

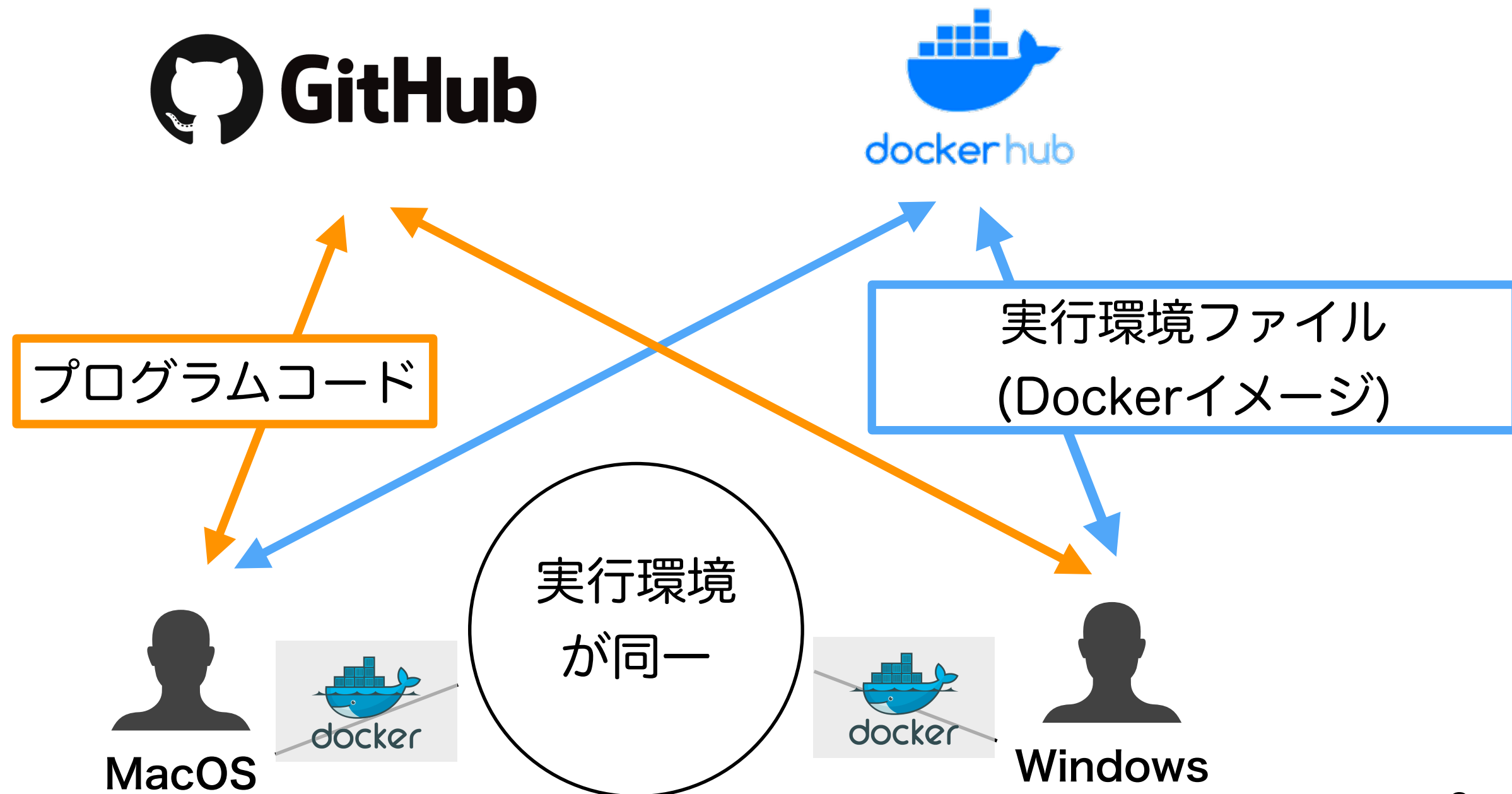
Dockerイメージサイズの削減に むけたIssue分類ツールの提案

オープンソースソフトウェア工学研究室
M2 60210045 作島大智

背景 | 対象技術

Docker

軽量なコンテナ型の仮想環境を
作成・配布・実行するためのプラットフォーム



背景

- ソフトウェア開発において開発者はDockerを盛んに利用している
 - 正確に素早く開発者間で開発環境を共有できるため
- プロジェクトは実行環境ファイルであるDockerイメージのサイズを小さくしたいと考えている
 - ex(Azure*等のクラウドサービスを用いる場合,
イメージサイズとビルド時間が大きくなるほど料金が高くなる
- 近年ではイメージサイズがどんどん小さくなっている [1]



* <https://azure.microsoft.com/ja-jp/>

[1] Lin, Changyuan & Nadi, Sarah & Khazaei, Hamzeh. (2020).
A Large-scale Data Set and an Empirical Study of Docker Images Hosted on Docker Hub.

背景

- イメージサイズを小さくするためにはイメージ内に含まれる
不必要なパッケージを削減する必要がある
- 公式Dockerイメージは大きく2種類存在する

通常
イメージ

一般的なパッケージ全てが
含まれるイメージ

スリム
イメージ

必要最低限のパッケージが
含まれるイメージ

イメージサイズに興味のある利用者は主にスリムイメージを使用する

イメージサイズが小さく，利用者が求める機能を持つDockerイメージ
を作成するにはDockerイメージをカスタマイズする必要がある

課題

- Dockerイメージをカスタマイズする
= 必要なパッケージと不要なパッケージを判別する
- スリムイメージのパッケージに関して体系的に示しているドキュメントはない
- 自身が必要とするパッケージとその依存関係等を満たすパッケージを明確にするためには関連するソースを調査し、自身で情報を収集する必要がある

課題

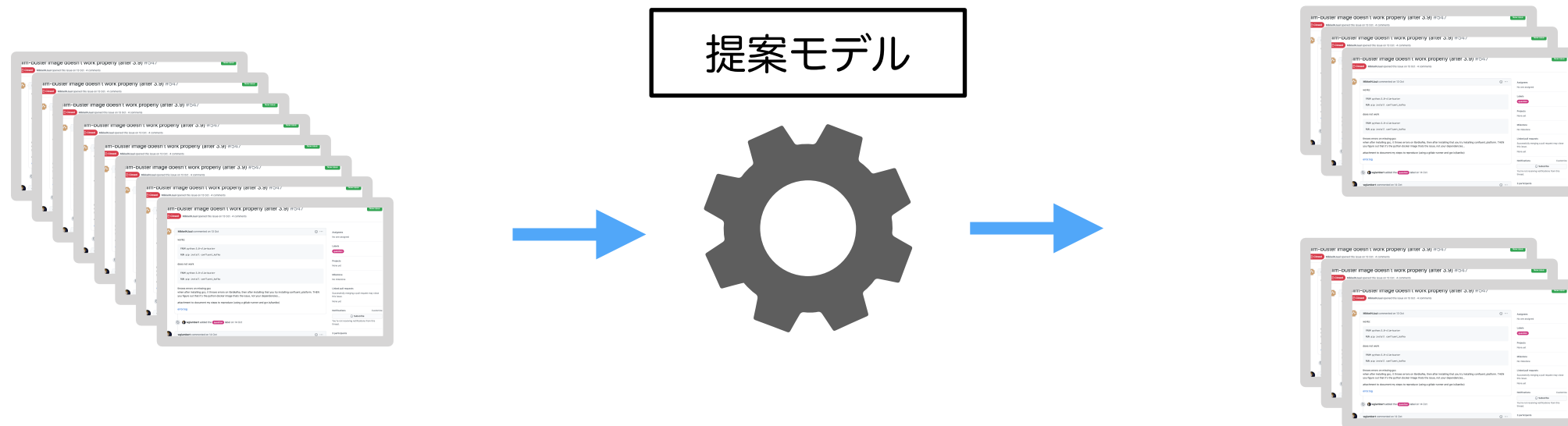
パッケージに関して調査する時間的コストが
非常に大きい

本研究の目的

目的

Dockerのパッケージ不足に関する情報を
効率よく多く集めること

→ パッケージ不足issue自動検知ツールの提案



多数のissue

パッケージ不足に関するissueと
その他のissueに分類

Issues (課題管理票)

- GitHub(共同開発環境)に実装されている機能
- 開発製品の不具合や機能改善要求などの様々な議論内容を管理

CVE-2021-23840 still in recent Debian Buster based Python images #580 タイトル

本文

質問者

回答者

コメント

コメント

アプローチ | メトリクス検討

- パッケージ不足issueを分類するモデルを構築するためには
分類に有効なメトリクスが必要になる
- 有効なメトリクスを検討するためにも、パッケージ不足issueと
その他issueの特徴の比較を以下の観点で行う
 - issueの単語数
 - コードスニペットの数
 - 会話数
 - 関わった開発者数
 - 強調文字
 - 太文字

アプローチ | 強調文字

I tried to loosen the `six` package from `six==1.10.0` to `six>=1.2,<=1.15.0`

And throws a different error for `requests` package this time. Here's the error

Conflicting with version dependencies when running pip install #585

 Closed pmventura opened this issue on 1 Mar · 3 comments



pmventura commented on 1 Mar · edited

Hi everyone,

Having issues with version dependencies when running `pip install` on docker. However, when installing on my mac without docker and just virtualenv, works perfectly fine.

Mac OS - macOS Mojave v10.14.

Python Version - v3.7.3

Docker Compose Version - version 1.27.4, build 40524192

Here's the first error I got when running the docker-compose up

```
ERROR: Cannot install -r ./python-project/requirements/base.txt (line 104), -r ./python-project/requirements/bas
```

The conflict is caused by:

```
The user requested six==1.10.0
bcrypt 3.2.0 depends on six>=1.4.1
cvxpy 1.0.25 depends on six
django-anymail 0.5 depends on six
django-compat 1.0.15 depends on six>=1.10.0
django-extensions 2.2.1 depends on six>=1.2
drf-yasg 1.16.1 depends on six>=1.10.0
fake-factory 0.6.0 depends on six
google-api-core 1.26.0 depends on six>=1.13.0
```

To fix this you could try to:

1. loosen the range of package versions you've specified
2. remove package versions to allow pip attempt to solve the dependency conflict

```
ERROR: ResolutionImpossible: for help visit https://pip.pypa.io/en/latest/user_guide/#fixing-conflicting-dependencies
ERROR: Service 'backend' failed to build: The command '/bin/sh -c pip install -r ./python-project/requirements/
```

I tried to loosen the `six` package from `six==1.10.0` to `six>=1.2,<=1.15.0`

And throws a different error for `requests` package this time. Here's the error

```
ERROR: Cannot install -r ./python-project/requirements/base.txt (line 1), -r ./python-project/requirements/base.
```

The conflict is caused by:

```
The user requested requests=2.22.0
```

目視した際に、パッケージ名が強調されていることが多く感じた

アプローチ | メトリクス検討

		Target	Other	P値 (5%)	P値 (1%)
単語数	タイトル	5.0	5.0		
単語数	本文	26.0	25.5		
単語数	コメント	70.5	44.5	*	*
コード数	本文	1.0	0.0	*	
コード数	コメント	1.0	0.0	*	*
太文字	本文	0.0	0.0		
太文字	コメント	0.0	0.0		
強調単語	本文	1.0	0.0	*	*
強調単語	コメント	3.5	1.5	*	*
会話数	全体	3.5	2.0	*	*
関わった開発者数	全体	3.0	2.0	*	*

アプローチ

○ 使用するメトリクス

- テキストデータ(タイトル, 本文, コメント)
- コメントの単語数
- コメントのコード数
- 本文の強調単語
- コメントの強調単語
- 会話数

○ 使用する機械学習アルゴリズム

- RandomForest
- SVM

データセット

- パッケージ不足であるとラベルがされているデータがない
= 自らデータを作成する必要がある
- Dockerの公式Githubリポジトリの中から
最もissue件数が多かったリポジトリのissueを目視

	パッケージ不足 issue	その他issue	計
python *	102	222	324

* <https://github.com/docker-library/python>

提案モデルの評価

○ テキストデータのみ (1-gram) + 有意差のあったメトリクス

	Precision	Recall
Random Forest	0.63	0.30
SVM	0.51	0.44

SVM	Predicted Negative	Predicted Positive
Actual Negative	18.0	4.6
Actual Positive	5.7	4.5
RF	Predicted Negative	Predicted Positive
Actual Negative	20.5	1.7
Actual Positive	7.2	3.0

まとめ

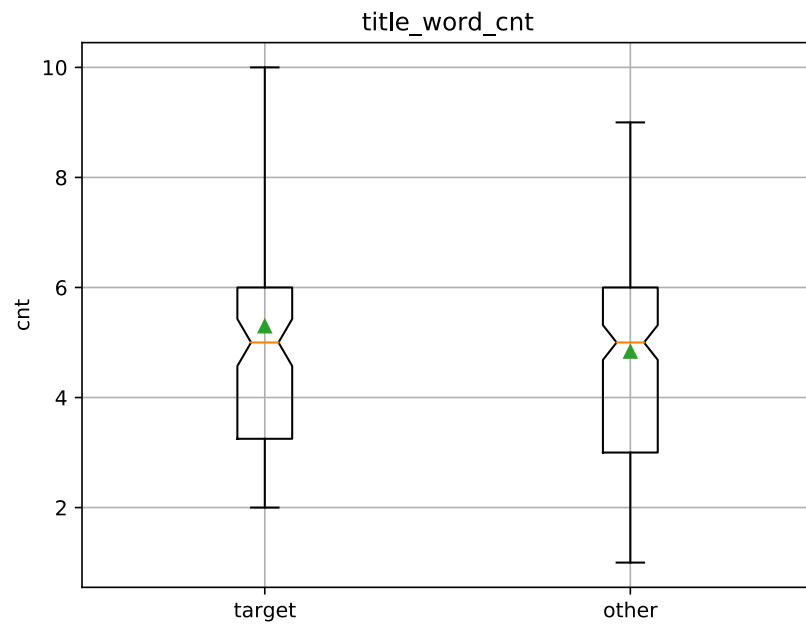
- 利用者はDockerイメージを小さくしたいと考えている
- Dockerイメージを小さくするためにパッケージを削減するが、
パッケージ不足等でエラーが生じる場合がある
- パッケージ不足に関する情報を収集するには
時間的なコストが大きく、効率的な収集方法がない
- パッケージ不足issueを効率的に収集するためのモデルを構築
 - テキストデータや有意差のあったメトリクスを使用

今後の予定

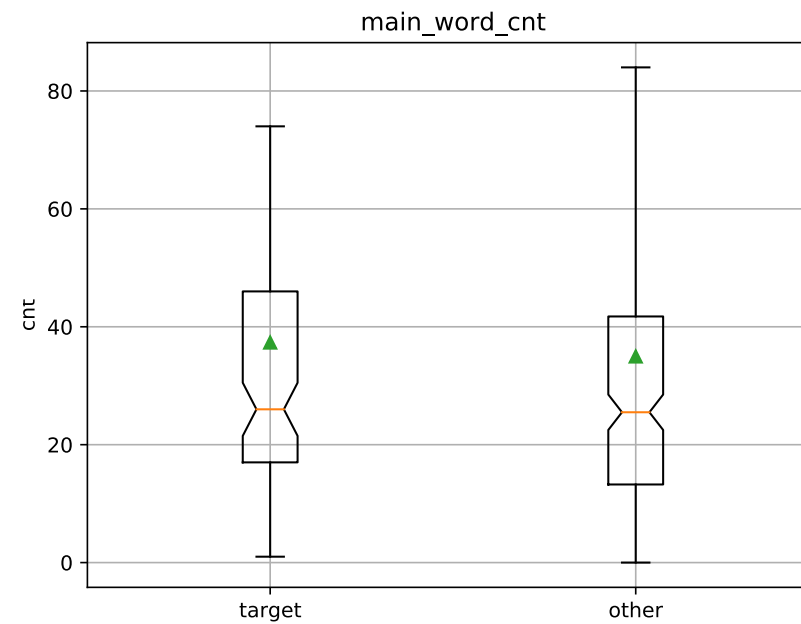
- データセットのデータ数増加
 - 現在はpythonリポジトリのみの300件程度
- 有効なメトリクスの思案
- 使用する機械学習アルゴリズムの吟味

付録

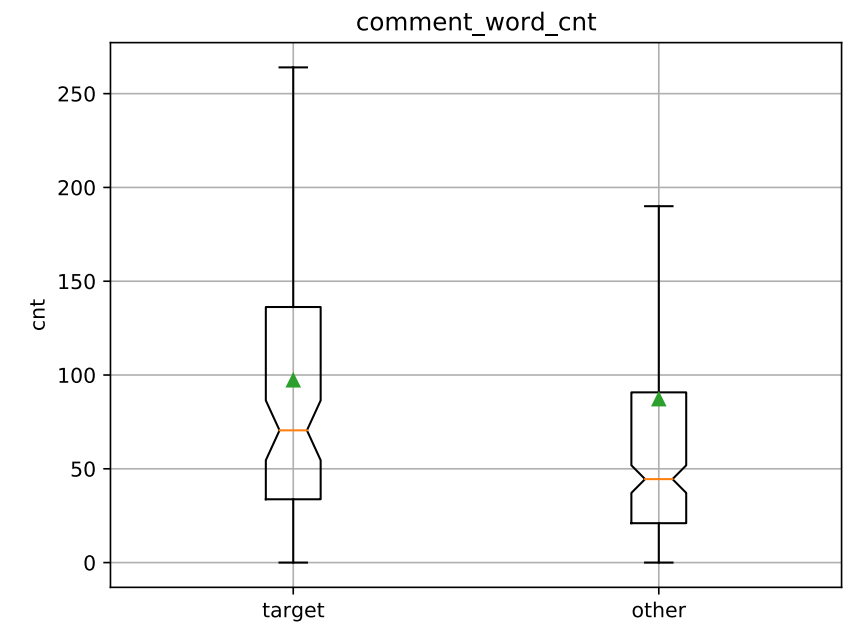
単語数



タイトル



本文



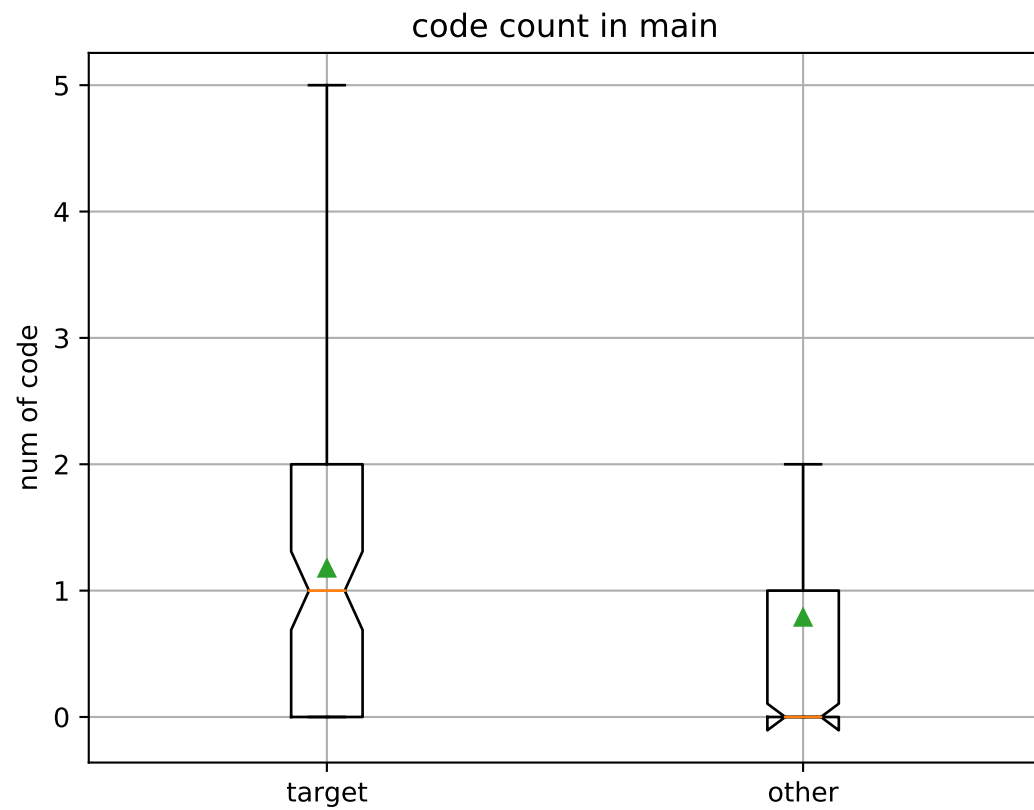
コメント

target	other	p値
5.0	5.0	0.19

target	other	p値
26.0	25.5	0.41

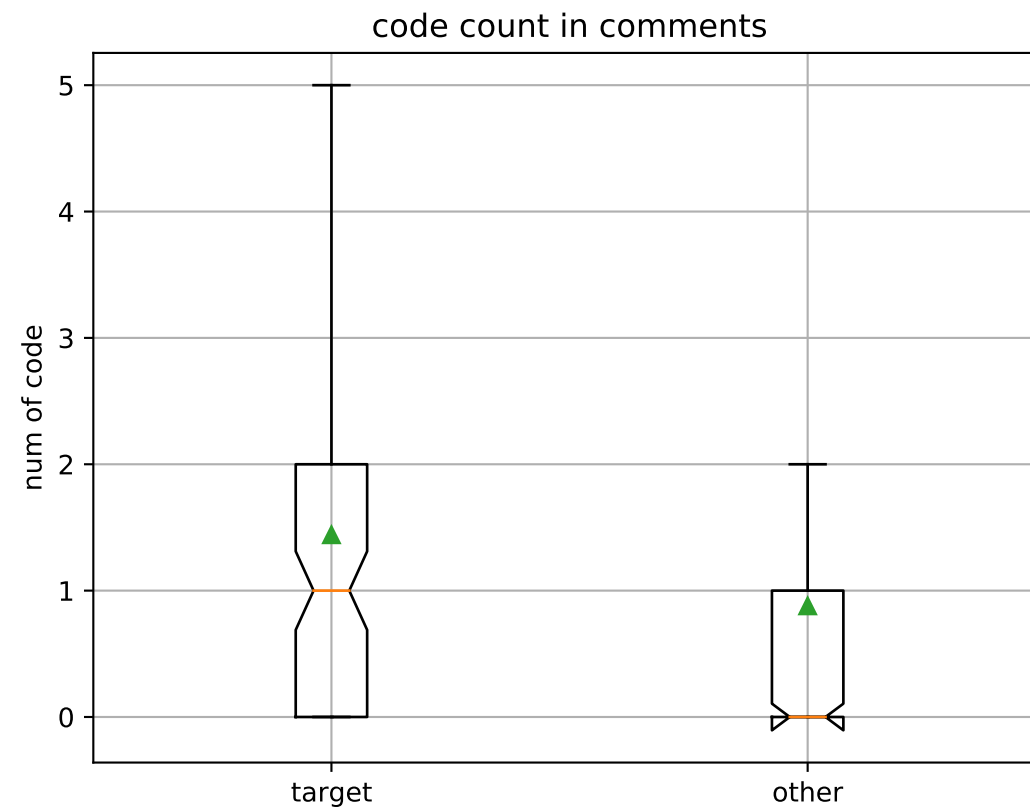
target	other	p値
70.5	44.5	0.00

コードスニペットの数



本文

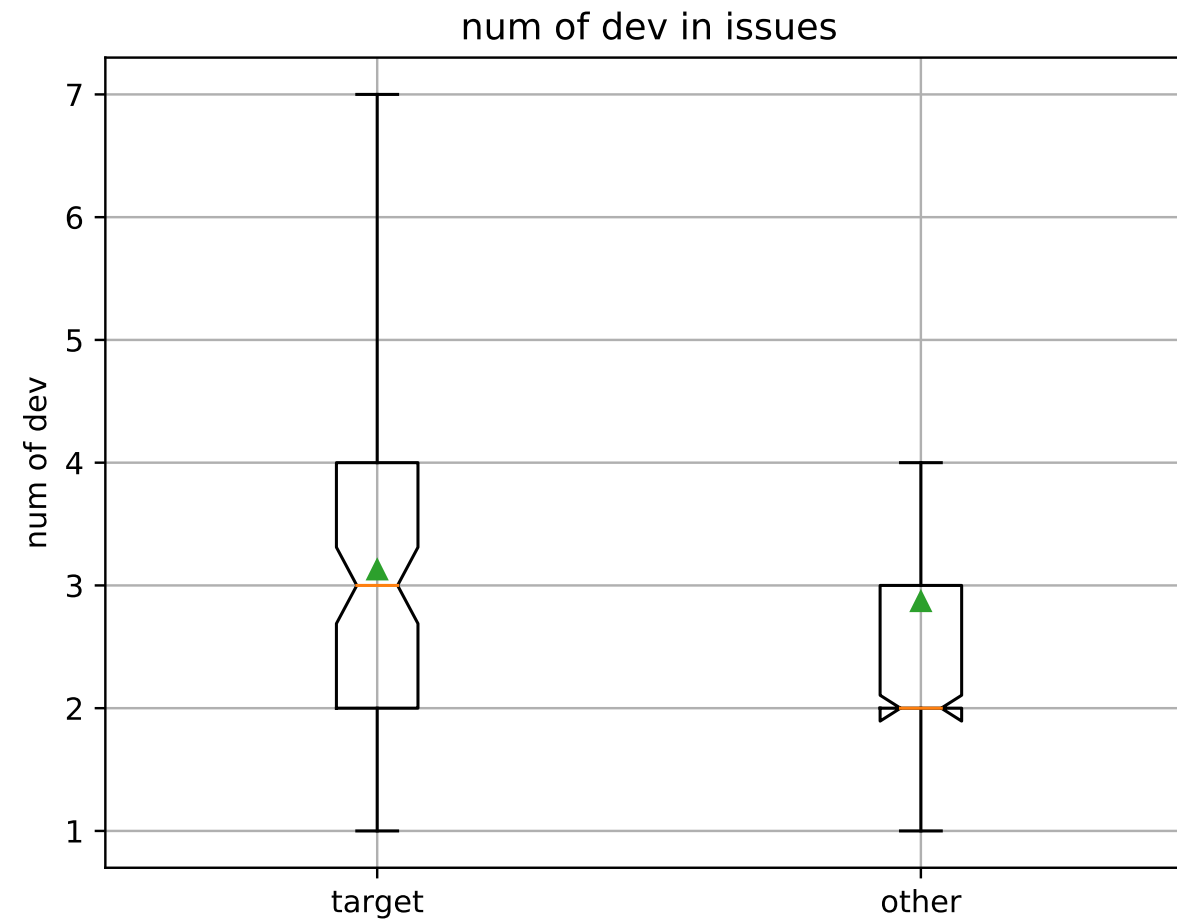
target	other	p値
1.0	0.0	0.03



コメント

target	other	p値
1.0	0.0	0.00

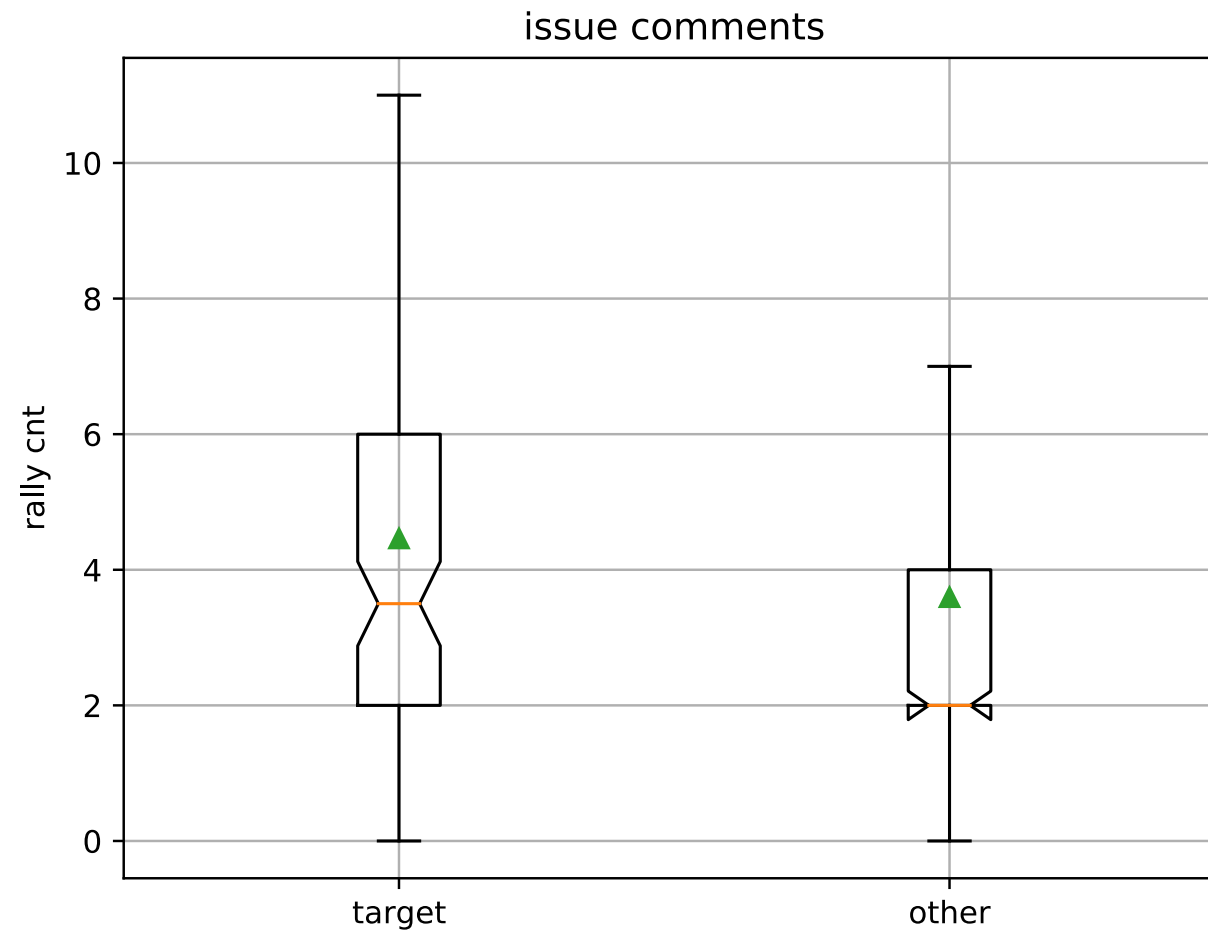
関わった開発者数



issue全体

target	other	p値
3.0	2.0	0.00

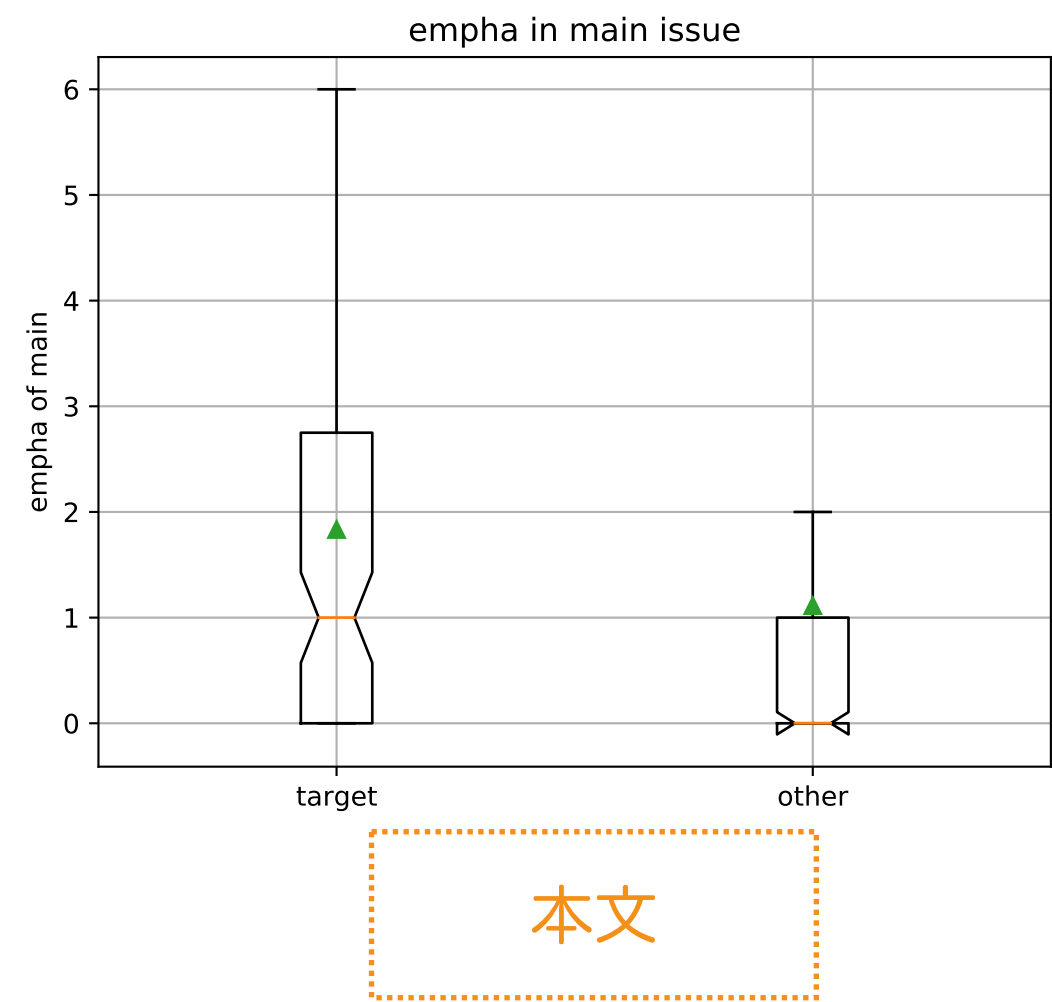
会話数



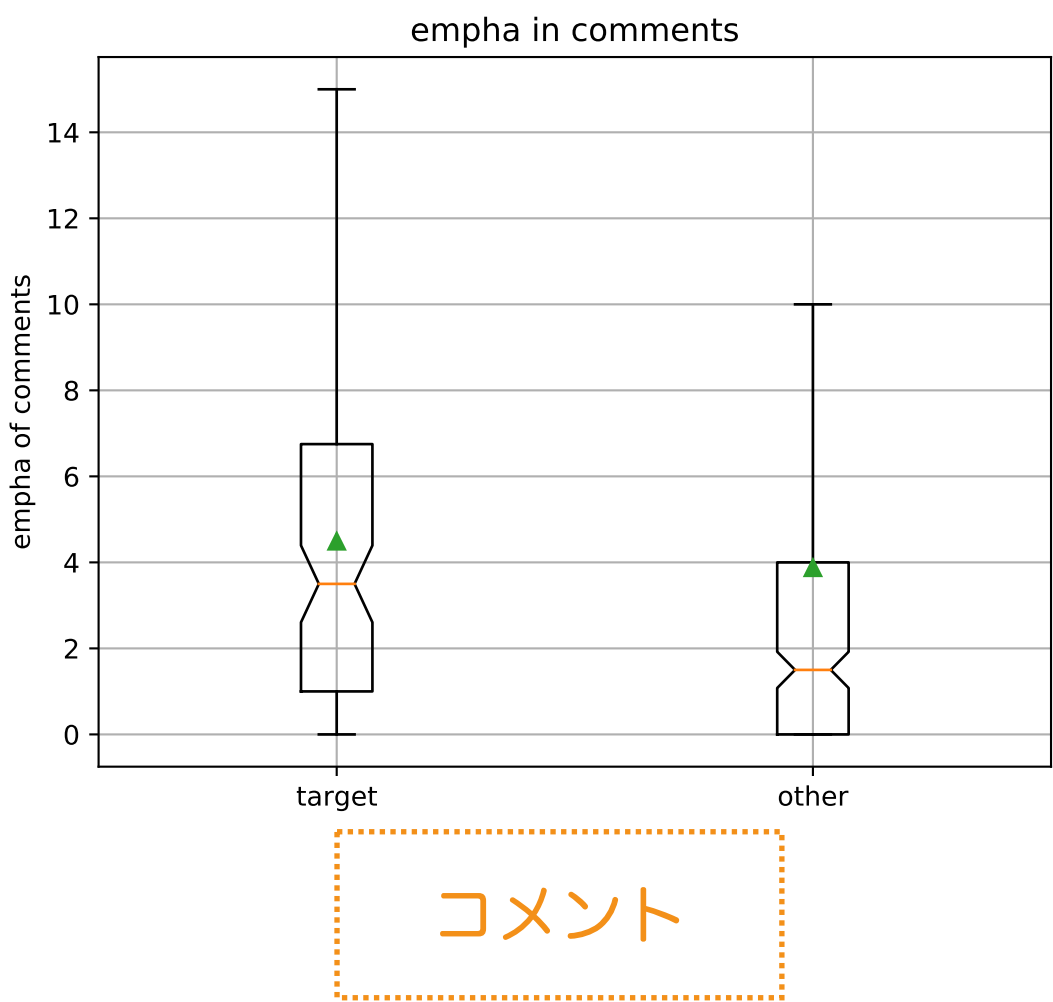
コメント

target	other	p値
3.5	2.0	0.00

強調単語

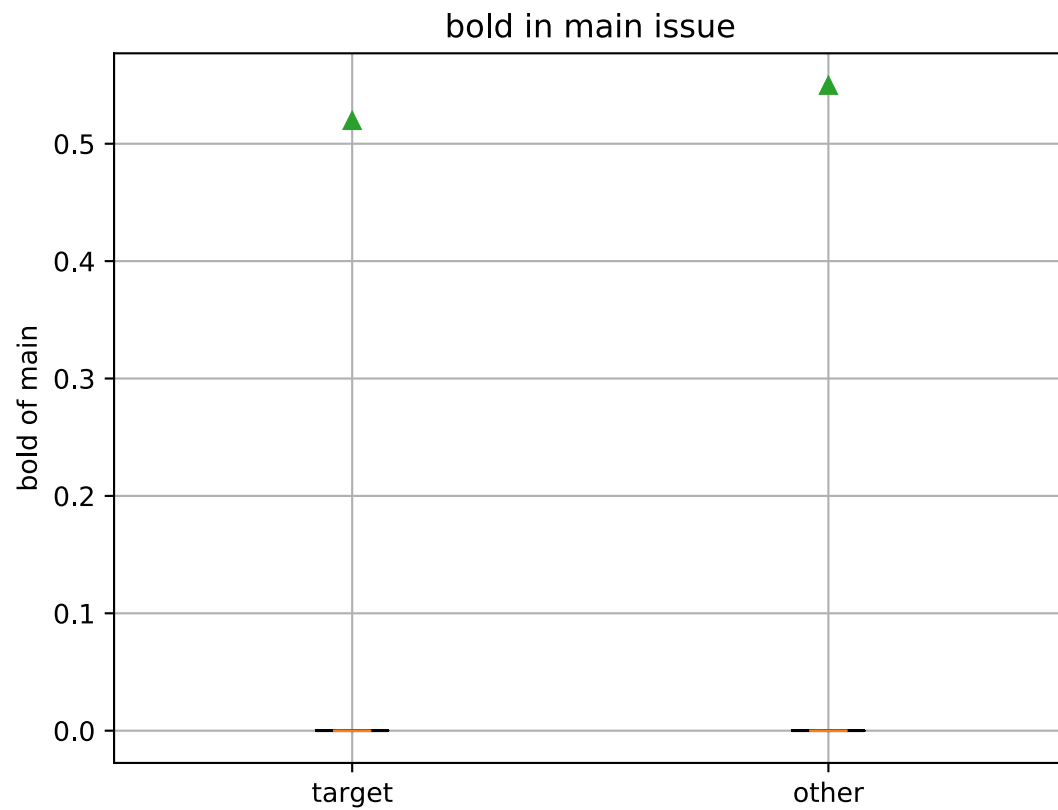


target	other	p値
1.0	0.0	0.00



target	other	p値
3.5	1.5	0.00

太文字の数



本文

target

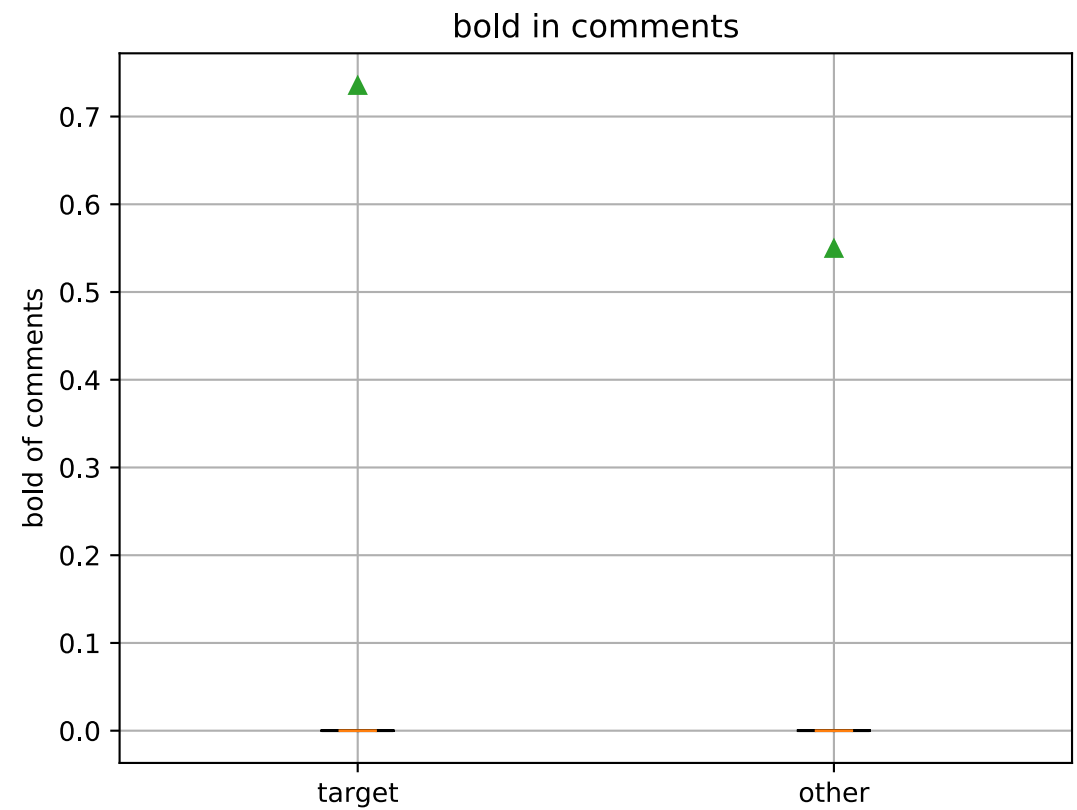
other

p値

0.0

0.0

0.59



コメント

target

other

p値

0.0

0.0

0.33