Mining GitHub and JIRA Repositories

TAPAJIT DEY

Task Overview

- Repositories to be mined:
  a. Apache-Camel --- open-source integration framework based on known Enterprise Integration Patterns.
  b. Apache-Cassandra --- open-source, distributed, no-SQL database
  c. Apache-Derby --- open-source Relational Database Management System (RDBMS)

- Additional constraint:
  a. Present findings about any commit with a message including the word ‘fix’

Mining commits from GitHub

- Using GitHub API (v3), and personal access token
- Input: Repo name, access token
- Output: JSON objects for all commits separated by newline in gzipped file
- Process:
  a. Get rate-limit info, and, at each step
     i. make sure to have more than 20 queries left
     ii. Wait until reset time if fewer than 20 queries left
  b. Get commit page for repo, extract last page from link (header), query each page sequentially
  c. Save failed queries, retry to get them once more
  d. Save each commit JSON object in separate lines in a gzipped output file
- Code: Python3 code --- GHminer.py

Filtering Commits with ‘fix’ in message

1. Read the JSON objects saved earlier to check if commit message has the word “fix”: `re.search(r'^fix|\sfix', msg, re.I)`

2. Save commit_sha, author_name, author_email, author_github_id, commit_timestamp, parent_commit_sha in a gzipped csv file

3. Code: Python3 --- GH_data_process.py

Commit Mining Result: as of 17th Feb. 2020

Project	Camel	Cassandra	Derby
No. of Commits	42696	25013	8269
No. of Commits with “fix” in the message	8845	3270	1538
Earliest commit with “fix” in the message	2007-03-19	2009-03-18	2004-09-30

Mining JIRA Repositories for Issues

- Using “jira” package in Python for querying the JIRA database, with username-password based authentication.
- Input: JIRA URL, project names, username, password
- Output: a gzipped csv file with issue related variables
- Process:
  a. Obtain all issues for a project, 1000 at a time, until query result comes as empty
  b. Extract related variables for each issue and save them in a gzipped csv file
  c. Any “,” in description and summary replaced with “;” for consistency in parsing result.

Mining JIRA Repositories for Issues

Variables extracted for each issue:

assignee name	assignee username	components	created	creator name
creator username	description	fixVersions	issuetype	key
last viewed	priority	project	reporter name	reporter username
resolution	resolution date	status description	status name	subtask
summary	time spent	updated		

Mining JIRA Repositories for Issues: Result (as of 17th Feb. 2020)

Project	Camel	Cassandra	Derby
No. of Issues	14527	15476	7056
Issues Unresolved	468	2225	1260
Earliest Issue	2007-04-18	2009-03-07	2004-09-24