Check Github License

Christian Külker

2023-03-18

Contents

The website https://github.com usually shows the license next to a git repository. Sometimes this information shows "other" and the license is not known to github.com. This document briefly describes how github.com detects licenses and how to verify (**without guarantee**) that a project license is detected.

1   The Quick Guide

1. Choose a recent operating system, such as Debian Bullseye
2. Install the `licensee` dependencies: `aptitude install ruby-rugged`
3. Install `licensee` using the Ruby gem: `gem install --user-install licensee`
4. Run `licensee` on the `LICENSE` file (not the uncommitted changes repository): `cd repository;licensee detect LICENSE`.
5. If files like `COPYING` and `LICENSE` are properly selected, dual-licensed projects are also possible: "CC-BY-SA-4.0, GFDL-1.3 licenses found".

2   The longer Explanation

1. When a license is selected during project creation, github.com automatically creates a license file named `LICENSE`. Selectable licenses are collected in the `_licenses` directory of the choosealicense.com repository: https://github.com/github/choosealicense.com/tree/gh-pages/_licenses. These licenses should be recognized. The license file from github.com contains metadata about the license, such as an id called `spdx`. Only licenses with an id will be detected. See https://github.com/spdx/license-list-data and https://github.com/spdx/license-list-XML for details.

2. Sometimes a license is added later with a `LICENSE` file, or the license file has a different name. For example, `COPYING` is the `GNU` standard. I do not know which files are scanned (other than `COPYING` and `LICENSE`), but to be sure it might be advisable to rename the file to either `LICENSE` or `COPYING`. If both files exist, the project will display both licenses: "CC-BY-SA-4.0, GFDL-1.3 licenses found".

3. The content of the `LICENSE` file should be **text** with **78** characters per line. The best way to ensure this is to copy the contents of the license file found in the `_licenses` directory to `LICENSE`.

4. The Licensee Ruby gem https://github.com/licensee/licensee is used to query the license. To use it, the following requirements are necessary:

   - Recent Ruby version (2.7 for example, 2.3 will not work) (faraday-net_http requires Ruby version >= 2.6.0.)

   - A recent version of Debian (Bullseye will work; Stretch will not, even from package). An older Debian could use `aptitude insall ruby-licensee` to install `licensee`, but newer licenses may not be recognized and this binary seems to broken on Debian Stretch.

   - Even if you are considering installing from source, it is convenient to install the heavy dependencies from package, install as root using the command `aptitude install ruby-rugged`.

   - Use the `gem install` command.

   - Use the `--user-install` switch or add `gem: --user-install` to `~/gemrc` (tested) or use the `root` user (untested).

   - Add `~/.local/share/gem/ruby/2.7.0/bin` to your `PATH` if you are using Ruby 2.7.

   - If you want to install `licensee` from source. The dependencies are at

least `rugby-dev`, `cmake` and maybe `libevent-pthreads-2.1-7`, but surely other dependencies need to be installed as well. This path has not been explored further on older systems or without installing `ruby-rugged`.

- After running `gem install licensee`, the `licensee` binary can be found in `~/.local/share/gem/ruby/2.7.0/bin` or, if installed via package, in `/usr/bin/licensee`.

5. For example, in the <github.com/ckuelker/quick-guide-en-us> repository, the license file is named `LICENSE.markdown` and the content is GFDL-1.3 in markdown format. The original idea was to use this file for a CMS to display the license, but this idea was later abandoned. However, https://github.com reports the license as `View License`. So while `licensee` can detect the license, either the 99.16% accuracy is not high enough or the `*.markdown` extension prevents automatic detection.

```
cd github.com/ckuelker/quick-guide
~/.local/share/gem/ruby/2.7.0/bin/licensee detect .
License:        GFDL-1.3
Matched files:  LICENSE.markdown
LICENSE.markdown:
  Content hash:  82840a8d2e4f2a671f952b01b4971fa8a0ca4082
  Confidence:    99.16%
  Matcher:       Licensee::Matchers::Dice
  License:       GFDL-1.3
  Closest non-matching licenses:
    GFDL-1.3 similarity:  99.16%
    LGPL-2.1 similarity:  36.80%
    ODbL-1.0 similarity:  31.60%
```

6. Other projects have a better match and are recognized by https://github.com, such as <github.com/ckuelker/gitolite-tools>.

```
cd github.com/ckuelker/gitolite-tools
~/.local/share/gem/ruby/2.7.0/bin/licensee detect .
License:        GPL-3.0
Matched files:  LICENSE
LICENSE:
  Content hash:  7d4cdf499d39e2e1ce27b2878e22872f0f5a74dd
  Confidence:    100.00%
  Matcher:       Licensee::Matchers::Exact
  License:       GPL-3.0
```

7. Even if the file is not `LICENSE.markdown` but `LICENSE` and the content is mark-
   down but not text, https://github.com will not detect it.

```
cd github.com/ckuelker/www.pankyll.org
License:        GFDL-1.3
Matched files:  LICENSE
LICENSE:
  Content hash:  82840a8d2e4f2a671f952b01b4971fa8a0ca4082
  Confidence:    99.16%
  Matcher:       Licensee::Matchers::Dice
  License:       GFDL-1.3
  Closest non-matching licenses:
    GFDL-1.3 similarity:  99.16%
    LGPL-2.1 similarity:  36.80%
    ODbL-1.0 similarity:  31.60%
```

After changing the format of the `LICENSE` file from markdown to text, the detec-
tion of the project is still 99.16%. My guess is that it is checking against the reposi-
tory and not against the uncommitted changes.

```
cd github.com/ckuelker/www.pankyll.org
~/.local/share/gem/ruby/2.7.0/bin/licensee detect .
License:        GFDL-1.3
Matched files:  LICENSE
LICENSrE:
  Content hash:  82840a8d2e4f2a671f952b01b4971fa8a0ca4082
  Confidence:    99.16%
  Matcher:       Licensee::Matchers::Dice
  License:       GFDL-1.3
  Closest non-matching licenses:
    GFDL-1.3 similarity:  99.16%
    LGPL-2.1 similarity:  36.80%
    ODbL-1.0 similarity:  31.60%
```

When a file is explicitly requested to be scanned, the scanning rate changes.

```
cd github.com/ckuelker/www.pankyll.org
~/.local/share/gem/ruby/2.7.0/bin/licensee detect LICENSE
License:        GFDL-1.3
Matched files:  LICENSE
LICENSE:
  Content hash:  164a858691ea0a6fb0dd06c5ca00e5dd7620eef8
  Confidence:    100.00%
  Matcher:       Licensee::Matchers::Exact
```

```
License:        GFDL-1.3
```

8. After committing the change to https://github.com/ckuelker/www.pankyll.org, the license value changed to "GFDL-1.3 license". So we can conclude that 99.16% is not enough. A subsequent git pull on another machine showed that project detection worked with the updated `LICENSE` file.

```
cd github.com/ckuelker/www.pankyll.org
~/.local/share/gem/ruby/2.7.0/bin/licensee detect .
License:        GFDL-1.3
Matched files:  LICENSE
LICENSE:
  Content hash:  164a858691ea0a6fb0dd06c5ca00e5dd7620eef8
  Confidence:    100.00%
  Matcher:       Licensee::Matchers::Exact
  License:       GFDL-1.3
```

9. These experiments are not complete, but may help.

3   History

Version	Date	Notes
0.1.2	2023-03-18	Improve writing
0.1.1	2023-02-25	Dual-license information
0.1.0	2023-02-24	Initial release

4   Disclaimer of Warranty

THERE IS NO WARRANTY FOR THIS INFORMATION, DOCUMENTS AND PROGRAMS, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE INFORMATION, DOCUMENT OR THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE INFORMATION, DOCUMENTS AND PROGRAMS IS WITH YOU. SHOULD THE INFORMATION, DOCUMENTS OR PROGRAMS PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

5   Limitation of Liability

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE INFORMATION, DOCUMENTS OR PROGRAMS AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE INFORMATION, DOCUMENTS OR PROGRAMS (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED

INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE INFORMATION, DOCUMENTS OR PROGRAMS TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.