TopSE SW開発実践演習（2018/11/1向け）

課題提出日　2018/11/1
学籍番号　30014
氏名　大内一哲

Logistic Regression（ロジスティック回帰）を調査することで原理を学ぶ

■Data Science Glossary on Kaggle
（https://www.kaggle.com/shivamb/data-science-glossary-on-kaggle-updated）

1. Regression Algorithms
   1.1 Linear Regression
   1.2 Logistic Regression ← 報告対象

Logistic Regression

	Kernel	Author	Language	Views	Comments	Votes
1	Logistic regression with words and char n-grams	Bojan Tunguz	Python	32822	98	369
2	Logistic Regression and ROC Curve Primer	Troy Walters	R	11714	18	83
3	Example: Attacking logistic regression	Allunia	Python	8991	5	82
4	Logistic regression with words and char n-grams	thousandvoices	Python	6251	14	80
5	Bayesian Logistic Regression with rstanarm	Aki Vehtari	R	20479	15	57
6	Logistic Regression Implementation	DATAI	Python	388	0	55
7	Simple logistic model - PORTO	Sudhir Kumar	Python	4496	26	54
8	Telco Customer Churn-LogisticRegression	Faraz Rahman	R	4138	30	49
9	Titanic: logistic regression with python	Baligh Mnassri	Python	13239	3	41

← 報告対象

TopSE SW開発実践演習（2018/11/1向け）

課題提出日　2018/11/1
学籍番号　30014
氏名　大内一哲

題材「Logistic regression with words and char n-grams」について

■概要

- 大量（56万行）のWiki文章を解析し、毒舌の種類に応じてクラス分けする。
- TfidfVectorizer（文章内に出現する単語の出現頻度と希少性の2ファクターを用いたクラス分類アルゴリズム）が使われている。

■この題材を選んだ理由

- Voteが高く多くの人から利用されている意味で題材として相応しいと考えた
- ロジスティック回帰だけでなく、「TfidfVectorizer」もついでに学びたい

■題材のURL

https://www.kaggle.com/tunguz/logistic-regression-with-words-and-char-n-grams

TopSE SW開発実践演習（2018/11/1向け）

課題提出日　2018/11/1
学籍番号　30014
氏名　大内一哲

クラス分けの基準（毒舌キーワード）

Keyword	意味
toxic	毒舌
severe_toxic	程度の強い毒舌
obscene	卑猥
threat	脅し
insult	侮辱
identity_hate	同一性＋憎しみ？？？

TopSE SW開発実践演習（2018/11/1向け）

課題提出日　2018/11/1
学籍番号　30014
氏名　大内一哲

訓練データと評価データについて

■概要

- 訓練データ（train.csv）、評価データ（test.csv）はWikiの文章をテキスト形式で構成。
- それぞれ、60MB近くあり Excelで表示しきれないほど。
- WCコマンドで行数を調べたところ、訓練データが56万1809行、評価データが56万2889行。

■訓練データの抜粋（行頭の属性と1番目のレコードのみ）

"id","comment_text","toxic","severe_toxic","obscene","threat","insult","identity_hate"

属性

"0000997932d777bf","Explanation
Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27",0,0,0,0,0,0

レコード

TopSE SW開発実践演習（2018/11/1向け）

課題提出日　2018/11/1
学籍番号　30014
氏名　大内一哲

評価結果

■評価の流れ

・Logistic regression with words and char n-gramsからソースコードを取得し私用のMacbookProのJupyterでRUNする。

待つこと約1時間

【参考】評価PC（MacbookPro）スペック情報

CPU	i7-6567U@3.3GHz x 4コア
メモリ	16GB
GPU	Intel Iris Graphics 550 ←処理中は使われず

■出力された内容

CV score for class toxic is 0.9692156350497833
CV score for class severe_toxic is 0.9875938591018891
CV score for class obscene is 0.9838696975489344
CV score for class threat is 0.9833764206713553
CV score for class insult is 0.9774264753220822
CV score for class identity_hate is 0.9739428722139429
Total CV score is 0.9792374933179979

詳細はソースコードに記載

■ソースコードの格納場所（Github）
https://github.com/topse2018-kaggle/team

TopSE SW開発実践演習（2018/11/1向け）

課題提出日　2018/11/1
学籍番号　30014
氏名　大内一哲

今回調査で参考にした書籍やWebサイトなど

■書籍

- Python Machine Learning　2版（Sebastian Raschka & Vahid Mirjalili）

■Webサイト（Kaggle以外で）

- TfidfVectorizerのよく使いそうなオプションまとめ
  http://moritamori.hatenablog.com/entry/tfidf_vectorizer

■今後の課題

- TfidfVectorizer（文章内に出現する単語の出現頻度と希少性の2ファクターを用いたクラス分類アルゴリズム）の原理を数式レベルまで落とし込んで理解する。