Name
Email
Github
LinkedIn
College
Phone No
IRC Handle
Timezone
Location

SPDX License Detection Package

GSoC Proposal 2021 for Dart Organization

ABSTRACT

This project is related to the development of Dart Package for detection of SPDX license from the LICENSE file. It addresses the issue of identifying the licenses used in the project. This dart package will automate the process of reading the license file and then gives a valid SPDX identifier which could be used to properly display the license on pub.dev website.

All the supported licenses are listed here :

https://github.com/spdx/license-list-XML/tree/master/src

PROPOSAL

Problem :-

- Software licensing is very important for any open-source software. Sometimes we use licenses that we do not recognize either, thus we don't understand what we could do with the software and what we couldn't.
- This problem gets more complex when we have multiple licenses in the same software project.
- Recognizing a license and displaying its valid spdx identifier is crucial to large as well as small projects.

My Approach :-

As the main project idea suggests that we could use 2 approaches for identifying the license used in the project.

1. **Following the SPDX License List Matching Guidelines :-**

   The advantage of using this approach is that it is the standard way for getting the solution to this problem. However this approach would require a lot of time and resources. Another reason for not using this approach is that it is based on guidelines which could be either followed in the project or not.

2. **Applying text similarity like licensee with optimized approach :-**

   This is a very good approach if we want to cover a wide range of licenses and do not care about guidelines since this approach is based on test similarity.

   There are two main things in text similarity algorithm that I am going to use :

   a. After normalizing the text according to spdx recommendation, there is a possibility that we could get the exact copy of the licenses given here :

      https://github.com/spdx/license-list-XML/tree/master/src

   b. If we still don't get it then we will apply the **Sorensen-Dice** coefficient which is really good at calculating the similarities between two strings. Then I will optimize the solution by **hashing** the licenses and its text before applying the Sorensen-Dice coefficient algorithm. This algorithm will also give the possibility of being the correct license.

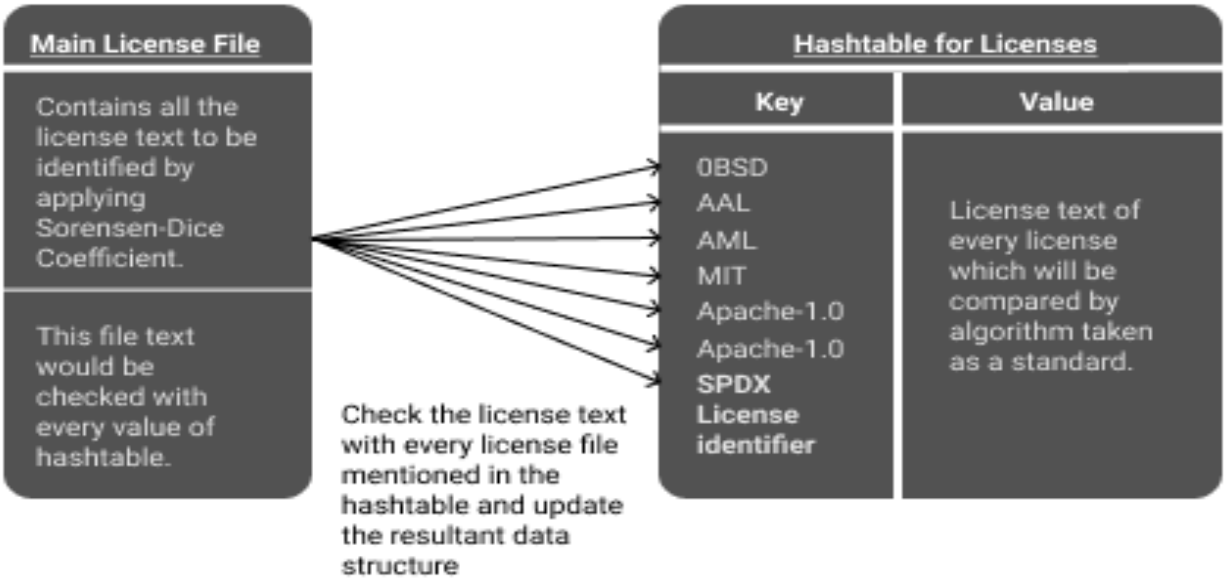
Features :-

- Native Dart package for identifying licenses with simple usage.

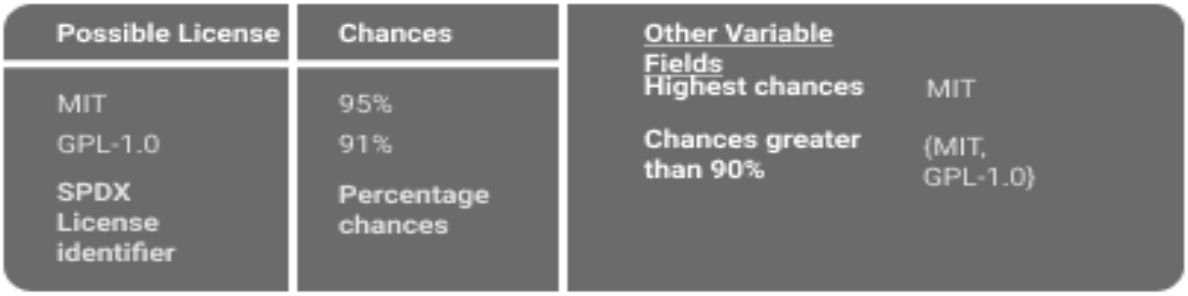
  import "package:Package-name/Package-name.dart";

- Supported command-line usage
  a) Package-name detect [PATH] # detect the license with the high chances
  b) Package-name help [PATH] # describe available commands
  c) Package-name version [PATH]# returns the license version
- Faster in execution and gives the best possibility of being the right license.

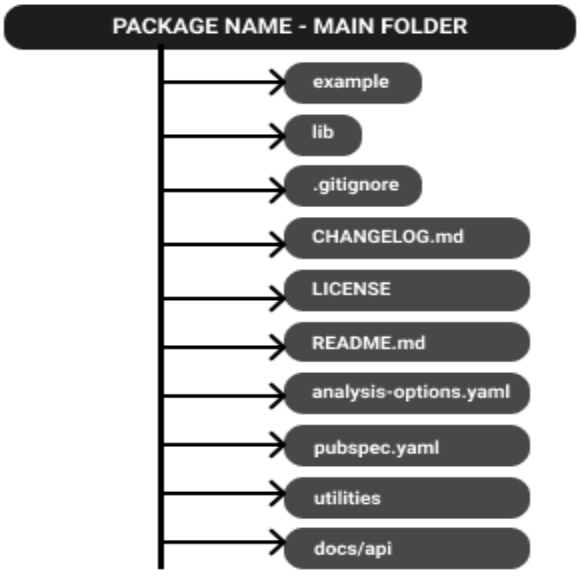
Basic Implementation Architecture of approach :-



Resultant Data Structure



Main Project Structure for developing the package :-



1. **example -** contains a set of examples on how the package could be used.
2. **lib -** main code used for license detection.
3. **.gitignore -** files to be ignored by the git
4. **CHANGELOG.md -** maintain the changes made over different versions of the package.
5. **LICENSE -** the license under which this package is published.
6. **README.md**
7. **analysis-options.yaml**
8. **pubspec.yaml**

9. **utilities -** contains setup.sh script and functionalities that are going to be added in the future.
10. **docs/api -** contains the documentation of the project.

DELIVERABLES

These are the things that I will deliver as an end-product before the completion of GSoC 2021.

1. **Dart Package** which would be published on pub.dev and it could also integrate with the existing package like pana. I could publish the package but due to time boundation I will integrate it with other packages later on.
2. **setup.sh** - A shell script for setting up command line usage of the package on Linux.
3. **Github Repository** for the development, issue-tracking and maintenance of the project. It is a must while developing an open-source software.
4. **Documentation** - Detailed documentation on how to use the package and describe the structure of the package so that others could also contribute to the project.
5. **Examples** - Sample examples showing how this package would be used.

WORK DONE TILL NOW / PREREQUISITES

I have successfully completed the prerequisites of the project.

- **Write a small sample program that can fetch one of the SPDX master files and print it to the terminal as markdown.**

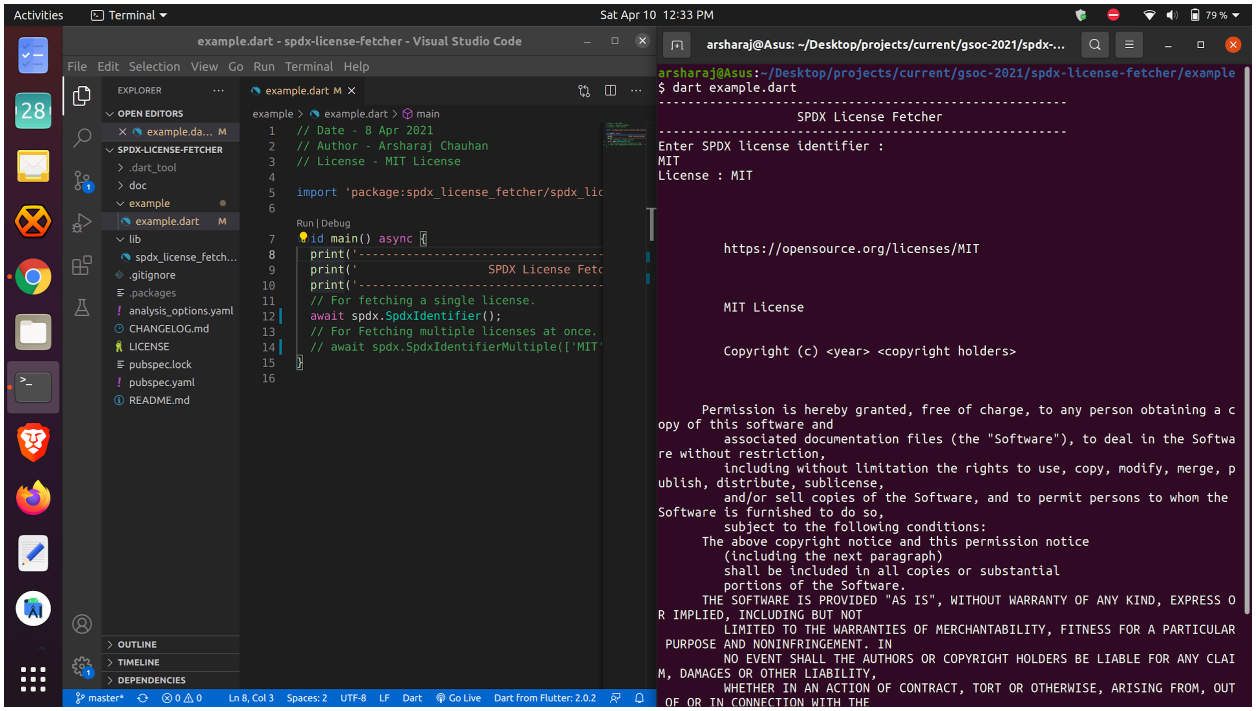
  Since I have already worked with packages, I made a package for doing this task and utilized it to print the content to the terminal after parsing the XML content.

  Here are the links to the project.

  Github : https://github.com/arsharaj/spdx-license-fetcher

  Pub : https://pub.dev/packages/spdx_license_fetcher

  Here is the screenshot of the output after using the package :



- **References to relevant documentation**
    a) https://github.com/licensee/licensee
    b) https://github.com/dart-lang/pana
    c) https://github.com/google/licenseclassifier
    d) https://spdx.dev/license-list/matching-guidelines/
    e) Sorensen-Dice Coefficient for text similarity
    f) Everything related to Dart Packages
    g) XML Parsing in Dart
    h) Algorithm implementation of Sorensen Dice Coefficient

PROJECT TIMELINE

1. April 13, 2021 - May 16, 2021 (Proposal Review Period)

- During my proposal review period, I am going to enhance my previous knowledge.
- Design and improve my package outline and design to every little bit of detail.

2. May 17, 2021 - June 7, 2021 (Community Bonding)

- Gain exposure to open-source software development by interacting with community members and introducing myself as an aspiring Flutter Developer.
- Learning and developing new things as well as helping out community members.

3. June 7, 2021 - July 16, 2021 (Working Phase 1)

- Setup project over Github and start working on it.
- Send the initial development release of the package after discussing my approach with my mentor.
- Initial testing of the package with all the possible test cases according to package outline.
- Keep on updating the package by minor bug fixes and document them in the changelog.
- Send my code for first evaluation.

4. July 16, 2021 - August 16, 2021 (Working Phase 2)

- Keep on fixing the bugs and errors for a more robust package.
- Make my codebase into standard format.
- Write the set of examples on how to use the package in the examples folder.
- Write the setup.sh file in the utilities folder.
- Work upon new features and issues suggested by the community members.
- Help out other community members during all of this time.
- Send my code for second evaluation.

5. August 16, 2021 - August 23, 2021 (Final Working Phase)

- Write appropriate documentation for use of the package and contributing to its development.
- Publish the documentation over Github pages.
- Publish the package over pub.dev.
- If possible, I would also try to integrate this package with packages like pana.
- Submit my final code and evaluations of the package.

6. August 23, 2021 - Onwards

- Keep on updating the package with new features and fixes.
- Contribute to open-source softwares and projects in some form or the other.

Other Commitments :-

Since I am in the third year of my B.Tech so it may be possible that I would not be able to give the required time during my end-semester examinations. But, I assure you that I will cover up my time before/after my semester exams. I could not give the exact timing of the exams due to the Covid disturbance. I will be in constant touch with my mentor organization.

RELATED WORKS

Spdx License Fetcher Package

- Language used - Dart
- Development of package from the scratch
- Console based Dart package that fetches license file by its SPDX identifier and then prints it to the terminal as markdown.
- Support multiple license fetching.
- Project idea : SPDX License detection Package
- Project link : https://github.com/arsharaj/spdx-license-fetcher
- Package link : https://pub.dev/packages/spdx_license_fetcher

- There were numerous projects and applications that I developed using Flutter but they were not open-source since I developed those applications in different hackathons at different colleges.

- Some of those applications are not even relevant to this project.

WHY DART ?

I really want to work with the Dart organization because your organization would give me the boost and exposure that I am craving for. I have been testing and trying out Flutter since its hotfix release. Flutter, Dart and this overall open-source community gave me a lot of things. Now it's my time to give back to the community.

I intend to work with the Dart organization for a long time and through GSoC I want to grab this opportunity to work with you guys and gain real world exposure to OSS development. As per my interests, intentions and skills, I have chosen the Dart Organization.

WHY ME ?

As for me I am a Computer Science student currently in third year of my B.Tech. I have good experience in the field of software and application development. I had already published many packages in different languages like Dart (obviously), Python. I have the knowledge and prerequisites required for this project idea. I intend to work with the Dart community for long time.

I will be waiting for your positive feedback.

Thanks and Regards,

Arsharaj Chauhan