Researchers compile a new database of executable Python code snippets on GitHub

August 23 2018, by Ingrid Fadelli

```python
import requests
import json

urlbase = 'http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address='
urlend = 'Zurich,Switzerland'

r = requests.get(urlbase+urlend) # request to google maps api

r=r.json()
if r.get('results'):
    for results in r.get('results'):
        latlong = results.get('geometry','').get('location','')
        latitude = latlong.get('lat','')
        longitude = latlong.get('lng','')
        break
    print latitude, longitude

else:
    print 'No results'
```

a: https://gist.github.com/10017416

```
FROM python:2.7.13
ADD snippet.py snippet.py
RUN ["pip", "install", "requests"]
CMD ["python", "snippet.py"]
```

b: Dockerfile

Figure 1: (a) Code snippet for using the Google Maps geocode API. (b) Dockerfile containing environment specification required to run code snippet.

(a) Code snippet for using the Google Maps geocode API (b) Dockerfile containing environment specification required to run code snippet. Credit: Horton & Parnin

A team of researchers at North Carolina State University has recently carried out an empirical analysis of the executable status of Python code snippets shared on GitHub. Their study, pre-published on arXiv, also presents Gistable, a new database of executable Python code snippets on

GitHub's gist system, which could enable reproducible studies in the field of software engineering.

Every day, software developers worldwide create and share code online to demonstrate and outline new programming concepts. GitHub is one of the largest online platforms on which developers can share their code snippets and collaborate on the development of software. Currently, it contains over 300,000 Python snippets and over 4.5 million gists in a variety of programming languages.

While code snippets published online can be very useful, sometimes they are not directly executable by others. This might be due to parse errors in the code or to issues with executing snippets in environments that contain unmet dependencies.

To gain a better understanding of how many code snippets hosted on GitHub's gist system are actually executable, researchers at North Carolina State University conducted a thorough evaluation of the executability of publicly available Python scripts hosted on the platform. Their study was aimed at identifying common issues with the execution of code snippets, which could provide valuable insight for further research on automated software configuration management.

In their study, the researchers also presented Gistable, a database and extensible framework built on GitHub's gist system. Gistable contains 10,259 Python code snippets, of which approximately 5,000 come with a Dockerfile to configure and execute them without import error.

"Our work on Gistable was motivated as part of a larger project concerning automated configuration of application environments," Eric Horton, one of the researchers who carried out the study, told Tech Xplore. "Given a codebase, such as the snippets studied in Gistable, we want to find a process which can build a sufficient execution

environment for them without requiring input from a developer. In order to do this, we first had to step back and answer a couple questions. First, is this a common use case? We needed to establish a baseline for how often existing applications need some sort of non-trivial configuration. Second, when not executable, what type of configuration is needed to enable execution?"

In their study, the researchers found that 75.6 percent of analyzed Python gists required substantial configurations to overcome issues such as missing dependencies, configuration files, reliance on a specific operating system, or other environment configuration challenges. In addition, the assumptions that developers make about resource names when trying to resolve configuration errors were found to be correct less than half of the time.

"We found that around 30 percent of our sample fell into the 'hard to configure' category, with the most common configuration difficulty being dependencies on external libraries," Horton explained. "Our research in the immediate future will focus on techniques for finding and installing these libraries. Afterward, we hope to address other common configuration difficulties discovered as part of Gistable."

Overall, an insufficiently configured environment was the primary factor preventing the Python code snippets from being executable. While in some cases, correct application environment configurations could be recovered automatically, others required further interventions. In future, the researchers plan to investigate strategies to consistently perform effective environment configurations.

"I think the most meaningful achievement of this study was our investigation into how developers perform configuration manually," Horton said. "Not only did the responses from participants confirm that this is in many cases a hard problem, but they also helped us categorize

things that can make configuration difficult. This is very useful, because it points us at a concrete list of items for future research."

More information: Gistable: Evaluating the Executability of Python Code Snippets on GitHub. arXiv:1808.04919v1 [cs.SE]. arxiv.org/abs/1808.04919

Abstract

Software developers create and share code online to demonstrate programming language concepts and programming tasks. Code snippets can be a useful way to explain and demonstrate a programming concept, but may not always be directly executable. A code snippet can contain parse errors, or fail to execute if the environment contains unmet dependencies.

This paper presents an empirical analysis of the executable status of Python code snippets shared through the GitHub gist system, and the ability of developers familiar with software configuration to correctly configure and run them. We find that 75.6% of gists require non-trivial configuration to overcome missing dependencies, configuration files, reliance on a specific operating system, or some other environment configuration. Our study also suggests the natural assumption developers make about resource names when resolving configuration errors is correct less than half the time.

We also present Gistable, a database and extensible framework built on GitHub's gist system, which provides executable code snippets to enable reproducible studies in software engineering. Gistable contains 10,259 code snippets, approximately 5,000 with a Dockerfile to configure and execute them without import error. Gistable is publicly available at this URL: github.com/gistable/gistable