How to Use Regex for SEO & Website Data Extraction by @RedmontBham

One technology skill that is too-often ignored is regex.

Regex searching and extraction is perhaps the single most useful, yet largely unknown skill, for web operations, SEO practitioners, web analytics teams, researchers, community managers, and digital marketers of all types.

This article will explain what regex is and how to get started using it with web crawlers.

In addition, you'll also learn five interesting real-world regex use cases.

What is Regex?

Regex, or Regular Expressions, is essentially a tool for pattern matching. Regex is a staple of search engines, find and replace utilities, and is a native or add-on capability of many programming languages.

Combining "Find" with a web crawler can be extremely powerful for identifying errors and extracting data.

This feature is still relatively new to commercially available SEO tools.

Custom extraction, using regex (or Xpath or CSSpath), was only added to Screaming Frog SEO Spider in the July 2015 release. Up until that time, SEOTools for Excel was perhaps the most accessible tool for regex extraction – but only when combined with a crawler like Screaming Frog to first collect URLs to analyze.

Google Sheets has also long had the ability to use Regex extraction but, like SEOTools for Excel, requires a user to crawl the site with another tool first.

After Screaming Frog released their custom extraction feature in 2015, the “enterprise” level SEO tools like BrightEdge and Conductor responded by adding similar features.

Despite the now widespread availability of regex functionality, the power of this feature is still largely ignored by many web professionals.

How You Can Learn Regex

I suggest you dive right in with a combination of Screaming Frog and tools like:

Try out some of the below use cases, or come up with your own.

Regex Use Cases

Here are a few examples of how this comes together – with real-world use cases that you can replicate today using the provided regex formulas.

1. Extract Metadata (or Any Data) from Any Website

For this example, let’s look at ESPN.

As I was writing this article, I saw this story about NBA star Blake Griffin.

Viewing Source of this page, I’m able to locate this bit of analytics-related metadata (similar metadata is found on most major websites):

{"omniture":{"columnist":"lowe_zach","league":"nba","countryRe

gion":"en-
us","hier1":"nba:story","section":"nba","source":"espn.com","p
ageName":"nba:story","storyInfo":"22258759+zach-lowe-blake-
griffin-trade-future-la-clippers-detroit-
pistons","sections":"nba:story","site":"espn","premium":"premi
um-
no","convrSport":"basketball","pageURL":"www.espn.com/nba/stor
y/_/id/22258759/zach-lowe-blake-griffin-trade-future-la-
clippers-detroit-
pistons","lang":"en_us","prop35":"2018-01-30","contentType":"s
tory","sport":"basketball","account":"wdgespcom","siteType":"f
ull","prop58":"isIndex=false"}

Now that we've identified what is likely a common data pattern, we can begin to construct the regex to extract this data.

Here I will copy and paste this code, along with a few lines before and after, into regexr.com (my preferred regex tool) and begin to work on some formulas:



Suppose I'm auditing the ESPN website and I want to get a list of all news stories, authors, and dates. I simply need to fire up the crawler with these 2 regular expressions included in the Custom Extraction filter:

```
"columnist":"(.*?)"
"prop35":"(.*?)"
```



And it works! I'm able to crawl a list of ESPN.com URLs, or crawl the entire site, and extract this information that will be critical for my content audit.

2. Trademark Enforcement

The registered trademark symbol (®) is a symbol that typically must appear after a trademarked word or words, in the first usage on a webpage.

A good example of this is "ITIL®" – which is an initialism that stands for IT Infrastructure Library, a trademark owned by Axelos.

Many websites that discuss ITIL fail to include the registered trademark symbol when referencing the concept.

One such website that discusses ITIL but may not always include the registered trademark symbol, is Cherwell.com.

Using two simple regular expressions, we can easily find the URLs in which Cherwell should consider adding a registered trademark symbol:



In this case, we've spotted four URLs that *do* include the correct symbol – and many more that do not.

This example shows just how powerful a simple regex string can be.

Here are the regular expressions for this use case:

```
ITIL®
ITIL(?!®)
```

Changing Names of Products & Finding Incorrect Capitalizations

In 2017, IBM officially changed the name of "DB2" to "Db2" – making the "b" lowercase.

The name DB2 was first used in 1983, so there are likely countless places on the web containing the incorrect usage.

This use case is common – either a word or phrase has a common mis-capitalization, r a product name changes.

In either case, the solution is to crawl and find the incorrect or outdated usages.

One example of a website that uses “DB2” frequently, and therefore ought to consider auditing their site to correct this usage, is Delphix.

Using the following regular expression, we are able to identify all incorrect instances of “DB2”, “db2”, or “dB2”:

(db|DB|dB)2



4. Find Uploaded Files on a Communities Site

Jive Software is perhaps the #1 Communities platform for enterprises. As such, it is used by companies like Cisco, ServiceNow, Adobe, BMC Software, McAfe, Wiley and many other major companies.

You can find other examples of companies using Jive’s platform with a Google query, such as: https://www.google.com/search?q=inurl:hosted.jivesoftware.com

Jive, much like WordPress or other web publishing platforms, has certain patterns that are typically followed.

One such pattern for Jive is that all uploads to the platform, when viewed from the document or blog that it is uploaded to, includes the following link:

<a class=”j-attachment-icon” href=”/servlet/JiveServlet/download/xxx/filename.ext”>

Therefore, we can use the following regex pattern to extract

all attachment URLs from any Jive hosted communities site:

"j-attachment-icon" href="(.*?)"

To see how this works, here's the results of briefly scanning the website https://community.servicenow.com:



Spell Check

Hopefully most web publishers are checking spelling either in Word or through other means before they publish, but common misspellings still slip through the cracks.

Not to worry, though.

Regex + Screaming Frog can help you identify commonly misspelled words. The only catch is, you need to know the commonly misspelled words.

Certain words are commonly misspelled in any given language.

In addition, there are words specific to an industry or brand which should be included along with the list of language-specific common misspellings.

Here is a simple regex string to match the most common U.S. English misspelled words:

(?i)accomodate | accomodation | acheive | accross | agressive | agression | apparantly | appearence | arguement | assasination | basicly | begining | beleive | belive | bizzare | buisness | calender | Carribean | cemetary | chauffer | collegue | comming | commitee | completly | concious | curiousity | definately | dilemna | dissapear | dissapoint | ecstacy | embarass | enviroment | existance | Farenheit | familar | finaly | florescent | foriegn | forseeable | fourty | foward | freind | futher | jist | glamourous | goverment | gaurd | happend | harrass | harrassment | honourary |

humourous | idiosyncracy | immediatly | incidently | independant | interupt | irresistable | knowlege | liase | liason | lollypop | millenium | millenia | Neandertal | neccessary | noticable | ocassion | occassion | occured | occuring | occurance | occurence | pavillion | persistant | pharoah | peice | politican | Portugese | posession | prefered | prefering | propoganda | publically | realy | recieve | refered | refering | religous | rember | remeber | resistence | sence | seperate | seige | succesful | supercede | suprise | tatoo | tendancy | therefor | threshhold | tommorow | tommorrow | tounge | truely | unforseen | unfortunatly | untill | wierd | whereever | wich

Let's fire up Screaming Frog once more and see if we can locate any of these errors on CNN:



In no time at all, we've already spotted a misspelling of the word "Fahrenheit":



Summary

Patterns are all over the web. Using regex, you can locate instances of patterns and extract associated data.

Regular expressions are not simple to master, but once you get started you'll find endless uses for this powerful technology. It was in this vein that Jamie Zawinski famously stated:

> *"Some people, when confronted with a problem, think "I know, I'll use regular expressions." Now they have two problems."*

http://tracking.feedpress.it/link/13962/8272051