

11/10

真鍋先生MTG

進捗報告

東 裕之輔

Done

- 研究ストーリーの作成
 - 別スライド
- Ninka評価データ(Debian5.0.2)から学習データを作成し分類テスト

ライセンス学習データの作成

- ライセンス記述の一部を指定してgrepを実行
 - “Copyright” “merchantability”など即席で作成
 - 100ファイル作成 9割がGPL系
 - 他のライセンス BSD2,BSD4,MIT,Apache

	term	frequency	NA.	density	occurrence
1088	software	28	FALSE	0.0022479127	0.97
48	andor	56	FALSE	0.0044958253	0.96
228	copyright	6	FALSE	0.0004816956	0.93
724	merchantability	54	FALSE	0.0043352601	0.91
948	purpose	54	FALSE	0.0043352601	0.91
372	fitness	53	FALSE	0.0042549775	0.90

ライセンス文学習データで発生確率が高い単語上位5つ

Non-ライセンス学習データの作成

- ライセンスに頻出する単語を指定し、該当しなかったセグメンテーションを抽出する
 - "software" "merchantability" "license"など
- 100ファイル作成

	term	frequency	NA.	density	occurrence
238	file	3	FALSE	0.001993355	0.11
108	code	2	FALSE	0.001328904	0.08
410	make	9	FALSE	0.005980066	0.07
730	use	9	FALSE	0.005980066	0.07
731	used	9	FALSE	0.005980066	0.07
755	will	10	FALSE	0.006644518	0.07

nonライセンス文学習データで発生確率が高い単語上位5つ

テストデータの作成

- OpenOffice v3.3.0からNonライセンスセグメンテーションを100ファイル抽出
- これにフィルターをかけてどれくらいFalseが出力されるかを調べる
 - 機械学習の参考書に書いてあったのでこの方法を取りました

実行結果

- False:True = 93:7

分別は結構できている

- 条件

- ライセンス文事前確率=20%

- 未知の単語の出現率30%

に設定した場合

7ファイルの原因

- 学習データが少ない
- 恐らく両方の辞書に載っていない単語が多い
- 学習データをさらに増やす必要がある

テストデータが全てライセンス文の場合

- fossologyのspdxのライセンスファイルを使用
 - <http://www.fossology.org/projects/fossology/wiki/MatchSPDXLicenceIDs>
- 全210ファイルのライセンス文がどれくらいTRUEになるか
- 実行結果
 - False:True=195:15
 - ライセンスは精度がでない
 - 学習データのライセンスの偏りが原因

議論

- 今後精度を上げるにはどうすればよいか
 - 少なくともライセンス文の判定は100%である必要がある
- 学習データを増やす
- 未知の単語の確率を変更する
 - ゼロ頻度問題
- Tf-idf
 - どの単語と一緒に使われているか