Canonical form of a domain (or URL)

URLs of resources are stored in the database in their canonical form, which means that before requesting a list of categories, the URL needs to be canonicalized. In other words, you need to get the canonical representation of the URL.

A single URL resource pointer consists of several parts, but we are only interested in two of them:

- Domain name (`domain`)
- Path (`path`)

Here is a random URL as an example:
`directory.google.com/example/test.php?key=value&one=1`

Where the domain is:
`directory.google.com`

and this is the URL path:
`/example/test.php?key=value&one=1`

Since the URL points to a resource, it can be written in various forms and in a variety of ways. For a direct search in the database, it is necessary to bring all the various forms of URLs pointing to one resource to one canonical form. This is very important for getting the correct categorization of the requested URL.

We use the standard version of URL canonicalization from the Google Safe Browsing project
https://developers.google.com/safe-browsing/ with the API
version higher 2.

This project describes techniques for resource identifier canonicalization, examples and algorithms can be found on the page:
https://developers.google.com/safe-browsing/v4/urls-hashing#canonicalization

---

Examples of canonicalization

Source URL	Canonical URL
http://evil.com/foo-bar-baz	http://evil.com/foo

http://host/%25%32%35	http://host/%25
http://evil.com/foo-bar-baz	http://evil.com/foo
http://test.com/path/../	http://test.com/

Canonicalization in the context of the url2cat library

As described above, you can get a canonical form of a URL. But due to the nature of URLs, you cannot rely on one URL. To find the best match for it in the database, you need to generate derived URLs by alternately discarding parts of the primary URL from the left and right edges.

This is a random URL as an example:
`directory.google.com/example/test.php?key=value&one=1`

Derived URLs will be as follows:

`directory.google.com/example/test.php`
`directory.google.com/example/`
`directory.google.com/`
`google.com/example/test.php?key=value&one=1`
`google.com/example/test.php`
`google.com/example/`
`google.com/`

The resulting URLs must be checked in the database.

Created 2022-08-22 11:52:44 UTC
Updated 2022-08-22 12:17:30 UTC