ScanCode
a plan for false positive license detection

Spring 2022

Philippe Ombredanne, nexB Inc.

Context

- ScanCode detects licenses using rules
  - Each rule is a license or notice text file and a YAML file
- Each rule is for a license expression and has tags such as: is_license_text, is_license_notice, is_license_intro, is_license_reference
- A rule is for a license expression
- The license expression of some rules "unknown" licenses

Problem

- There are false positive or inaccurate license detections.
  - partial or incorrect detections
  - license rules that are too broad
  - license matches that could be combined

More problems

- False detection of **very short** and weak license detection
- **Partial detection** of a license text or notice fragment which is too weak to represent a bona fide license detection alone.
- Detection of longer unknown license references such as **introduction** and references
- Lack of proper detection of a structured license tag found in a package manifest, returned as an unknown license
- Fragments of the same license are detected with only copyrights added in between
- Sequence of SPDX licenses id are found in license detection tools

nexB

Solution elements

- Fix bugs
- Add new way to combine matches
  - (say a license intro followed by a license)
- Integrate the existing ML-based ScanCode Analyzer as standard in SCTK
  - Can help spot weak licenses matches and suggest new license detection rules
- Add other specific filtering heuristics (using data provided)
  - Similar to the SPDX license id lists filter
- Report mere clues separately from actual license matches
  - Rename and reclassify "unknown"

Solution elements

- Make it easier to report, review and curate license detections
  - Web app to scan a text (or a whole package)
  - UI to easily report as correct or incorrect and select the text that should be detected and what license this should be
  - Also work towards shared, reusable scans and peer reviews
  - Pending project ideas for contributors

Solution elements

- Report the primary license
  - Already merged in the develop branch
  - Goal is not to hide secondary licenses, but rather to surface the primary license as found in key files and package manifests
  - This is not directly a false positive solution but it contributes

Solution elements

- New license clarity scoring
  - Already merged in the develop branch
  - Evolution of the clarity score we crafted for ClearlyDefined
  - Complete and improved re-design to better capture the notion of license clarity as a hint to whether further review is needed
  - This is not directly a false positive solution but it contributes to better understand if there could be false positive

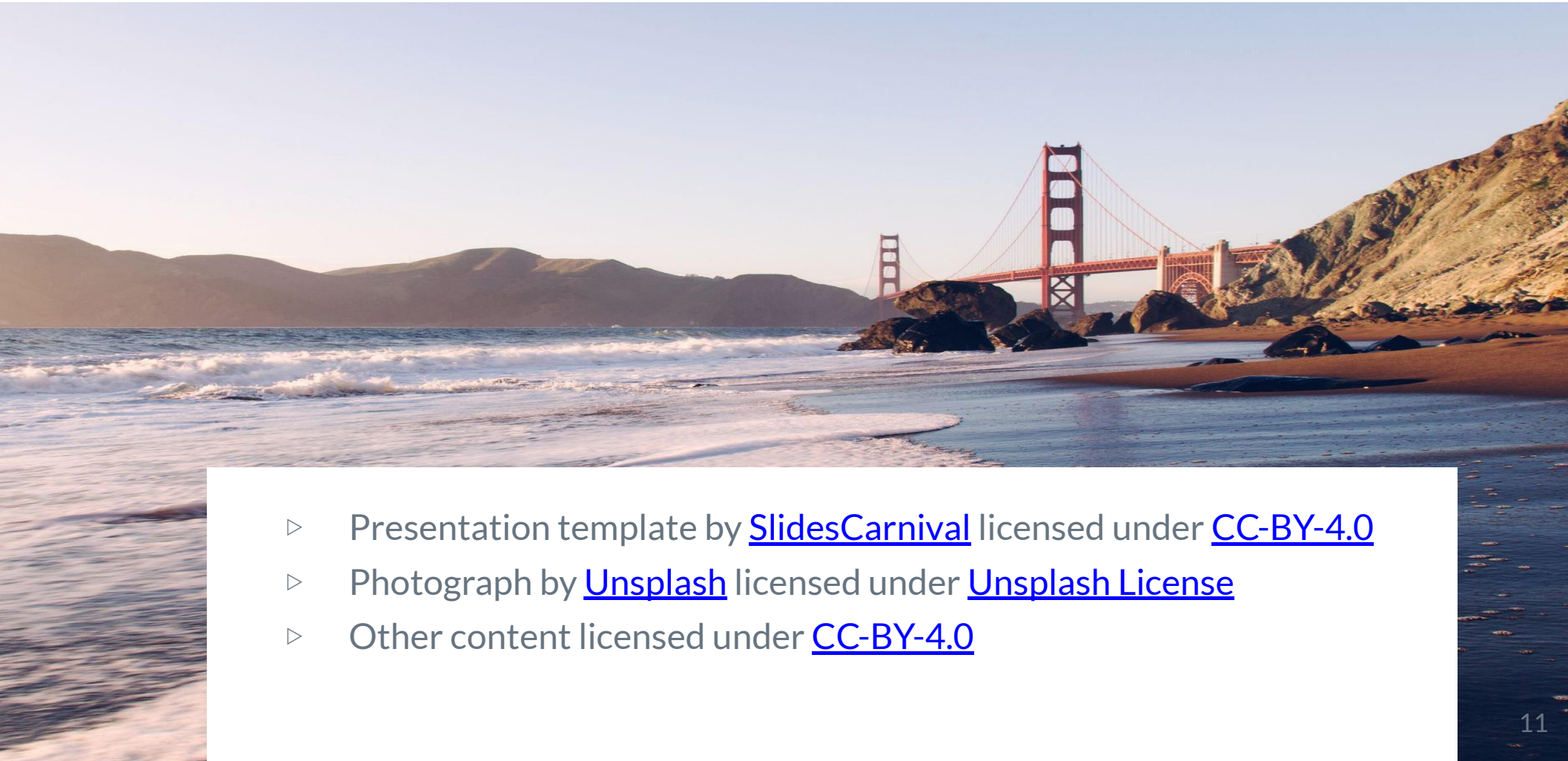
Solution elements

- Key {{phrases}} in license text rules
  - Already merged in the develop branch
  - Used to tag key phrases of a license rule that must be present for the rule to be considered as matched
  - Makes it easy to correct ambiguous license detections
  - ~ 500 rules already tagged
  - More to come

Logistics

- Can you help funding some work?
- If yes, what would be the best way for you?
  - Support contract?
  - Grants to open source project?

Credits



- Presentation template by SlidesCarnival licensed under CC-BY-4.0
- Photograph by Unsplash licensed under Unsplash License
- Other content licensed under CC-BY-4.0