Bot Busters

Big question:
How to identify bots in WoC?

Lots of reasons why useful:

- Potentially biased statistics of WoC authors
- Census of bots / automation practices
- Human - bot interaction research
- ...

fossabot

fossabot

Follow

Your friendly neighborhood badge bot. Sends PRs to your READMEs when integrating tools from **@fossas** to track scan status. Feedback? Contact support@fossa.io!

@fossas

Dependency Heaven

support@fossa.io

http://fossa.io

Block or report user

Popular repositories

qemu-ios
Forked from nvsio/qemu-ios

QEMU-based iOS Emulator

C ★ 8

myAut2Exe

myAut2Exe - The Open Source AutoIT Script Decompiler

Visual Basic ★ 6

react-crontab-input

a crontab.guru/ replica as a react component, with i18n support

JavaScript ★ 4 ⑂ 4

MVC5HtmlTable

A simple MVC5 c#.NET Framework library to convert lists of objects into HTML tables using HtmlHelpers

C# ★ 1

php-helpers

PHP helpers influenced by underscore.js and laravel helpers.

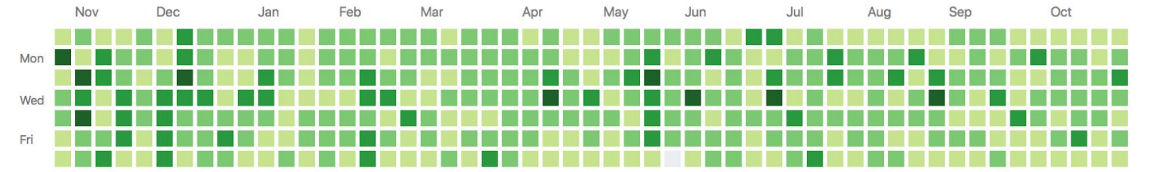
PHP ★ 1

ShopKeeper

A framework for creating and managing Shopify stores with TypeScript

TypeScript ★ 1

5,786 contributions in the last year



Codacy Badger

codacy-badger

Follow

Review less, merge faster. Check code style, security, duplication, complexity and coverage on every change while tracking code quality throughout your sprints.

Codacy

Lisbon

https://www.codacy.com

Block or report user

Overview | Repositories 1.8k | Projects 0 | Stars 0 | Followers 42 | Following 0

Popular repositories

Portfolio
My first portfolio create with Symfony 4
JavaScript ★1 1

test-dash
JavaScript ★1

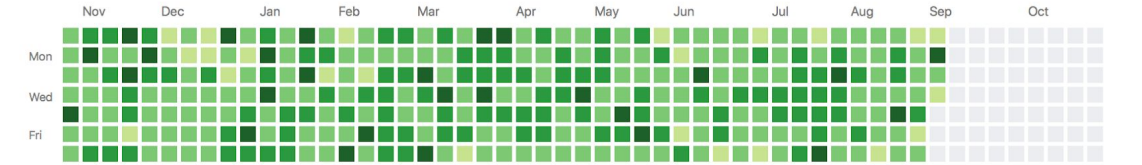
HawkBot
Forked from HawkDiscord/HawkBot
Hawk is a powerful and feature-rich Discord bot.
JavaScript ★1

steem
Forked from SteemCommunity/steem
The blockchain for Smart Media Tokens (SMTs) and decentralized applications.
C++ ★1

react-changelog
Forked from boxgames1/react-changelog
Easy include your changelog as a React component
CSS ★1 1

dart-fss
Web-scraping http://dart.fss.or.kr
Python ★1

8,696 contributions in the last year



Weekend goal:
Explore signals to identify bots

- Commit messages
- Time Stamp of the commit
- Files touched by the bot authors
  - Type of files
  - Number of files
- Number of repos to which the author opens Issues, PRs, or commits to
- The name or bio of author
  - Maybe the profile picture of the authors

Collect some ground truth data

Crawled WoC for commit authors with the term ‘bot’ (~12k total)

- *{Helper Bot <h3lperb0t@gmail.com>
- @dhis2-bot <ci@dhis2.org>
- ACG Build Bot (123D-ACG-Hickory_write) <Dinu.Bunduchi@autodesk.com>
- AUR Archive Bot <arch@carsten-teibes.de>
- Actor Bot <bot@actor.im>
- Agile CI Bot <bot@http://agile-iot.eu/>
- Anton Gustafsson Bot <antag99bot@gmail.com>
- Authenticator Bot <47196923+authenticator-bot@users.noreply.github.com>
- ...

Num distinct repos where user opened PRs

GHTorrent ~1 month in May 2019

id	login	num_repos
13810377	**snyk-bot**	4081
11000206	**pyup-bot**	1816
41425235	**regro-cf-autotick-bot**	670
32562891	**codacy-badger**	386
46203337	**hail-ci-test-1**	329
45952018	**scala-steward**	276
36774104	**fossabot**	246
2167527	BendingBender	246
46416266	hail-hephaestus	211
46062067	jenkins-x-bot-test	161
32280	olleolleolle	119
7954512	GulajavaMinistudio	107
286192	arlac77	106
210251	bastelfreak	105
58018	aschnell	103

Num distinct repos where user opened PRs

Interesting: Different results using GitHub Archive

New GitHub feature: apps (**dependabot[bot]**, **renovate[bot]**, ...)

Apps are not users! e.g., will not appear in GHTorrent

actors	login	distinct_repos
49699333	**dependabot[bot]**	514832
10810283	**direwolf-github**	64660
27856297	**dependabot-preview[bot]**	23396
39814207	**pull[bot]**	9956
19733683	**snyk-bot**	7896
29139614	**renovate[bot]**	6915
31301654	**imgbot[bot]**	5517
23040076	**greenkeeper[bot]**	3224
7571158	xmo-odoo	1916
42819689	**whitesource-bolt-for-github[bot]**	1892
16239342	**pyup-bot**	1861
37936606	**github-learning-lab[bot]**	1398
36490558	**regro-cf-autotick-bot**	1166
43237426	**jenkins-x-bot-test**	833
23717796	**depfu[bot]**	796
43047562	**scala-steward**	711
53303753	**daniel-beck-bot**	571

Explore new WoC signals

Signal:
Similarity of Commit Messages

Levenshtein Distance:

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j)+1 \\ \operatorname{lev}_{a,b}(i,j-1)+1 \\ \operatorname{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases}$$

... or simply the minimum number of character edits required to change one string into another. Very handy for automatically generated strings!

```
'Bot from 128.199.174.123 updated org.sugarlabs.StoryActivity',
'Bot from 180.200.156.187 updated org.laptop.sugar.ReadActivity',
'Bot from 128.199.174.123 updated org.worldwideworkshop.PollBuilder',
'Bot from 128.199.174.123 updated org.laptop.sugar.Jukebox',
'Bot from 128.199.174.123 updated org.laptop.TurtleArtActivity',
'Bot from 128.199.174.123 updated org.laptop.community.Spirolaterals',
'Bot from 128.199.174.123 updated org.laptop.HelpActivity',
'Bot from 127.0.0.1 updated org.sugarlabs.MazeWebActivity',
'Bot from 180.200.144.216 updated org.laptop.TurtleArtActivity',
'Bot from 128.199.174.123 updated org.laptop.sugar.ReadActivity',
'Bot from 127.0.0.1 updated me.samdroid.sugar.2048',
'Bot from 203.129.53.91 updated org.sugarlabs.BibliographyActivity',
'Bot from 128.199.174.123 updated vu.lux.olpc.Maze',
'Bot from 128.199.174.123 updated org.sugarlabs.IngeniumMachina',
'Bot from 180.200.155.64 updated org.worldwideworkshop.PollBuilder',
'Bot from 127.0.0.1 updated org.laptop.Chat',
'Bot from 128.199.174.123 updated org.laptop.ImageViewerActivity',
'Bot from 127.0.0.1 updated org.laptop.TurtleArtActivity',
'Bot from 127.0.0.1 updated org.sugarlabs.FractionBounceActivity',
'Bot from 203.129.53.91 updated org.laptop.AbiWordActivity',
'Bot from 127.0.0.1 updated vu.lux.olpc.Speak',
'Bot from 127.0.0.1 updated org.laptop.community.Spirolaterals',
'Bot from 180.200.156.187 updated org.laptop.community.Finance',
'Bot from 180.200.156.175 updated vu.lux.olpc.Maze',
'Bot from 128.199.174.123 updated org.sugarlabs.FractionBounceActivity',
'Bot from 180.200.158.1 updated org.sugarlabs.MusicKeyboard',
'Bot from 180.200.156.175 updated org.laptop.Terminal',
'Bot from 180.200.178.84 updated org.laptop.community.TypingTurtle',
'Bot from 180.200.158.1 updated org.worldwideworkshop.PollBuilder',
'Bot from 128.199.174.123 updated org.sugarlabs.MusicKeyboard',
'Bot from 128.199.174.123 updated me.samdroid.sugar.slides',
```

Signal:
Similarity of Commit Messages

Turns out, this naive approach works pretty well...

In [217]:
```
print(subset['score'][90])
subset['commit_msg'][90]
```

```
98.6
```

Out[217]:
```
array(['Updating cli to 2.0.0-preview3-006761 and shared runtime to 2.0.0-preview3-25513-02',
       'Updating cli to 2.0.0-preview3-006785 and shared runtime to 2.0.0-preview3-25516-01',
       'Updating cli to 2.0.0-preview3-006796 and shared runtime to 2.0.0-preview3-25518-01',
       'Updating submodule(s)\\n\\nEntityFrameworkCore => c81d716cc27013d6ff5ca3eb5fa60cdbaa782bd3 KestrelHttpServer
=> 7d712f58aae2b78305992d45b60d66fee383a2ca\\n\\n[auto-updated: submodules]',
       'Updating cli to 2.0.0-preview2-006391 and shared runtime to 2.0.0-preview3-25413-01',
       'Updating cli to 2.0.0-preview3-006736 and shared runtime to 2.0.0-preview3-25511-03',
       'Updating cli to 2.0.0-preview3-006770 and shared runtime to 2.0.0-preview3-25516-01',
       'Updating cli to 2.0.0-preview3-006764 and shared runtime to 2.0.0-preview3-25514-02',
       'Updating cli to 2.0.0-preview1-005783 and shared runtime to 2.0.0-preview1-001967-00',
       'Updating cli to 2.0.0-preview1-005893',
       'Updating cli to 2.0.0-preview3-006750 and shared runtime to 2.0.0-preview3-25513-02',
       'Updating shared runtime to 2.0.0-preview1-001894-00',
       'Updating cli to 2.0.0-preview3-006800 and shared runtime to 2.0.0-preview3-25518-01',
       'Updating cli to 2.0.0-preview3-006729 and shared runtime to 2.0.0-preview3-25510-01',
       'Updating CLI to 1.0.0-preview3-003178',
       'Updating cli to 2.0.0-preview1-005825 and shared runtime to 2.0.0-preview1-002059-00',
       'Updating cli to 2.0.0-preview3-006770 and shared runtime to 2.0.0-preview3-25515-01',
       'Updating cli to 2.0.0-preview3-006770 and shared runtime to 2.0.0-preview3-25514-02',
       'Updating shared runtime to 1.2.0-beta-001299-00',
       'Updating cli to 2.0.0-preview3-006655 and shared runtime to 2.0.0-preview3-25502-01',
       'Updating cli to 2.0.0-preview3-006734 and shared runtime to 2.0.0-preview3-25511-03',
       'Updating cli to 2.0.0-preview3-006841 and shared runtime to 2.0.0',
       'Updating submodule(s)\\n\\nHttpAbstractions => 3e3772eecd4cc57399c28a3f899e6b0406ef2e1b\\n\\n[auto-updated: s
ubmodules]',
       'Updating submodule(s)\\n\\nEntityFrameworkCore => 3885333f04d22a9ea75ec9e8357deb02ef74d112 JsonPatch => 70c81
33fce523b6eb42f39cc4da78e9ac11e99dc Razor => 801ad075601be8f3d04ff49a0ddf411fb43d32e6\\n\\n[auto-updated: submodule
s]',
       'Updating CLI to 1.0.0-preview2-003118',
       'Updating cli to 2.0.0-preview2-006461 and shared runtime to 2.0.0-preview3-25413-01',
       'Updating CLI to 1.0.0-preview2-003024',
       'Updating submodule(s)\\n\\nSignalR => 06475270ec845b5be95c46857d33599e41585e86\\n\\n[auto-updated: submodule
s]',
       'Updating cli to 2.0.0-preview3-006787 and shared runtime to 2.0.0-preview3-25516-01',
       'Updating shared runtime to 1.2.0-beta-001289-00',
       'Updating cli to 2.0.0-preview1-005825 and shared runtime to 2.0.0-preview1-002061-00',
       'Updating cli to 2.0.0-preview1-005807 and shared runtime to 2.0.0-preview1-005807 and 2.0.0-preview1-002028-0
0',
       'Updating cli to 2.0.0-preview2-006349 and shared runtime to 2.0.0-preview2-25406-03',
       'Updating cli to 2.0.0-preview1-005917',
       'Updating submodule(s)\\n\\nJavaScriptServices => 2d98a1808ce49f9696c61099689e022d23774c75\\n\\n[auto-updated:
submodules]',
       'Updating cli to 2.0.0-preview3-006766 and shared runtime to 2.0.0-preview3-25514-02',
       'Updating cli to 2.0.0-preview3-006783 and shared runtime to 2.0.0-preview3-25516-01',
```

In [211]:
```
print(subset['score'][5])
subset['commit_msg'][5]
```

```
42.4
```

Out[211]:
```
array(['THE BASIC APP IS DONE, PAGE IS CENTERED NOW. THE CENTERING ISSUE WAS THE AUTOLAYOUT PARAMS.',
       'Initial Commit', 'UPDATED FONT.', 'UPDATED WITH A TITLE.',
       'UPDATED FONT', 'Changed UI colors.',
       'FINISHED FIRST VERSION. LOAD JPG FILES INTO THE JPG FOLDER. THE APP WILL DISPLAY ALL THE JPG IMAGES IN THAT F
OLDER IN A TABLE VIEW.',
       'UPDATED WITH A PRETTIER UI AND ROUNDED CORNERS ON BUTTONS.'],
      dtype=object)
```

In [214]:
```
print(subset['score'][38])
subset['commit_msg'][38]
```

```
25.5
```

Out[214]:
```
array(['Cleaned up h1-h5 tags',
       'Merge d64d3e1bbe43dee8e5ae1e0bd4b81572c0cf14aa into 69fa384bd3d751685a26d92823f3bb8e38f831f6',
       'Merge a1edb6ee93d62f8608bc3381166e9d1a1a06313a into 69fa384bd3d751685a26d92823f3bb8e38f831f6'],
      dtype=object)
```

Commit timestamps

Bots should have:

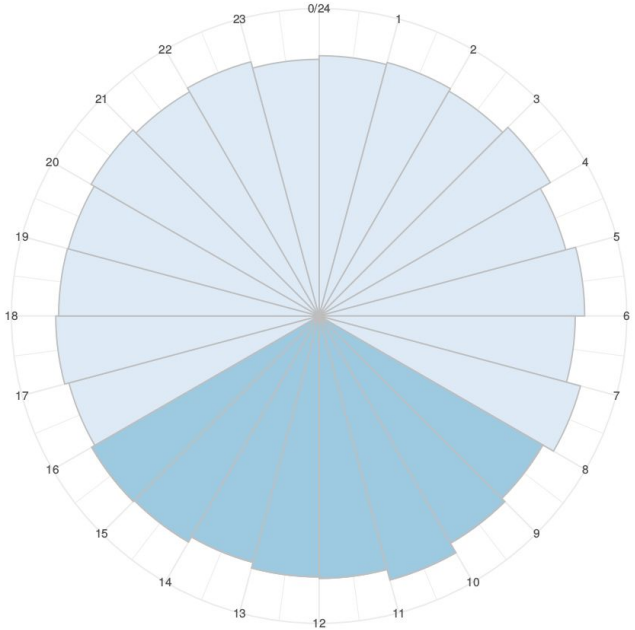
- Either round the clock activity, or
- Activity at specific times of the day

Humans should have:

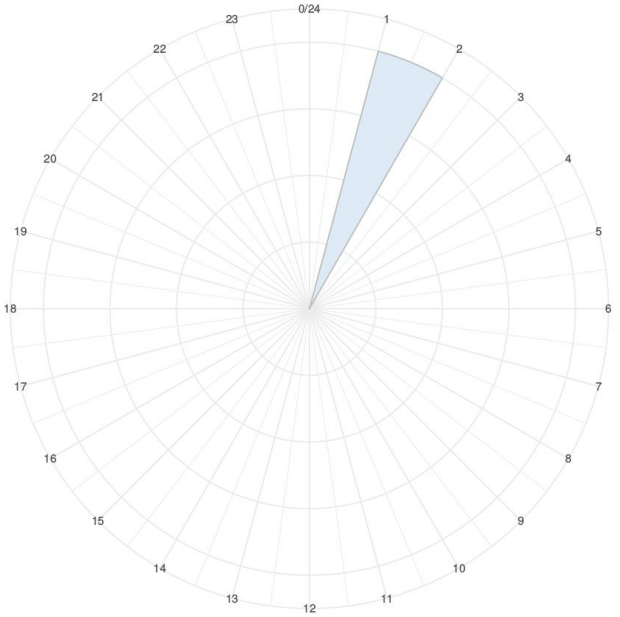
- Activity during the day, likely concentrated on the working hours

Example of activities of suspected bots

Ninja.org i18n Bot <long.huynh+bot@autonomous.nyc>

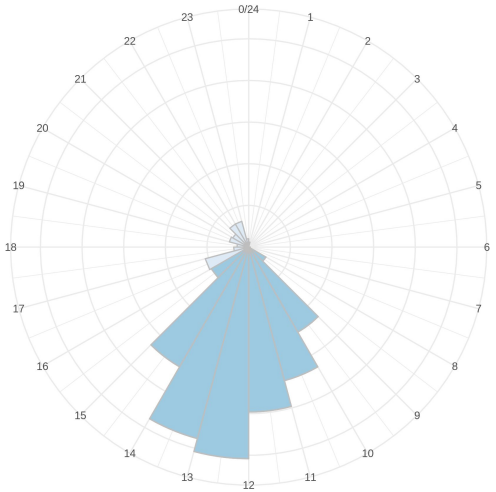


A Bot <bot@example.com>

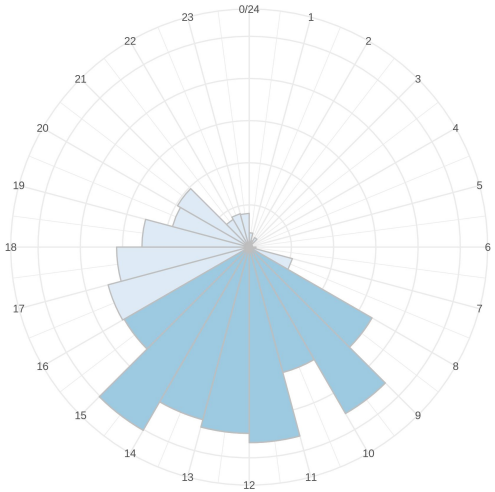


Example of activities of suspected non-bots

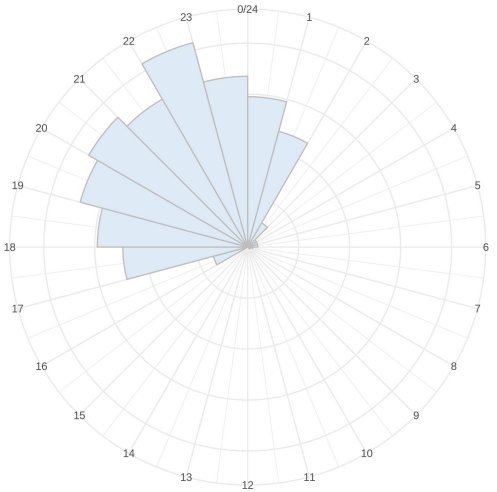
Dinesh497 <dinesh.bhagwandin@hva.nl>



mnissler@chromium.org <mnissler@chromium.org@0039d316-1c4b-4281-b5



Danielle Roberts <1051443@85c56323-6fa9-3386-8a01-6480fb634889>



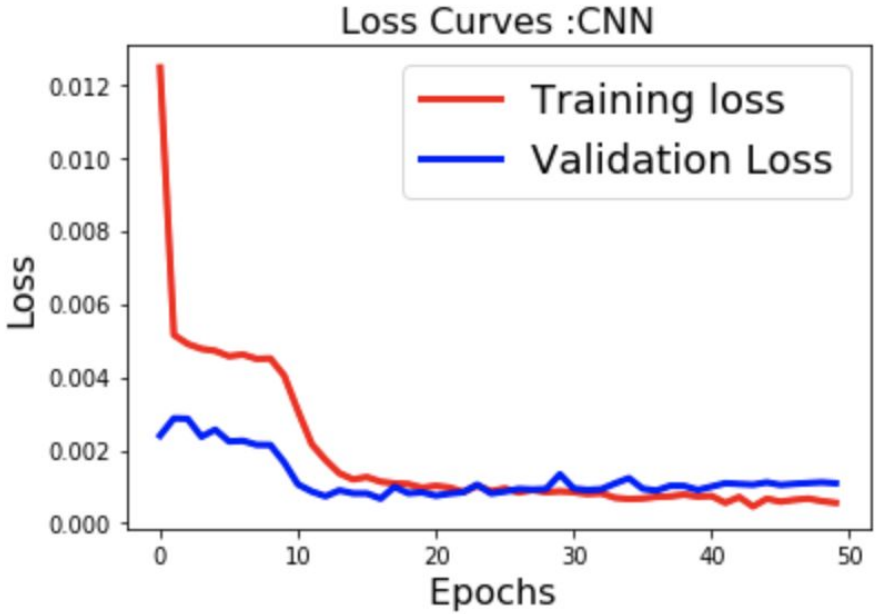
Files Touched by (suspected) bots vs. non-bots

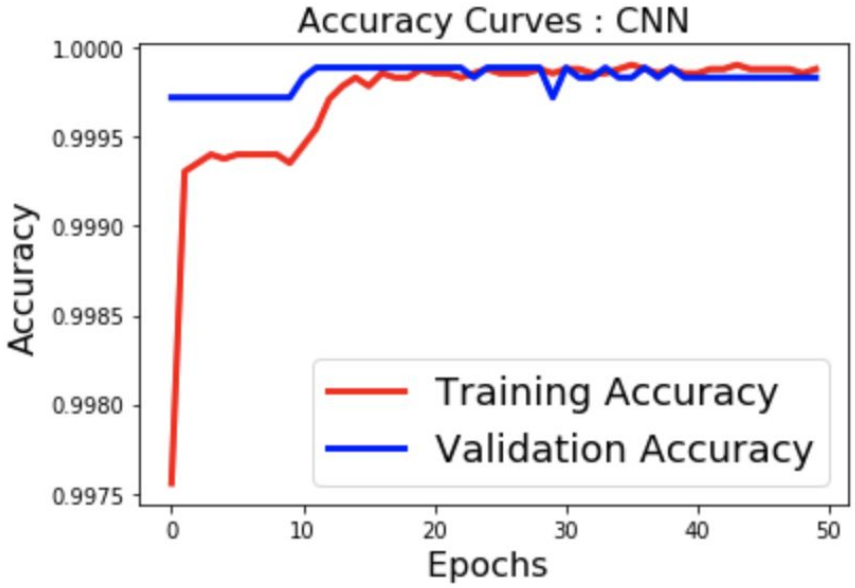
Bots

Non-Bots

Learning the patterns via CNN

- Extract the training data
  - A subset of bot commit message
  - A subset of human commit messages
- Generate embeddings for the selected messages
  - GloVe: Obtaining vectors representation for words
- Train a CNN network on the selected data


Loss Curves :CNN
Training loss
Validation Loss
Loss
Epochs
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
10
20
30
40
50


Accuracy Curves : CNN
Training Accuracy
Validation Accuracy
Accuracy
Epochs
1.0000
0.9995
0.9990
0.9985
0.9980
0.9975
0
10
20
30
40
50