December 2021

Open source license text matching and reporting

Armijn Hemel

Open source license text matching and reporting

Abstract

When using open source software it is important to find out under which open source license the software was released under, as this determines what can and cannot be done with the software. There are many open source software licenses with different license terms that are not always compatible with each other. Different pieces of software released under incompatible software licenses cannot be combined with each other. It is therefore necessary to find out which licenses are declared in source code and correctly report these.

Source code repositories or releases that are open source licensed almost always contain a text file with the text of the license that the code has been released under, such as the GNU General Public License (various versions), the Apache License (various versions), and so on.

While open source license texts are meant to be immutable, they are frequently changed. Many times the changes are purely cosmetic, but sometimes the license are changed in such a way that the changes could affect the meaning of the license. It is important to be able to detect such changes.

In this article a very lightweight method for comparing license texts found in source code archives with official license texts is presented.

Keywords

open source, license texts, scanning

Background

Most open source license texts that can be found in source code archives or source code repositories are verbatim unchanged copies of the license text, but sometimes there are changes. Reasons for changes are:

- cosmetic changes
  - reformatting lines to fit within a certain amount of characters
  - different line endings on different platforms
- adding extra headers/footers with project specific information
- accidental changes, like global “search and replace” having gone wrong
- adding extra clauses to the license text (effectively changing the license text)

Whatever the reason for the changes it is important to detect changes, as some of these changes might change the meaning of the license text. Many license scanners focus on detecting the type of license, without looking at whether or not the license has been changed, making it easy to miss if clauses were changed or if any clauses were added or removed.

Some licenses were also changed by the license stewards themselves, creating different versions of licenses. A good example is version 2 of the GNU General Public License, of which there are at least

16 official versions that all slightly vary[1] and of which there are countless hybrids of the different versions.

Method

The method describes a simple method to detect if license texts are different from the official license texts or not. It first tokenizes the license texts and then compares the tokens to official license texts to determine the closest match.

1. tokenize the official license texts to get rid of whitespace and line endings
2. store the list of tokens in the same order as in which they appear in the original license text
3. for each license text that has to be compared to the official license texts do:
    a. tokenize the license text to get rid of whitespace and line endigs
    b. store the list of tokens in the same order as in which they appear in the license text
    c. compare the list of tokens to the stored list of tokens of the license text. For each of the stored official license texts do:
        1. determine if the list of tokens of the scanned license text is the same as the offical license text. If so, report and continue with the next license text, as it is an exact match.
        2. determine if the list of tokens of the scanned license text is a subset of the tokens of the official license text, taking the order of tokens into account. If so, the license text has been modified and parts of the original text have been left out.
        3. if the list of tokens of the scanned license text is not a subset of the tokens of the official license text determine if it is a superset and all of the tokens of the original license text are found in the scanned license text, taking the order of tokens into account. If so, there is extra text in the license text.
        4. if the license is not an exact match (subset, superset or neither) determine the difference between the two sets. Store the difference between the original license text and the scanned license text.
    d. report the token difference of the closest match for license texts that are not an exact match.
    e. report the closest match for the license text (exact matches and not exact matches)

References

[1] GPL History project: https://github.com/pombredanne/gpl-history/