Trust Source

DeepScan - Analysis of license texts

Grigory Markin, EACG GmbH

Agenda

Why do we analyse license texts ?

License text analysis

- License recognition based on a license text
- Extraction of copyright information
- Modification of original license texts

Analysis of comments in source code

1. Copyright clauses
2. License keys
3. SPDX keys
4. Links to license texts

Agenda

Trust Source

Why do we analyse license texts ?

License text analysis

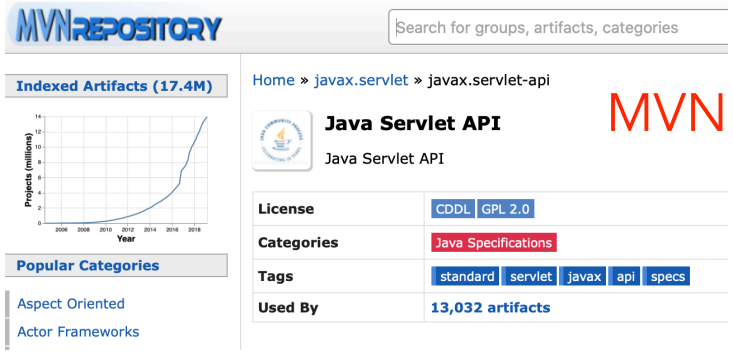
- License recognition based on a license text
- Extraction of copyright information
- Modification of original license texts

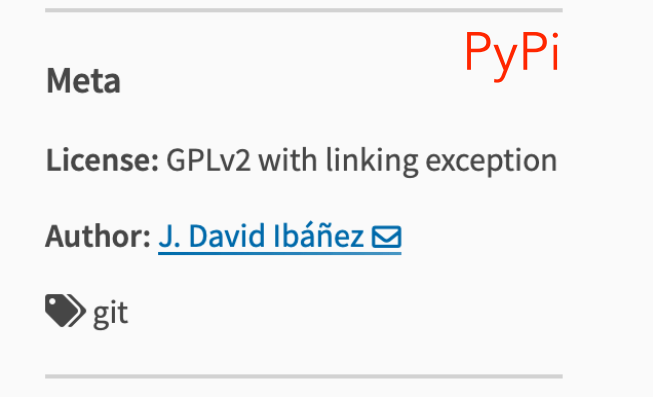
Analysis of comments in source code

1. Copyright clauses
2. License keys
3. SPDX keys
4. Links to license texts

License information sources for open source components

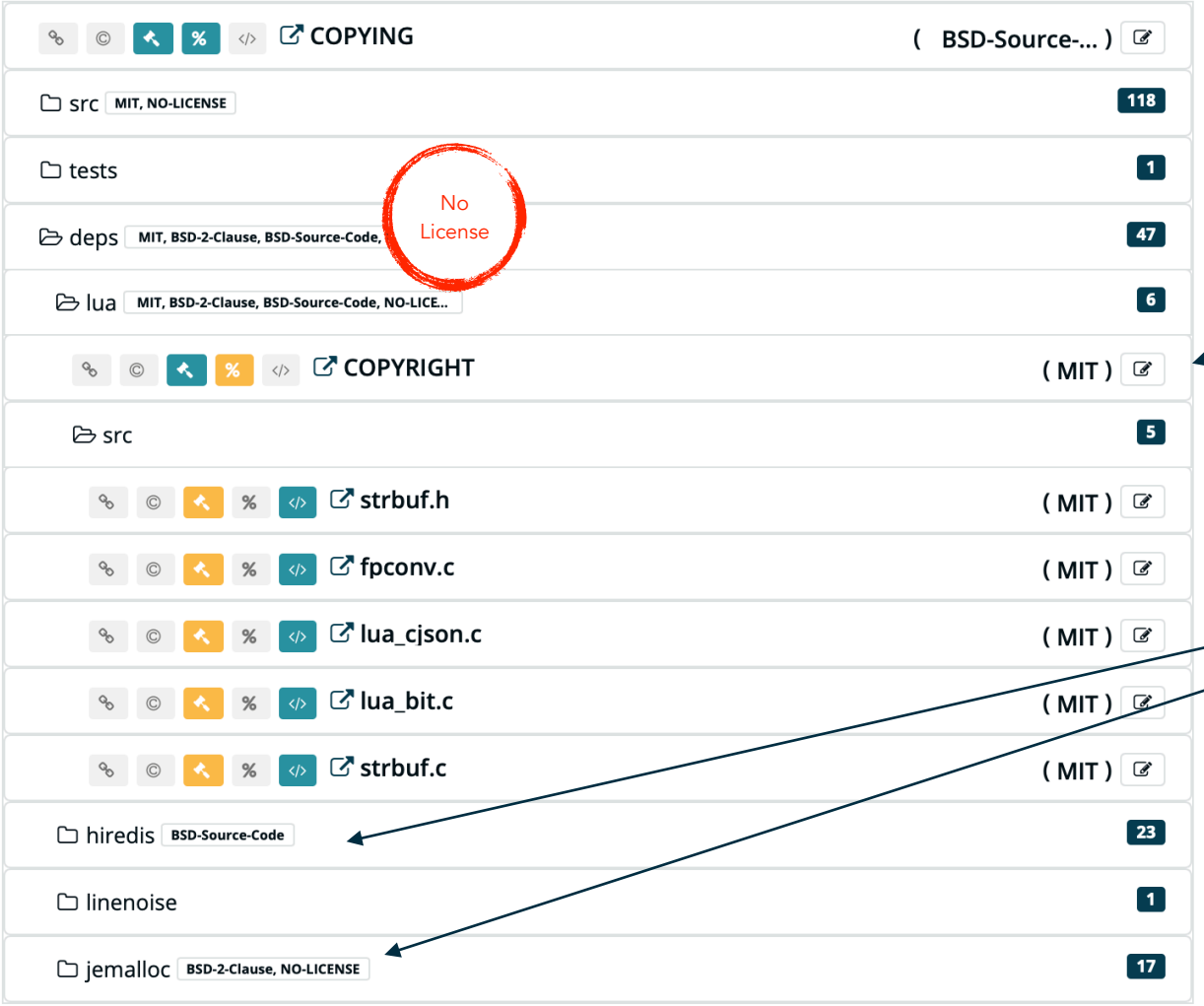
- Packagemanagers (Maven, NuGet, NPM, PyPI, SPM, etc.)
  - Often available but can contain old information
  - Declared license can differ from the actual license
- Sourcecode
  - License file
  - License information in comments (License key, SPDX, URL)
  - License notice in README





Why do we analyse license texts ?

Example: https://github.com/antirez/redis.git



BSD-3-Clause

MIT within an external component (Lua)

Further licenses in the external components

Screen Shot taken from TrustSource DeepScan: https://app.trustsource.io

Agenda

Why do we analyse license texts ?

License text analysis

- License recognition based on a license text
- Extraction of copyright information
- Modification of original license texts

Analysis of comments in source code

1. Copyright clauses
2. License keys
3. SPDX keys
4. Links to license texts

License recognition based on a license text

- Given a text, our goal is to determine whether it corresponds to a license text from our database
- 100% match is for most license impossible (e.g. modified copyright clause)
- There are licenses that differ from each other minimally (MIT vs. JSON)
- One text compared to another can be structured differently but be semantically 100% identical (Move parts around)
- Detect text manipulations
- Hence, consider similarity analysis

MIT:

```
Copyright <YEAR> <COPYRIGHT HOLDER>

Permission is hereby granted, free of charge, to any person obtaining a copy of this
software and associated documentation files (the "Software"), to deal in the Software
without restriction, including without limitation the rights to use, copy, modify,
merge, publish, distribute, sublicense, and/or sell copies of the Software, and to
permit persons to whom the Software is furnished to do so, subject to the following
conditions:

The above copyright notice and this permission notice shall be included in all copies
or substantial portions of the Software. …
```

The JSON License:

```
Copyright (C) 2020 JSON.org

Permission is hereby granted, free of charge, to any person obtaining a copy of this
software and associated documentation files (the "Software"), to deal in the Software
without restriction, including without limitation the rights to use, copy, modify,
merge, publish, distribute, sublicense, and/or sell copies of the Software, and to
permit persons to whom the Software is furnished to do so, subject to the following
conditions:

The above copyright notice and this permission notice shall be included in all copies
or substantial portions of the Software.

The Software shall be used for Good, not Evil.   …
```

License recognition based on a license text

- Text preparation
    1. Remove all special characters
    2. All words are lower cased and in infinitive
    3. Assign an unique number to every word
    4. Compute hash for all preprocessed license texts
- Compare based on hashes -> 100% match
- Otherwise, compute **Sørensen–Dice Coefficient (DSC)**
- All texts with DSC > 0.85 are seen as candidates
- Text with the largest DSC is chosen

- **Sørensen–Dice Coefficient (DSC)**

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

- Describes similarity between sets X and Y
    - X - set of words from the text 1
    - Y - set of words from the text 2
- DSC is between 0 and 1
    - 0 - nothing similar between two texts
    - 1 - two texts consist of the same set of words

License recognition based on a license text

- **Possible optimisations**
  - Consider weights for words (some words are less important than others)
  - Consider sentences (sets of sets of words -> moving a word from one sentence to another can significantly change the text semantics)

- **Current state**
  - The DSC-based approach is very simple and fast but also delivers surprisingly good results without additional optimisations
  - Adding additional computation steps would increase the complexity
  - The current state is a good preliminary step for further analysis

License recognition based on a license text

- **Current work**
  - Identify modifications made to an original license text
  - Simple diff computation between texts is almost useless
  - Make use of natural language processing techniques
    - Identify sentences boundaries
    - Rule out "common" sentences
  - DSC-based approach is good for choosing candidates to detailed analysis
- **Future work**
  - Determine legal information in plain texts (README files)

The JSON License:

```
Copyright (c) 2002 EACG GmbH

Permission is hereby granted, free of charge, to any person obtaining a
copy of this software and associated documentation files (the
"Software"), to deal in the Software without restriction, including
without limitation the rights to use, copy, modify, merge, publish,
distribute, sublicense, and/or sell copies of the Software, and to
permit persons to whom the Software is furnished to do so, subject to
the following conditions:

The above copyright notice and this permission notice shall be included
in all copies or substantial portions of the Software.

The Software shall be used for Good, not Evil or any religious
undertakings.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS
OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF
MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT.
IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY
CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT,
TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE
SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.
```

Agenda

Why do we analyse license texts ?

License text analysis

- License recognition based on a license text
- Extraction of copyright information
- Modification of original license texts

Analysis of comments in source code

1. Copyright clauses
2. License keys
3. SPDX keys
4. Links to license texts

Analysis of comments in source code

- Copyright clauses
  - Needed for attribution requirements
  - Cut copyright information from license texts to increase the precision (some texts allows to achieve 100% match by only comparing hashes)
  - Experimented with techniques based on natural language processing algorithms
    - Poor precision because of the typical structure of clauses
  - Chosen to use the ScanCode algorithm mostly based on regular expressions
- License keys
  - Search for known keys/aliases/names in comments
  - Collect keys/aliases/names in our own database
- SPDX
  - Search and parse SPDX expressions
- URLs
  - Search for know URLs leading to license texts
  - Example: Licensed under https://www.apache.org/licenses/LICENSE-2.0.txt
  - Collect URLs in our own database

Conclusion

- The need of analysis for effective licenses remains high
- Our work showed that similarity analysis for the identification of licenses provides excellent results
- However, to know identified effective licenses does not finally answer the licensing question
- Further work is to be done to identify changes in original license texts

We invite you to participate in further work

Please find our sources at https://github.com/TrustSource

More information about Open Source Compliance can be found at https://support.trustsource.io/hc/en-us

Trust Source

TrustSource ist eine Marke der EACG
EACG GmbH – Enterprise Architecture Consulting Group
EACG Operations Services GmbH
Taunustor 1 (TaunusTurm)
60310 Frankfurt am Main

T: +49 69 153 22 77 50
F: +49 69 153 22 77 51
W: https://www.eacg.de