settings Tools

About 46,300 results (0.34 seconds)

downloads.bbc.co.uk › london pdf

XXXX Dear XXXX, RE: Freedom of Information Request ... - BBC
 5 Jan 2011 - The detailed information on services outside the Top 10, relating to those services and passengers in excess of capacity, that is being withheld ...

downloads.bbc.co.uk › commissioning › site › pasc1 PDF

BBC PasC
 20 Mar 2002 - PDU PRODUCTIONS (AS ABOVE) ?? State where relevant whether the Producer/Director is Continuing Staff (CS) Guest Staff (GS) or Short ...

downloads.bbc.co.uk › spanish › manual_biodigestor PDF Translate this page

biodigestor - Producción Animal
 by RB Botero - Cited by 72 - Related articles
 Se resume la experiencia adquirida por los autores durante la instalación y puesta en funcionamiento de biodigestores del tipo Taiwán (flujo continuo) Estos se.

www.bbc.co.uk › oxford › glyme PDF


Glyme Valley Way - Oxfordshire Cotswolds
 The Glyme Valley Way was devised by BBC Oxford and Oxfordshire County Council's Countryside Service as part of Oxfordshire 2007 which is celebrating a ...

Great, now we've refined our search for Google to query for all publically accessible PDFs on "**bbc.co.uk**" - You wouldn't have found files like this "Freedom of Information Request Act" file from a wordlist!

Here we used the extension **PDF**, but can you think of any other file formats of sensitive nature that **may** be publically accessible? (Often unintentionally!!) Again, what you do with any results that you find is where the legality comes into play - this is why "Google Dorking" is so great/dangerous.

Here is simple directory traversal.

I have blanked out a lot of the below to cover you, me, THM and the owners of the domains:



[All](#)

[Images](#)

[News](#)

[Videos](#)

[Books](#)

[More](#)

[Settings](#)


[Tools](#)

Index of /downloads

Index of

Index of /

Index of /

Name	Last modified	Size	Description
<hr/>			
Parent Directory		-	
			
<hr/>			