

11/25眞鍋先生MTG

東 裕之輔

Done

- 研究ストーリースライドの修正
- ライセンスの特定ができる分類器を作成する
- テスト
 - OpenOffice v330のUnknownファイルのライセンスを特定

分類器の作成

- 対応ライセンス19種類

ライセンス			
AGPL v3.0+	BSD3clause	GPLv2+	MIT
APL v1.0	BSD4clause	GPLv3+	MPLv1.1
APL v1.1	CPLv1.1	LesserGPL 2.1+	MPLv2.0
APL v2.0	EPLv1.1	LesserGPL 3.0+	NPLv1.1
BSD2clause	GPLv1+	LibraryGPL 2.0+	

- 各ライセンス100記述
 - 計1900ファイル

学習データ収集先

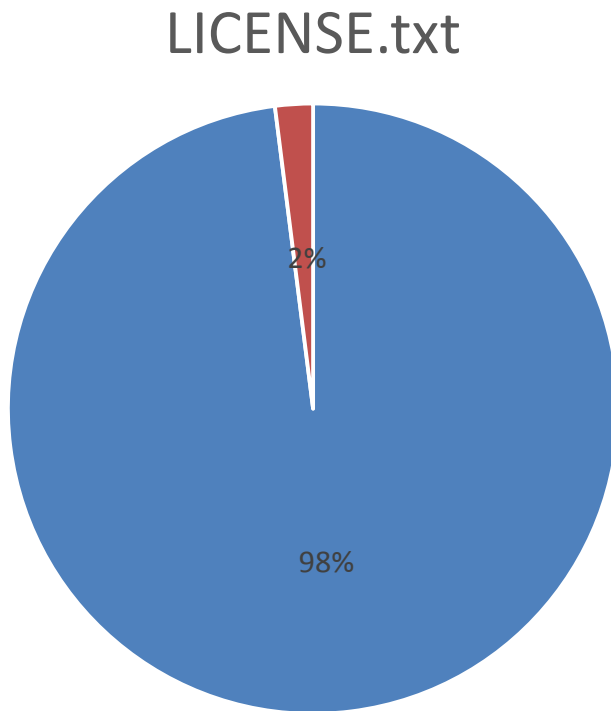
ライセンス	ossプロジェクト
GPLv1+	Gcc v2.9.0
GPLv2+	Debian(Ninka評価データ)
GPLv3+	Gcc4.9.2
LibraryGPL v2.0+	Gcc+Glibc+firefoxv9
LesserGPL v2.1+	Glibc v2.20
LesserGPL v3.0+	OracleOpenOffice v3.3.0
AfferoGPL v3.0+	scambio
CPLv1.1	Eclipse r14641
EPLv1.1	Eclipse MARS
MIT	Firefox v33

ライセンス	ossプロジェクト
BSD4clause	OpenBSD v5.3
BSD3clause	j-git v3.5.1 (201410131835-r)
BSD2clause	OpenSSH v6.7p1
NPLv1.1	Firefox v1.0
MPLv1.1	Firefox v9.0
MPLv2.0	Firefox v33
APL v1.0	Httpd 1.3.0
APL v1.1	Httpd 2.0.35
APL v2.0	OpenOffice v3.4.0

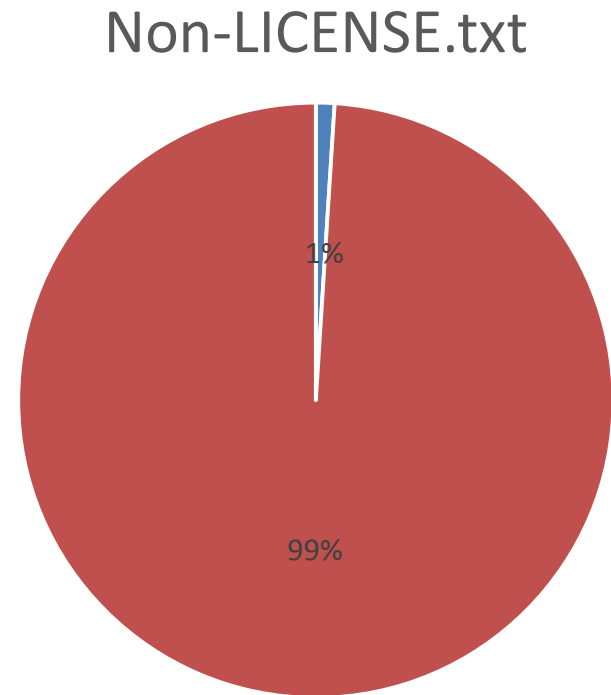
分類テスト結果

- ライセンスとノンライセンス,100ファイルを分類

TRUE:FALSE=98:2



TRUE:FALSE=1:99



2. ライセンスの分類

- テストデータ

- 19のライセンスセグメンテーションのファイル

Length	Class	Mode		
19	character	character		
AGPLv3.txt		APLv1.0.txt	APLv1.1.txt	APLv2.0.txt
"AGPLv3"		"APLv1.0"	"APLv1.1"	"APLv2.0"
BSD2clause.txt		BSD3clause.txt	BSD4clause.txt	CPLv1.0.txt
"BSD3clause"		"BSD3clause"	"APLv1.0"	"CPLv1.0"
EPLv1.0.txt		GPLv1.txt	GPLv2.txt	GPLv3.txt
"EPLv1.0"		"GPLv1"	"GPLv1"	"GPLv1"
LGPLv2.1.txt		LGPLv3.0.txt	LibraryGPLv2.0.txt	MIT-1.txt
"GPLv1"		"LGPLv3.0"	"GPLv1"	"APLv1.1"
MPLv1.1.txt		MPLv2.0.txt	NPLv1.1.txt	
"MPLv1.1"		"MPLv2.0"	"MPLv1.1"	

- 正解率: $11/19=57\%$

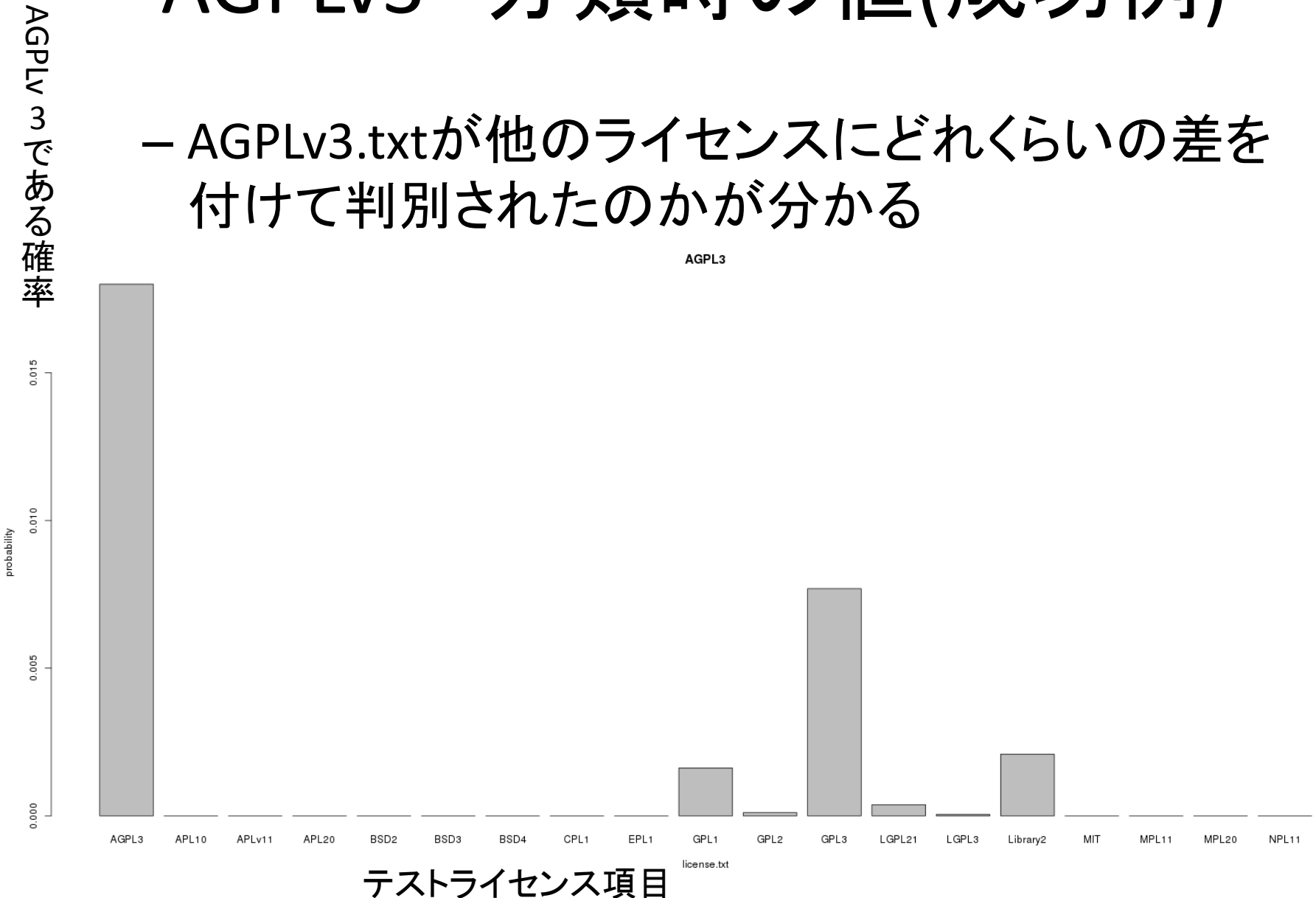
あまりよくない

ライセンス分類の課題

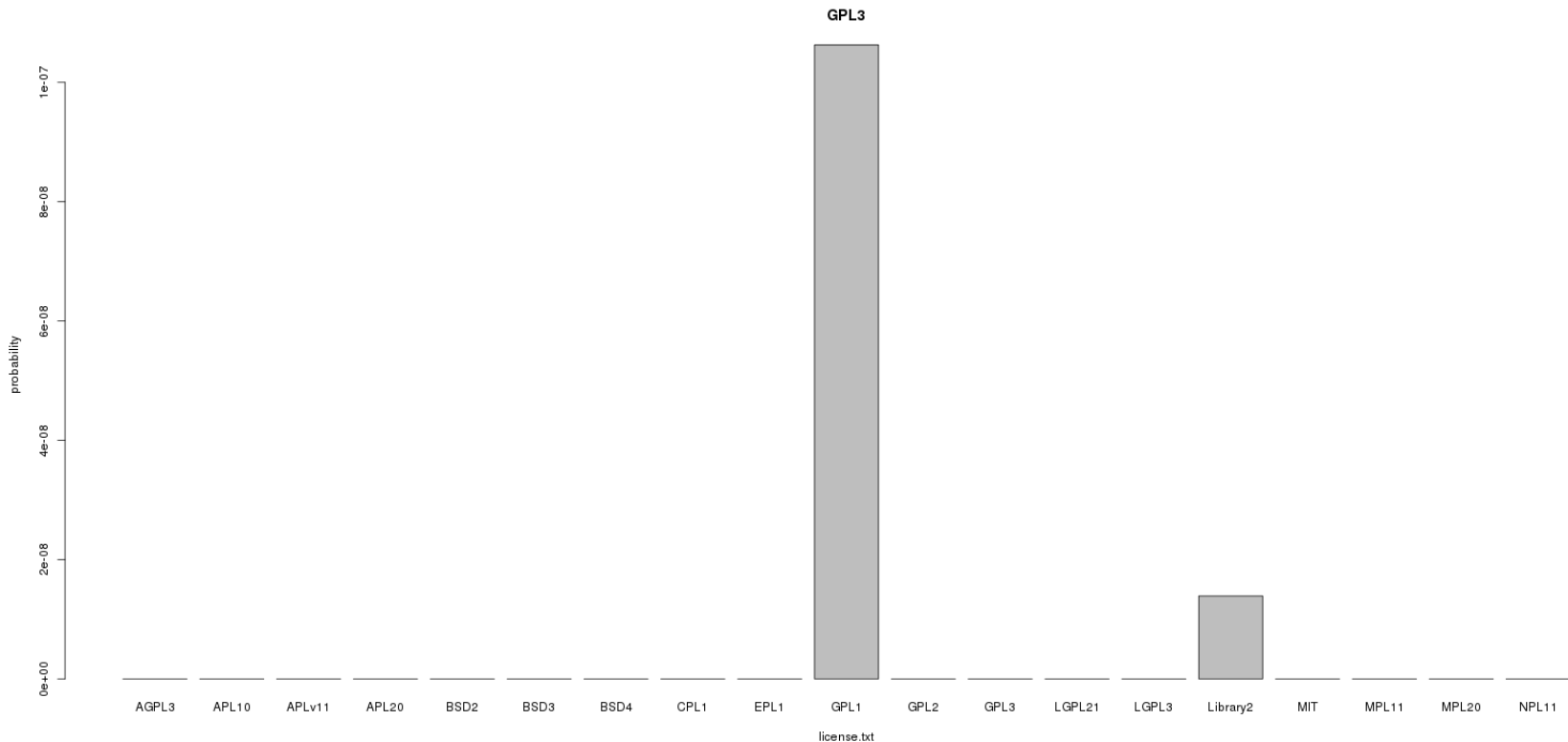
- 包括関係にあるライセンス文の判別が難しい
 - 違うライセンスに同じ記述が含まれている
 - GPL系・・・GPL,LGPL,AGPL
 - BSD系・・・BSD,APL,MIT
 - バージョンの違い
- 学習データの取り方を考える必要がある
 - 個人名,団体など消去していない
 - 色々なプロジェクトから取得する
 - セグメンテーションで学習しない

AGPLv3 分類時の値(成功例)

— AGPLv3.txtが他のライセンスにどれくらいの差を付けて判別されたのかが分かる



GPLv3 分類時の値(失敗例)



分類失敗の原因

- なぜかGPLv3やv2の値が低い
- LGPLもあまり確率が高くない

※他のライセンスプロット画像はplot_valuesに

3.Unknownファイルの分類(参考)

- OpenOfficev330のUnknownファイルを分類
 - 対象は50ファイル
- Unknownファイル群にはSeeFileがたくさん存在する(dmake)
 - 現段階ではSeeFileの学習データを集められていない
 - 追加でOpenOffice v3.3.0のUnknown群から即席のSeeFile学習データを作成

dmakeのライセンス記述(SeeFile)

```
/* RCS $Id: putenv.c,v 1.1.1.1 2000-09-22 15:33:34 hr Exp $
```

```
-- SYNOPSIS
```

```
--   My own putenv for BSD like systems.
```

```
-- DESCRIPTION
```

```
--       This originally came from MKS, but I rewrote it to fix a bug with  
--       replacing existing strings, probably never happened but the code  
--       was wrong nonetheless.
```

```
-- AUTHOR
```

```
--   Dennis Vadura, dvadura@dmake.wticorp.com
```

```
-- WWW
```

```
--   http://dmake.wticorp.com/
```

```
-- COPYRIGHT
```

```
--   Copyright (c) 1996,1997 by WTI Corp. All rights reserved.
```

```
--   This program is NOT free software; you can redistribute it and/or  
--   modify it under the terms of the Software License Agreement Provided  
--   in the file <distribution-root>/readme/license.txt.
```

```
-- LOG
```

```
--   Use cvs log to obtain detailed change logs.*/*
```

一般的なSeeFileとは少し違う

分類結果

- License_bayes_test_000330.xlsx参照
- 50ファイル中
 - ライセンス文を抽出できなかった:5ファイル
 - ライセンス文を抽出できた:45ファイル
 - 成功率=90% よく取得できている
 - ライセンス文の抽出かつ特定に成功:31ファイル
 - 精度=62%
 - ※ただしSeeFileは学習データとテストデータが同一

考察

- ライセンス文が包括関係にあるライセンスの特定は難しい
 - バージョン,GPL系,BSD系
- ノイズに弱い
 - 同じようなライセンス文でも少し違うだけで別のライセンスとされてしまう(dmake,aggなど)
 - セグメンテーションで抽出している

議論

- ベイズフィルター分類器の改善
 - 単語の形態素を統一して学習
 - 学習データクレンジング
 - 1ライセンス複数プロジェクトで学習データを取得する
- クラスタリングを行うかどうか
 - ソフトクラスタリング
 - NMF(Non-Negative Matrix Factorization)