

ライセンスルール自動生成を目的とした ライセンス記述パターン抽出手法

オープンソースソフトウェア工学研究室

60160087 東 裕之輔

OSSの再利用

- 近年，開発コスト削減のため，**オープンソースソフトウェア(OSS)**を再利用した開発が行なわれている



- ただし，OSSの再利用には，**OSSライセンスを遵守することが義務付けられている**

OSSライセンス

- OSSライセンスとは
 - OSSを利用する上で、著作者から許可されていること、義務付けられていることをまとめたもの
- OSSライセンスには、多くの種類が存在する
 - OSSライセンスは、2400種類以上存在する (BlackDuck社による調査)
 - 幅広く利用されているOSSライセンス



GPLv3



LGPLv3



MPLv2.0



MIT License
by ©ExcaliburZero

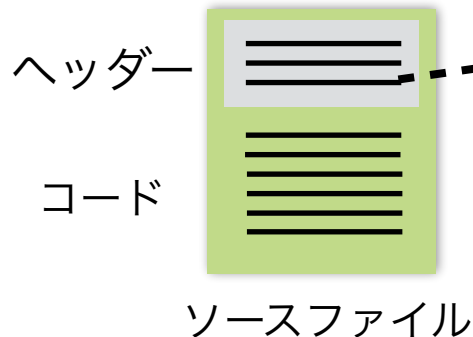


BSD License

ソースファイルのライセンス記述

- OSSライセンスは、ソースファイルのヘッダー部分に記述されている

BSD2のライセンス記述



```
#####  
# Redistribution and use in source and binary forms, with or without  
# modification, are permitted provided that the following conditions are met:  
#  
# 1. Redistributions of source code must retain the above copyright notice,  
#   this list of conditions and the following disclaimer.  
# 2. Redistributions in binary form must reproduce the above copyright  
#   notice, this list of conditions and the following disclaimer  
#   in the documentation and/or other materials provided  
#   with the distribution.  
#  
#####
```

- OSSライセンスを遵守するには、全てのソースファイルのライセンス記述を確認しなければならない

OSSに含まれるソースファイル数

- OSSには、非常に多くのソースファイルが含まれている



Linux

ドメイン : OS

ソースファイル数 : 22,310



ドメイン : DB

ソースファイル数 : 1,452



ドメイン : Server

ソースファイル数 : 291

- 全てのソースファイルのライセンスを手作業で確認するのは、労力がかかる

ライセンス特定ツール[1][2]

- ライセンスルール（正規表現）を用いて，自動的にソースファイルのライセンスを特定する



[1] Ninka : Daniel M German, Yuki Manabe, and Katsuro Inoue, "A Sentence-Matching Method for Automatic License Identification of Source Code Files" In Proc. of ASE'10, pp. 437–446

[2] Fossology : Robert Gobeille, "The fossology project," In Proc. of MSR'08, 2008, pp. 47–50.

ライセンスルール

- ライセンス記述を構成するライセンス文を特定するための正規表現

一般的なBSD2のライセンス記述

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

BSD2のライセンスルール

文名 : BSDpre

正規表現 : Redistribution(s)?
and use in source and binary[...]

文名 : BSD source cond.

正規表現 : Redistribution(s)? of
source code[...]

文名 : BSD binary cond.

正規表現 : Redistribution(s)? in
binary form[...]

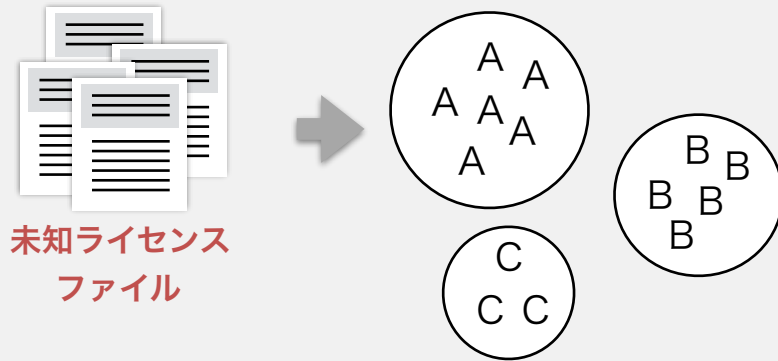
ライセンス特定ツールの課題

- 新しいライセンスや、未知の表記ゆれは、年々増加する
- ライセンス特定ツールは、**コングスタント**に新しいルールを追加し続けていくことが重要



手作業でのライセンスルール作成手順

STEP1. 未知ライセンスファイル集合を ライセンス毎に分類



STEP2. 各クラスターの頻出記述と 表記ゆれを調査する

頻出記述

This[...] is free software.
it under the term of the A license

[...]部の表記ゆれ

1	program
2	software
3	library

STEP3. ライセンスルール作成

1	This program is free ~
2	This software is free~
3	This library is free~

新ライセンスルール

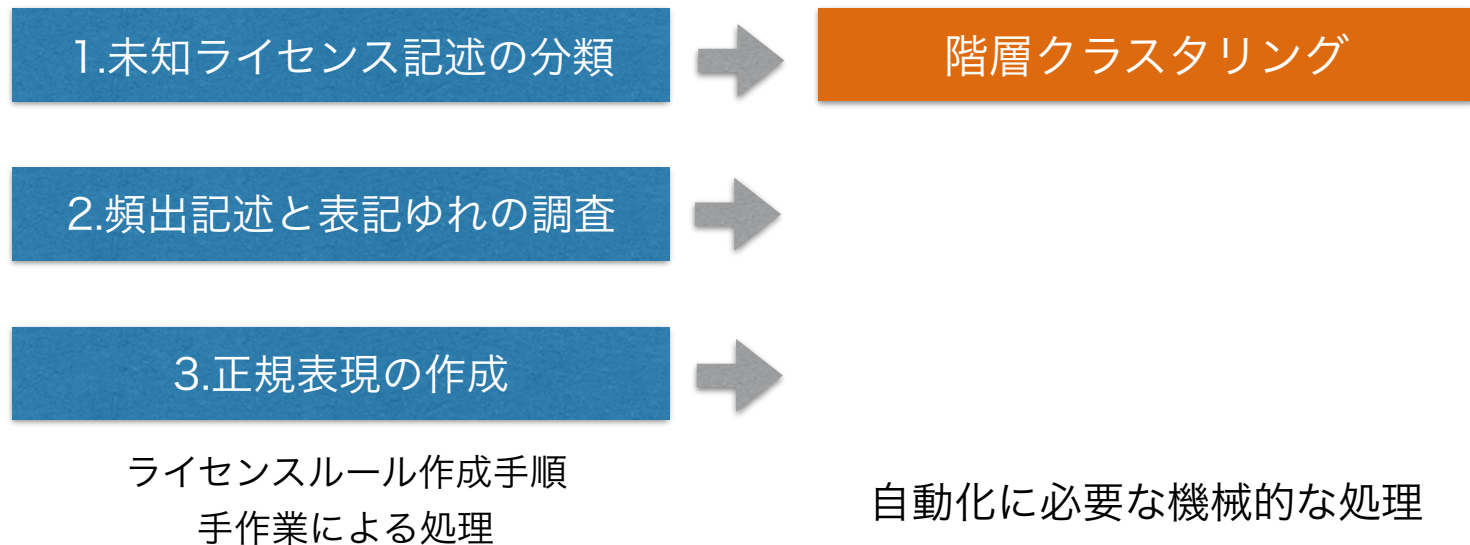
This (program|software|library) is ~

問題点

大量の未知ライセンスファイルが検出
された場合、大変な労力がかかる

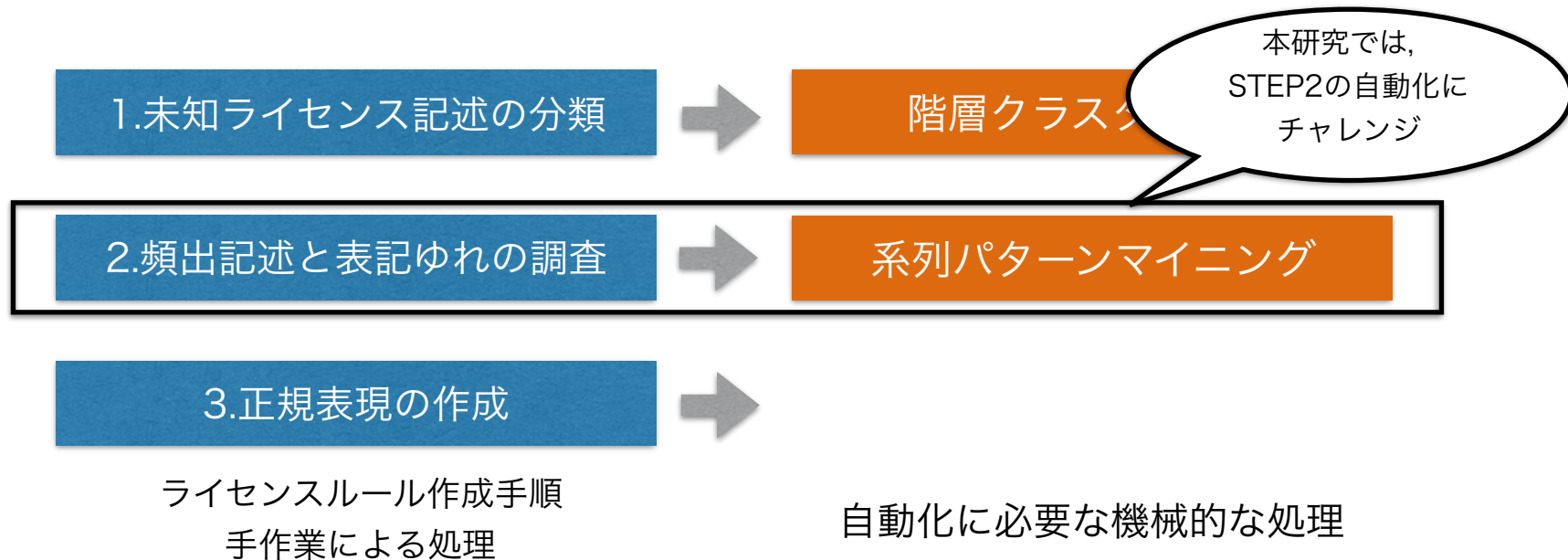
最終的なゴール

- ✧ ゴール：ライセンスルール自動生成手法の構築
- ✧ アプローチ
 - ライセンスルール作成手順における各ステップを、機械的処理に置き換える



最終的なゴール

- ✧ ゴール：ライセンスルール自動生成手法の構築
- ✧ アプローチ
 - ライセンスルール作成手順における各ステップを、機械的処理に置き換える



本研究の目的

- **本研究の目的** : 既存手法によって作成されたクラスタから**ライセンス記述パターン**を抽出する
- **ライセンス記述パターン** : 特定のライセンス記述に頻出する語句の出現順序

クラスタAの
未知ライセンス記述

This **program**
is free
software. It
under the term
of the A license

This **software**
is free
software. It
under the term
of the A license



系列パターンマイニング
(Step.2の自動化)

クラスタAから抽出された
ライセンス記述パターン

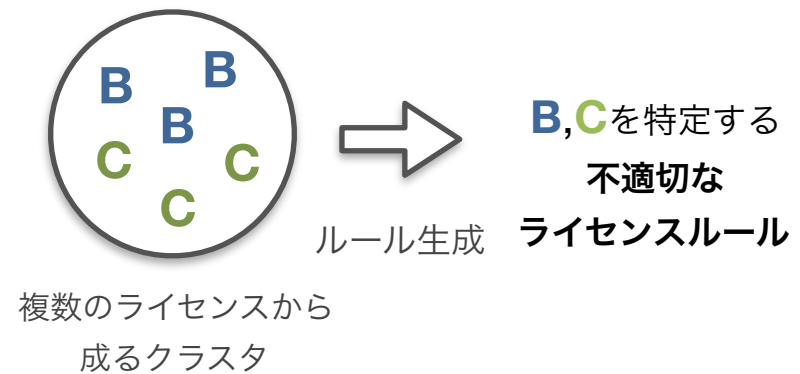
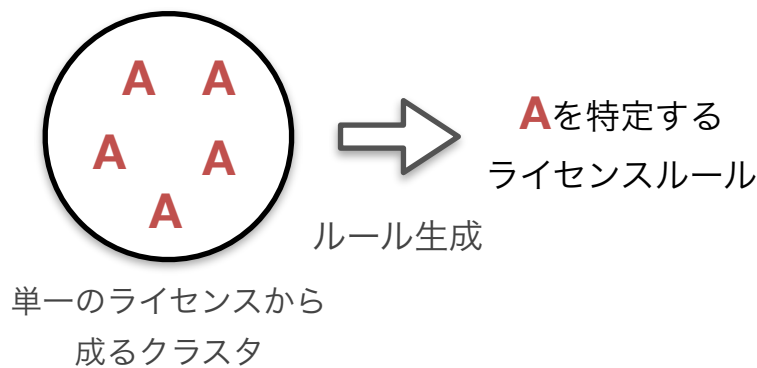
(This),(is),(free),(software.),(is),(free)
(It),(under),(the),(term),(of),(the),(A),
(license)

系列パターンマイニング

- 系列とは
 - 時系列データ
 - 系列の例：(コーヒー, パン), (コーヒー, 牛乳, パン)
- 系列パターンマイニングとは
 - 複数の系列に頻出する系列（系列パターン）を全て列挙する技術
 - 入力：(コーヒー, パン), (コーヒー, 牛乳, パン)
 - 出力：(コーヒー), (パン), (コーヒー, パン)
 - 本研究では, 系列を文章に出現する単語として捉え, ライセンス記述パターンの抽出に応用

Step2自動化における技術的課題1

- 1.単一のライセンスからなるクラスタから記述パターンを抽出したい
 - 複数のライセンスから成り立つクラスタからは、複数のライセンスを特定してしまう**不適切なライセンスルール**を生成する
 - 既存手法は、どうしても複数のライセンスから成るクラスタを作成してしまう



Step2自動化における技術的課題2

- 2. ライセンスルール作成に有用なライセンス記述パターンのみ抽出したい
 - 系列パターンマイニングの欠点：ルール作成に不要な大量の系列パターンを出力してしまう

例：LGPLのクラスタに頻出する系列パターン（一部）

```
-(This),( program),( is),( distributed),( in),( the),( hope),( that),( it),( will),  
( be),( useful),( but),( WITHOUT),( ANY),( WARRANTY),( without),( even),  
( the),( implied),( warranty),( of),( MERCHANTABILITY),( or),( FITNESS),  
( FOR),( A),( PARTICULAR),( PURPOSE).  
-(This),( is),( or)  
-(This),( is),( the),( the)  
-(This),( is),( it),( the),( of)  
-( Free),( Software)  
-(See),( the),( GNU),( Lesser),( General),( Public),( License),( for),( more ),  
( details.)  
-( by)  
-( of),( the),( GNU),( Lesser),( General),( Public)
```

- “This is the the”, “by”のような系列パターンは、どのライセンス記述でも成り立ってしまうため、ルール作成には**不適切**である

提案手法の概要

編集距離を用いたフィルタリング

- クラスタの中心との類似度が94%以上の記述以外をフィルタリング
 - クラスタの中心となる記述 : 編集距離による類似度が**94%**以上ある記述を最も多く持つ記述

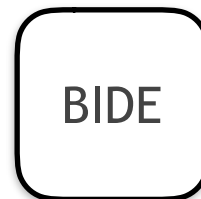
BIDEによる系列パターンマイニング

- BIDEアルゴリズム

- 系列パターンマイニングアルゴリズムの1つ
- 出現記述数が同じで、かつ、他の系列パターンの部分集合となる系列パターンを抽出しない

アイテム列集合

ABCD
ABC
BCDEFG
BCDEXG



出現記述数
2以上



BIDEによって抽出される系列パターン

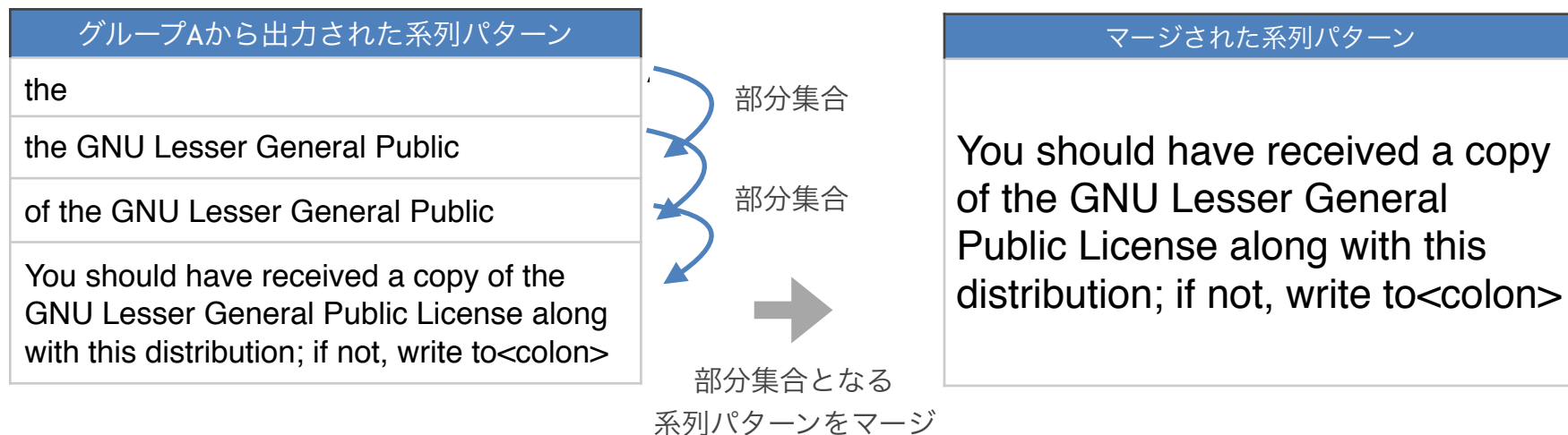
系列	出現記述数
BC	4
BCD	3
ABC	2
BCDEG	2

~~AB~~	~~2~~
~~BCDE~~	~~2~~

抽出しない

系列パターンのマージ

- 他の系列の部分集合になるような系列パターンをマージする
 - BIDEは、出現記述数が異なる系列間では、他の部分集合となる系列パターンを出力してしまう



- 不要なパターンをマージしたものをライセンス記述パターンとする

ケーススタディ

- 実験の目的：提案手法は、2つの技術的課題をどの程度解決できるか2種類の実験を行い検証する
 - 評価1. データセットとなるクラスタのうち、単一のライセンスから成るクラスタの割合
 - 評価2. 抽出したライセンス記述パターンのうち、ライセンスルール作成に有用なライセンス記述パターンの割合
- 既存手法によるクラスタリング結果
 - Debian v7.8.0で検出された2,446の未知ライセンスファイルを既存手法によりクラスタリングした結果を用いる

Debian-7.8.0	クラスタ数	単一のライセンスから成るクラスタ	複数のライセンスから成るクラスタ
2,446ファイル	246	170(69.1%)	76(30.9%)

評価1の実験結果

- 実験方法

- 編集距離を用いたフィルタリングを行い、単一のライセンスから成るクラスタの割合がどれくらい向上したかを確認する

- 実験結果

- 全体の95%のクラスタを単一のライセンスから成るクラスタにできた

Debian-7.8.0	外れ値除去	クラスタ数	単一のライセンスから成るクラスタ	複数のライセンスから成るクラスタ	フィルタリングされた記述
2,446ファイル	外れ値除去前	246	170(69.1%)	76(30.9%)	-
	外れ値除去後	126	120(95%)	6(5%)	1960

評価2の実験方法

1. 抽出したライセンス記述パターンを正規表現化し，データセットの未知ライセンス記述と，grepによるマッチングを行う
2. 各ライセンス記述パターンが**マッチした未知ライセンス記述のライセンスの種類が1種類であったかどうか**を確認する

ライセンス記述パターン

(This),(is),(free),(software.),(is),(free)
(It),(under),(the),(term),(of),(the),(A),
(license)

正規表現化



正規表現化されたライセンス記述パターン

This.* is.* free.* software.* is.* free.*
It.* under.* the.* term.* of.* the.* A.*
license

.は任意の1文字,

*は0回以上の繰り返しを表す

評価2の実験結果

	生成されたパターン数	1種類のライセンスに マッチしたパターン数	割合
ライセンス記述パターン	371	240	64.7%

ライセンス記述パターン	ライセンス名	#Lic
See the License for the specific language governing rights and limitat[...] If you do not delete the provisions above, a recipient may use your ve[...] If you wish to allow use of your version of this file only under the t[...] The Original Code is AOLserver Code and related documentation distribu[...] The Initial Developer of the Original Code is America Online, Inc. Por[...] The contents of this file are subject to the AOLserver Public License [...] Alternatively, the contents of this file may be used under the terms o[...] Software distributed under the License is distributed on an "AS IS" ba[...] You may obtain a copy of the License at http://aolserver.com/ .	AOLserver AOLserver AOLserver AOLserver AOLserver AOLserver AOLserver AOLserver AOLserver	6 1 1 1 1 1 1 5 1
This software/database is a "United States Government Work" under the [...] PUBLIC DOMAIN NOTICE National Center for Biotechnology Information The National Library of Medicine and the U.S. Government have not plac[...] The NLM and the U.S. Government disclaim all warranties, express or im[...]	PublicDomain PublicDomain PublicDomain PublicDomain	1 1 1 1
2 - This Source Code Form is subject to the terms of the Mozilla Publi[...] If a copy of the MPL was not distributed with this file, You can obtai[...]	MPLv2.0 MPLv2.0	1 1
This program is distributed in the hope that it will be useful, but WITH[...] You should have received a copy of the GNU Lesser General Public Lic[...]	LGPL LGPL	32 7

系列を示す括弧を外して表示。 #Lic はライセンス記述パターンが対応するライセンス種類数を示す。

まとめと今後の課題

- まとめ

- 未知ライセンスのクラスタから、ライセンス記述パターンを抽出する手法を提案した
- Debian v7.8.0を対象としたケーススタディで分かったこと
 - 編集距離を用いてフィルタリングすることにより、クラスタリング結果から単一のライセンスから成るクラスタの割合を増加させることができた
 - 系列パターンマイニングを用いて、ライセンスルール作成に有用なライセンス記述パターンを抽出できた

- 今後の課題

- フィルタリングされたファイルが多い
 - 課題1において、1つのクラスタから複数のライセンス記述グループを取り出せるようにする
- 複数のライセンスにマッチしたライセンス記述パターンの活用
 - 同じクラスタから抽出されたその他のライセンス記述パターンと同時に用いることで、活用できないか検証する

今後の予定

- STEP3の自動化を実装し，提案手法の有用性を評価する
- 提案手法の有用性：何ファイルの未知ライセンスファイルのライセンスを特定できるようになったのか

		10月	11月			12月			1月			2月		
		下	上	中	下	上	中	下	上	中	下	上	中	下
STEP3の自動化	実装													
	実験													
	評価													
修論執筆	目次作成													
	背景													
	問題点													
	目的とアプローチ													
	ケーススタディ													
	考察													
	結論													
	スライド作成													
修論発表	発表練習													

STEP3自動化に必要な処理（実装中）

- バリエーションチェック行列の作成し，正規表現を生成する

ライセンス記述 パターン	This	.*	is	.*	free	.*	software
ライセンス文1	This	software	is		free		software
ライセンス文2	This	library	is		free		software
ライセンス文3	This	program	is		free		software
ライセンス文4	This		is		free		software

生成されるべきライセンスルール

This (?: (?: software| library)| program)? is free software

STEP3自動化実装の進捗

- 前スライドの例を実現するところまでは完了
 - あと数日で実装完了できそう

```
> m<-get_variants_matrix("This.* is.* free.* software",c("This softw  
are is free software","This library is free software","This program  
is free software","This is free software"))
```

	[,1]	[,2]	[,3]
[1,]	"software"	""	""
[2,]	"library"	""	""
[3,]	"program"	""	""
[4,]	""	""	""

```
> make_lic_rules("This.* is.* free.* software",m)
```

```
[1] "This(?:(?: software| library)| program)? is free software"
```