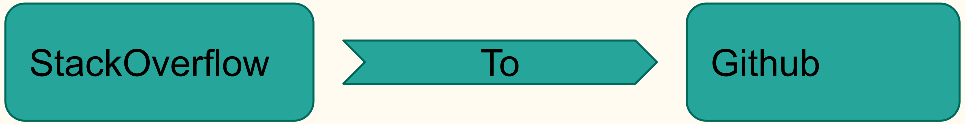
Tracking of Flow of Code from Stackoverflow to Github

Rutwik Kulkarni
Kulendra Kumar Kaushal

CS 5704: Software Engineering, Spring 2020
Instructor: Dr. Na Meng
Department of Computer Science, Virginia Tech
Blacksburg, VA 24061

Problem Statement

We plan to conduct an empirical study to investigate the presence of Stack Overflow snippets in the source code of popular public Github repositories.

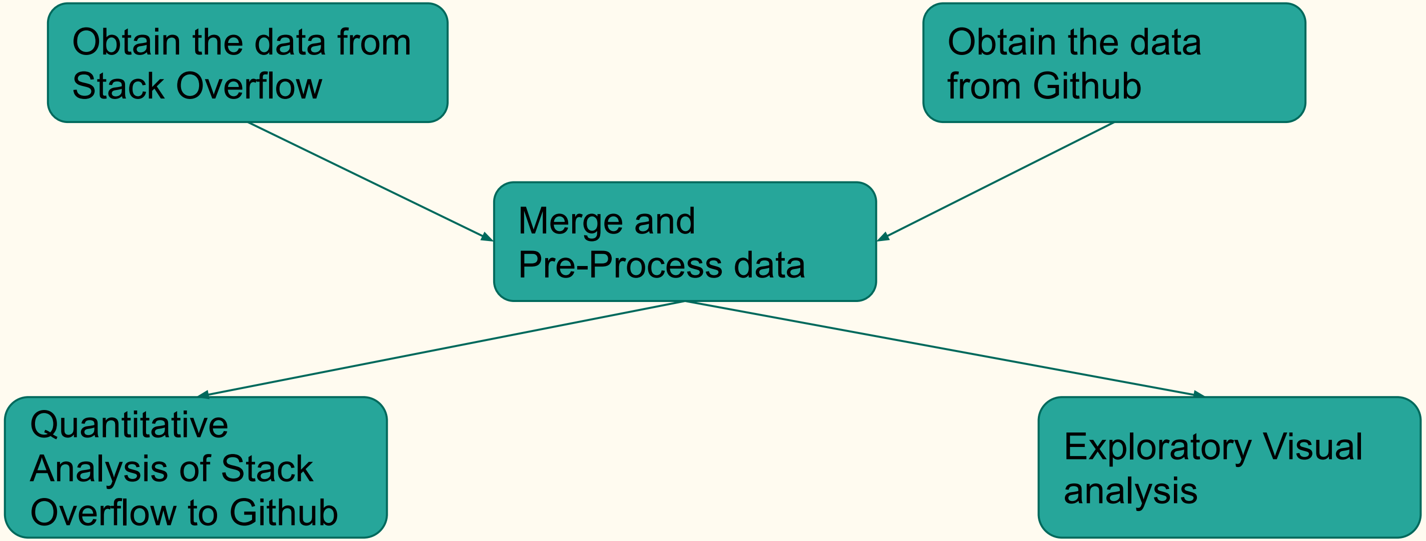


Research Questions

This Study will help to address the following Research Questions

- RQ1: To find the most referenced answer to any Stack Overflow question and recommend it to the developer
- RQ2: To analyze whether the most favourite answer is the most referenced answer
- RQ3: To analyze the extent to which the code is copied from Stack Overflow to Github
- RQ4: Cross language analysis of the above research questions.

Proposed Solution



Obtain the data from Stack Overflow

Obtain the data from Github

Merge and Pre-Process data

Quantitative Analysis of Stack Overflow to Github

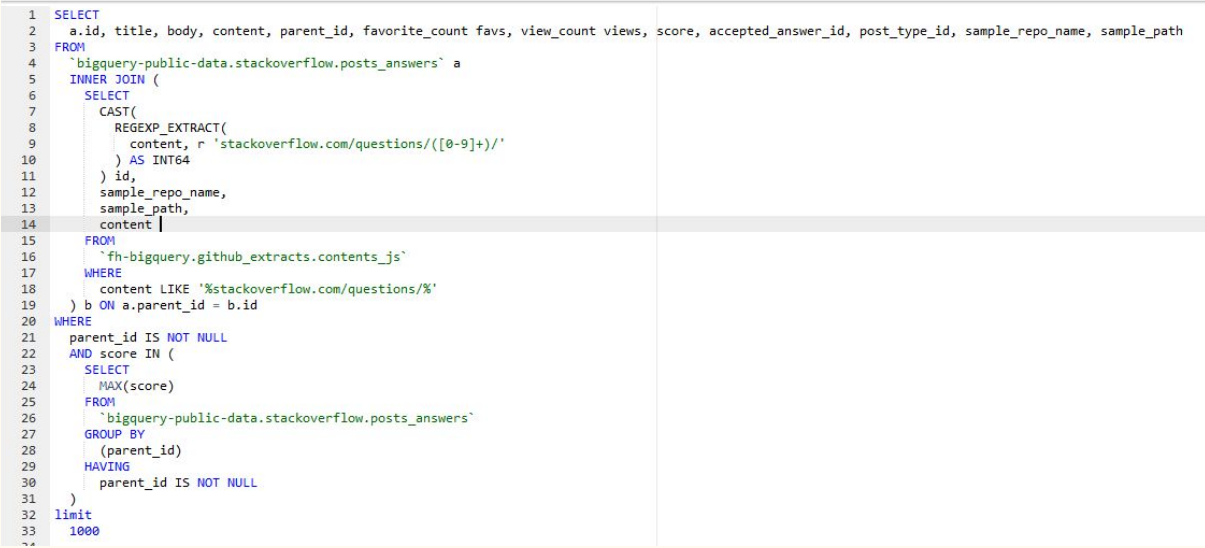
Exploratory Visual analysis

STEP 1: Dataset Creation

- Multiple open data corpus of StackOverflow dump and Github extracts are available.
- We decided to choose Google BigQuery public datasets over Stack Exchange Data Explorer[SEDE]
- SEDE can only output up to 50,000 rows
- SEDE protects itself from abuse with CAPTCHAs, and has no API.
- Plenty of other datasets shared on BigQuery including Github extracts.
- BigQuery allows us to add our own dataset and take multiple joins.
- Used Google BigQuery to extract most upvoted answer for each stack overflow question.
- Obtained all Github file referencing the above StackOverflow posts.

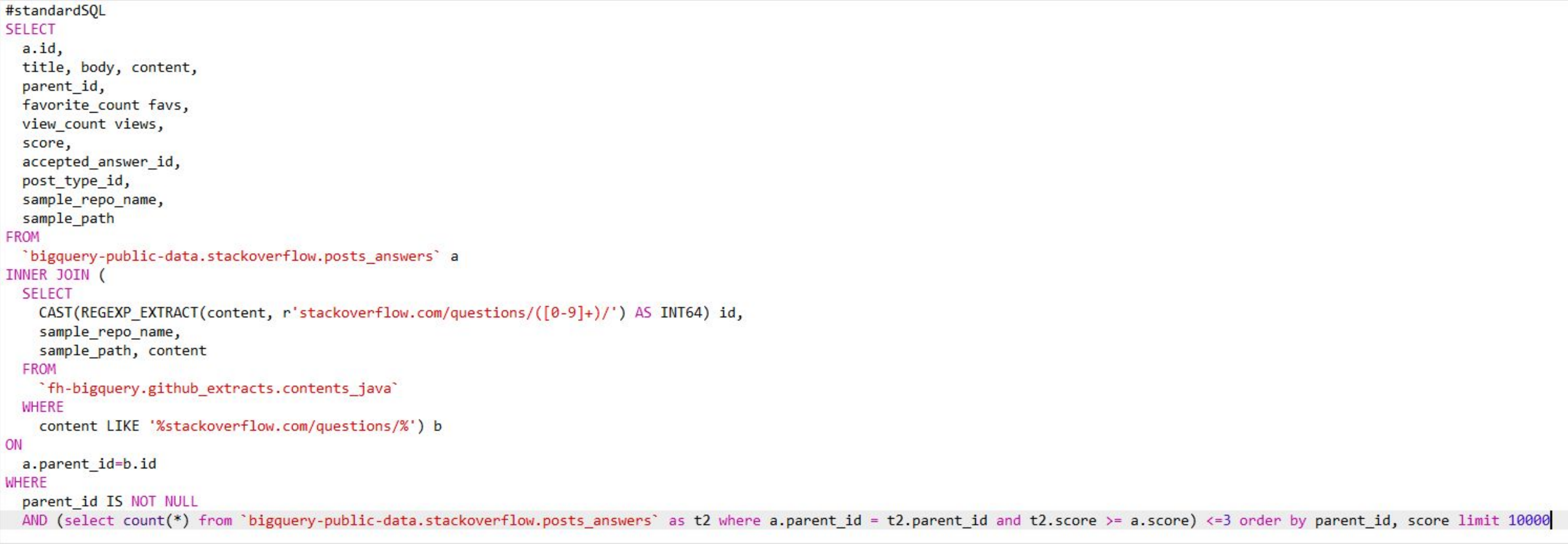
Query used for analysis

To extract the most favourite answer from stackoverflow and the reference of its corresponding question on github we used the following query:



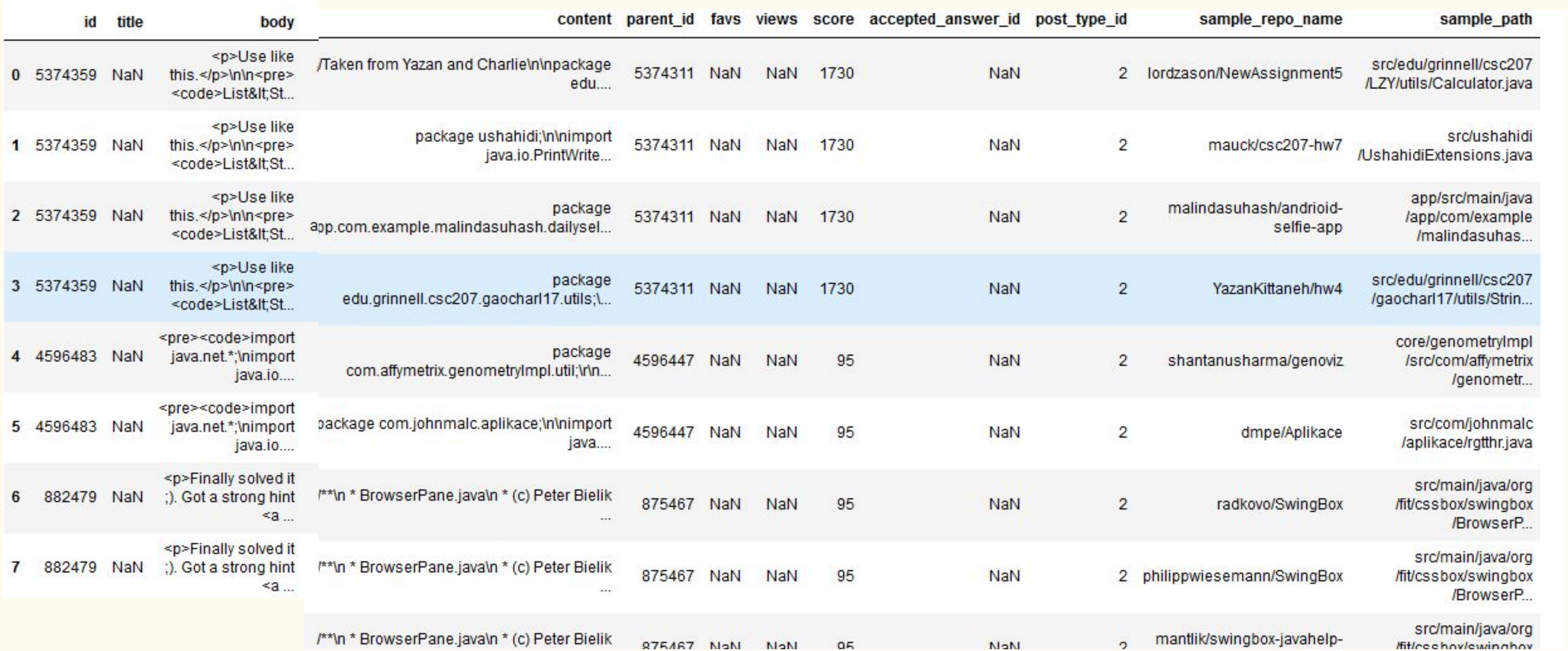
```sql
SELECT
  a.id, title, body, content, parent_id, favorite_count favs, view_count views, score, accepted_answer_id, post_type_id, sample_repo_name, sample_path
FROM
  `bigquery-public-data.stackoverflow.posts_answers` a
  INNER JOIN (
    SELECT
      CAST(
        REGEXP_EXTRACT(
          content, r 'stackoverflow.com/questions/([0-9]+)/'
        ) AS INT64
      ) id,
      sample_repo_name,
      sample_path,
      content 
    FROM
      `fh-bigquery.github_extracts.contents_js`
    WHERE
      content LIKE '%stackoverflow.com/questions/%'
  ) b ON a.parent_id = b.id
WHERE
  parent_id IS NOT NULL
  AND score IN (
    SELECT
      MAX(score)
    FROM
      `bigquery-public-data.stackoverflow.posts_answers`
    GROUP BY
      (parent_id)
    HAVING
      parent_id IS NOT NULL
  )
limit
  1000
```

Query for Extracting multiple answers



```
#standardSQL
SELECT
  a.id,
  title, body, content,
  parent_id,
  favorite_count favs,
  view_count views,
  score,
  accepted_answer_id,
  post_type_id,
  sample_repo_name,
  sample_path
FROM
  `bigquery-public-data.stackoverflow.posts_answers` a
INNER JOIN (
  SELECT
    CAST(REGEXP_EXTRACT(content, r'stackoverflow.com/questions/([0-9]+)/') AS INT64) id,
    sample_repo_name,
    sample_path, content
  FROM
    `fh-bigquery.github_extracts.contents_java`
  WHERE
    content LIKE '%stackoverflow.com/questions/%') b
ON
  a.parent_id=b.id
WHERE
  parent_id IS NOT NULL
  AND (select count(*) from `bigquery-public-data.stackoverflow.posts_answers` as t2 where a.parent_id = t2.parent_id and t2.score >= a.score) <=3 order by parent_id, score limit 10000
```

Sample Dataset Obtained from the Query.



	id	title	body	content	parent_id	favs	views	score	accepted_answer_id	post_type_id	sample_repo_name	sample_path
0	5374359	NaN	<p>Use like this.</p>\n\n<pre><code>List&lt;St...	/Taken from Yazan and Charlie\n\npackage edu....	5374311	NaN	NaN	1730	NaN	2	lordzason/NewAssignment5	src/edu/grinnell/csc207/LZY/utils/Calculator.java
1	5374359	NaN	<p>Use like this.</p>\n\n<pre><code>List&lt;St...	package ushahidi;\n\nimport java.io.PrintWrite...	5374311	NaN	NaN	1730	NaN	2	mauck/csc207-hw7	src/ushahidi/UshahidiExtensions.java
2	5374359	NaN	<p>Use like this.</p>\n\n<pre><code>List&lt;St...	package app.com.example.malindasuhash.dailysel...	5374311	NaN	NaN	1730	NaN	2	malindasuhash/andrioid-selfie-app	app/src/main/java/app/com/example/malindasuhas...
3	5374359	NaN	<p>Use like this.</p>\n\n<pre><code>List&lt;St...	package edu.grinnell.csc207.gaocharl17.utils;\...	5374311	NaN	NaN	1730	NaN	2	YazanKittaneh/hw4	src/edu/grinnell/csc207/gaocharl17/utils/Strin...
4	4596483	NaN	<pre><code>import java.net.*;\nimport java.io....	package com.affymetrix.genometryImpl.util;\r\n...	4596447	NaN	NaN	95	NaN	2	shantanusharma/genoviz	core/genometryImpl/src/com/affymetrix/genometr...
5	4596483	NaN	<pre><code>import java.net.*;\nimport java.io....	package com.johnmalc.aplikace;\n\nimport java....	4596447	NaN	NaN	95	NaN	2	dmpe/Aplikace	src/com/johnmalc/aplikace/rgtthr.java
6	882479	NaN	<p>Finally solved it ;). Got a strong hint <a ...	/**\n * BrowserPane.java\n * (c) Peter Bielik ...	875467	NaN	NaN	95	NaN	2	radkovo/SwingBox	src/main/java/org/fit/cssbox/swingbox/BrowserP...
7	882479	NaN	<p>Finally solved it ;). Got a strong hint <a ...	/**\n * BrowserPane.java\n * (c) Peter Bielik ...	875467	NaN	NaN	95	NaN	2	philippwiesemann/SwingBox	src/main/java/org/fit/cssbox/swingbox/BrowserP...
				/**\n * BrowserPane.java\n * (c) Peter Bielik	875467	NaN	NaN	95	NaN	2	mantlik/swingbox-javahelp-	src/main/java/org/fit/cssbox/swingbox

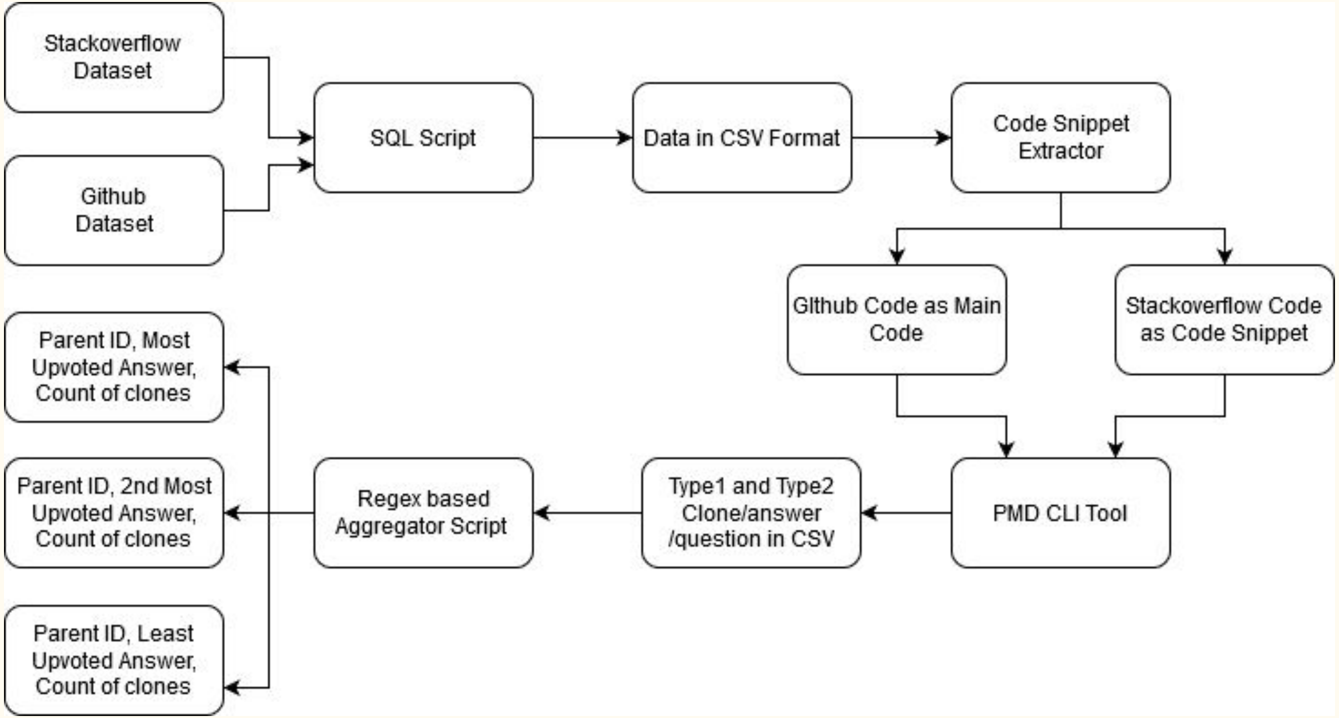
STEP 2: Data Preprocessing

- StackOverflow post contents are present in XML format and code snippet is present between code-tag. e.g.

```
'<p>Use like this.</p>\n\n<pre><code>List&lt;String&gt; stockList = new
ArrayList&lt;String&gt;();\nstockList.add("stock1");\nstockList.add("stock2");\n\nString
[] stockArr = new String[stockList.size()];\nstockArr =
stockList.toArray(stockArr);\n\nfor(String s : stockArr)\n
System.out.println(s);\n</code></pre>'
```

- Github code is present in normal text format.
- Unknown characters removal from the data.
- Each code snippet from stack overflow answer and github code repo stored in a separate file for further analysis.

Implemented Architecture

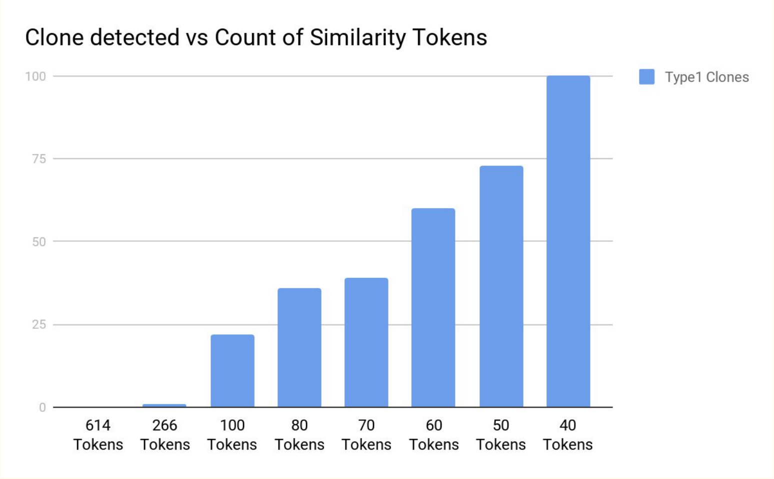

Stackoverflow Dataset
Github Dataset
SQL Script
Data in CSV Format
Code Snippet Extractor
GIthub Code as Main Code
Stackoverflow Code as Code Snippet
PMD CLI Tool
Type1 and Type2 Clone/answer /question in CSV
Regex based Aggregator Script
Parent ID, Most Upvoted Answer, Count of clones
Parent ID, 2nd Most Upvoted Answer, Count of clones
Parent ID, Least Upvoted Answer, Count of clones

STEP 3: Clone Detection

- Used PMD 5.4.1 duplicate code detector to identify the presence of type 1 and type 2 clones.
- PMD does file to file matching for clone detection.
- PMD detects clone based on the number of similar tokens.
- Clone is defined based on optimal threshold value for number of similar tokens.
- Used PMD to detect SO to Github and Github to Github clone detection.
- Manual validation of the results obtained.

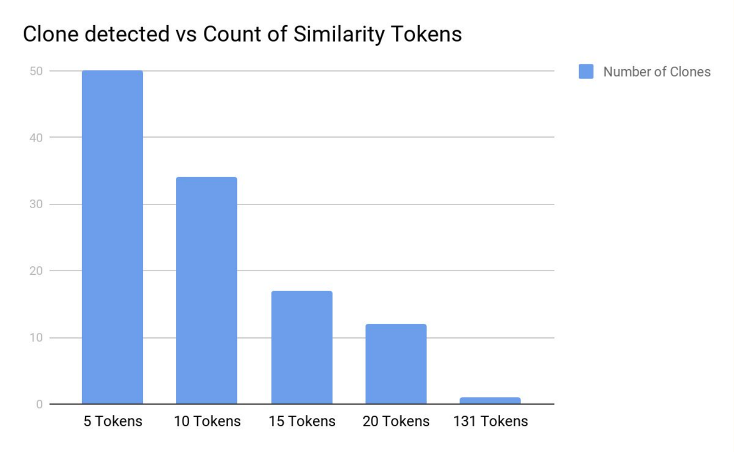
RESULTS: Java Type 1 Clone

- Total 1000 Github files referencing StackOverflow questions analyzed.
- 334 Unique questions were referred.
- Clones found for different number of matching tokens.
- Our initial choice of tokens for estimating closes was 10 based on manual observation.



RESULTS: Javascript Type 1 Clone

- Total 1000 Github files referencing StackOverflow questions analyzed.
- 212 Unique questions were referred.
- Total Clones(not grouped by questions) found for different number of matching tokens.
- Our initial choice of tokens for estimating closes was 10 based on manual observation.



Rosetta Baseline Tokens For Different Tasks

Program	Java	Java Script	Ratio
BST	502	146	3.43
Hello World-Text	35	8	4.375
Hello World-Server	154	70	2.2
Hello World Graphical	417	8	52.125
Swap	74	39	1.8974
While Loop	47	34	1.382
For loop	77	89	0.8651
MatMul	277	338	0.819
Null obj	26	44	0.59

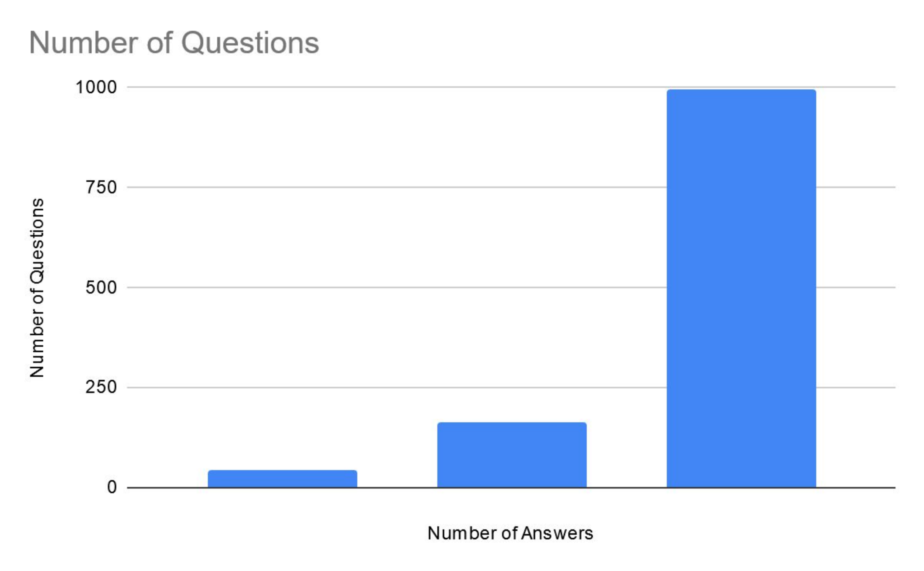
Median of ratios of Java to Javascript = 1.89

Rosetta Baseline Tokens For Different Tasks

1. From the previous manual analysis, we found that the copied program with more than 10 tokens in javascript made sense
2. Additionally as the smallest program that can be copied is Hello World and it contains 8 tokens in JavaScript
3. Therefore on the basis of based on previous manual observation and Hello world analysis 2 programs are considered to be clones when number of tokens are equal to 10 (Approximation of 8)
4. For java we calculated number of tokens so that a program can be considered as clones based on median ratio i.e 10*1.89 ~ 19 tokens

Results :Java

1. A new data of 10,000 questions containing multiple answers of the same question and their references was queries
2. Basic Analysis of the data:
    a. Unique Questions :1200
    b. Unique Answers: 3552
    c. Questions with only 1 answer 43
    d. Questions with 2 answers 162
    e. Questions with 3 answers 995



Results : Type 1 Clones Java

1. Out of all the questions analyzed 66 Questions had their most upvoted answer referenced once or more.
2. 56% (37/66) of times most upvoted answer has highest number of referenced github repositories while 31% (20/66) of times, 2nd most upvoted answer has higher number of referenced github repositories for the same question.
3. It was very interesting to know that for 13% (9/66) questions the least favourite Answer has highest number of referenced github repositories as compared to other answers of the same question.

Results: Type 2 Clone Java

1. Out of all the questions analyzed 47 Questions had some or the other answer referenced
2. Most favourite Answer was referenced 63.8% (30/66) number of times
3. For 36.17% (17/47) of questions the most referenced answer was not the most favorite one
4. It was very interesting to know that for 19% (9/47) questions the least favourite Answer is the most referenced.

Analysis: Java

- Two code snippets are being concluded to be type 1 clone if number of matching tokens are 19 or more.
- Two code snippets are being concluded to be type 2 clone if number of matching tokens are 30 or more.
- For type 2 clones, We didn’t chose 19 because PMD requires a very high computing resource.
- Out of 3300 referenced github repositories, only 66 (20 when minimum tokens was 30) SO Questions had referenced answers were type 1 while 47 question’s answers were type 2 cloned in github repositories => RQ3

Analysis: Java continued

- Found most referenced answer to the question =>RQ1
- For both type 2 and type 1 clones, most upvoted answer need not always be most referenced => RQ2

Results :Javascript

1. A new data of 1000 questions containing multiple answers of the same question and their references was queries
2. Basic Analysis of the data:
   a. Unique Questions : 212
   b. Unique Answers: 227

Analysis: Javascript

- Two code snippets are being concluded to be type 1 clone if number of matching tokens are 10 or more.
- Out of 1000 referenced github repositories, only 34(total) were type 1 clones. => (Partly RQ2)
- Out of 1000 referenced github repositories, only 17 SO Questions had referenced answers were type 1 cloned in github repositories => RQ3
- PMD doesn’t allow us to find type 2 clones for Javascript => (Partly RQ2)

Cross Language Analysis (RQ4)

- Based on empirical study conducted using tasks and their codes available on Rosetta in different programming languages, 2 tokens of Javascript is equivalent to 1 token of Java.
- For Java, 5.5% (66/1200) of the questions have their most upvoted answer referenced once or more while in case of Javascript, the percentage is 15%(17/112).
- As data is taken from a normal distribution. Considering Rossetta equivalent token as for Java and Javascript as baseline, we can say that for Javascript most upvoted answer is referenced more as compared to Java.

Challenges

1. Normal string matching of stackoverflow answers and github extracts doesn’t give any good results
2. PMD does not support Type 2 clones for Javascript
3. Limit on download of data for Google BigQuery (Free account)
4. StackOverflow and Github are getting updated constantly, so there could be a lack in consistency of the data. (i.e The answer which is mentioned to be most upvoted the dataset, may have more upvotes in dataset or may not be the topmost answer)

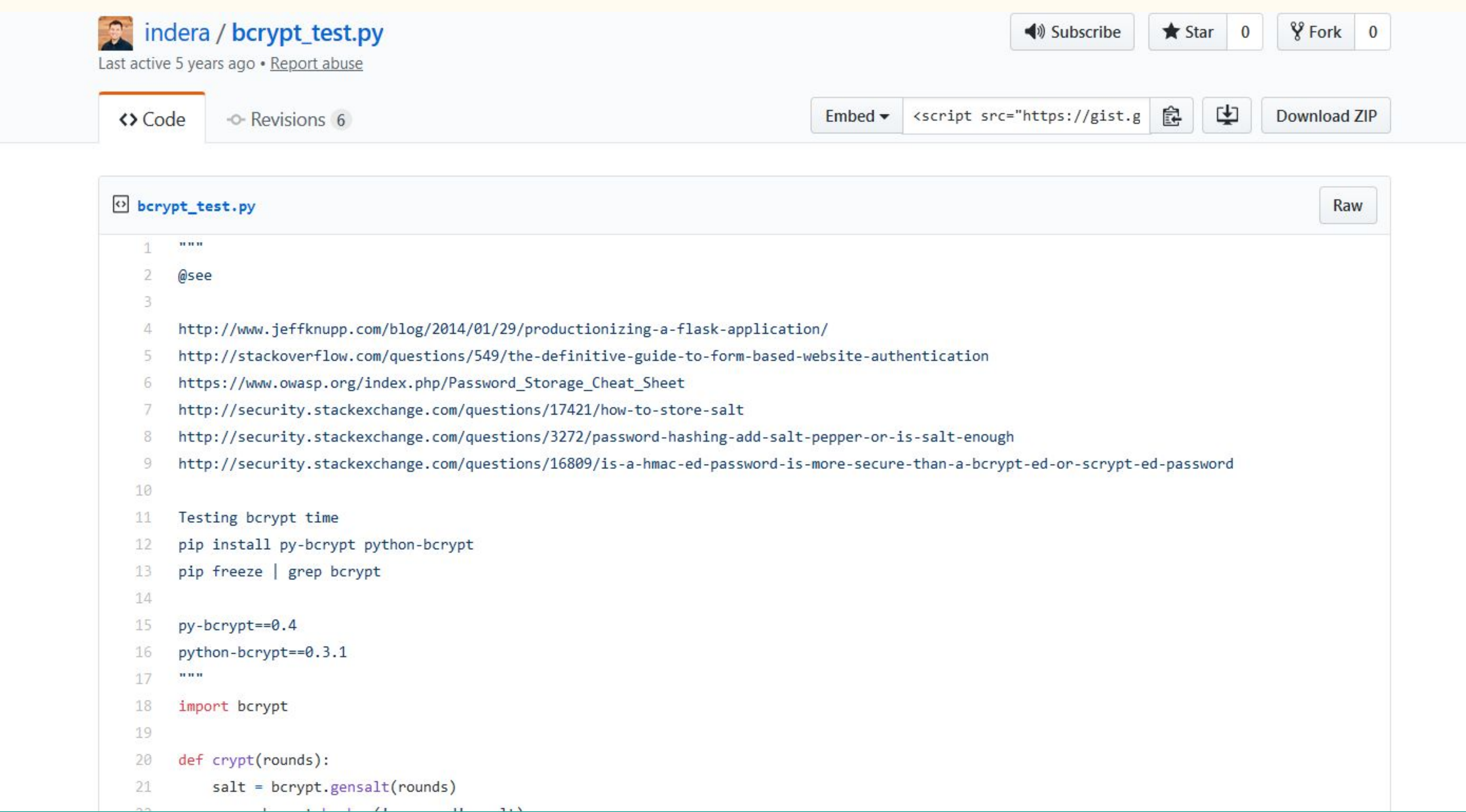
Future work

1. Analyze the flow of code from the github repositories to stackoverflow answers
2. Machine learning based clone detection using TF-IDF/ text similarity metrics
3. Adding more languages to perform cross language analysis for completing RQ4
4. Build UI to recommend the most referenced answer for a given stackoverflow question.

Questions??

References

- https://cmustrudel.github.io/papers/socialcom13.pdf
- https://empirical-software.engineering/assets/pdf/emse18-snippets.pdf



indera / bcrypt_test.py

Last active 5 years ago • Report abuse

bcrypt_test.py

```
"""
@see

http://www.jeffknupp.com/blog/2014/01/29/productionizing-a-flask-application/
http://stackoverflow.com/questions/549/the-definitive-guide-to-form-based-website-authentication
https://www.owasp.org/index.php/Password_Storage_Cheat_Sheet
http://security.stackexchange.com/questions/17421/how-to-store-salt
http://security.stackexchange.com/questions/3272/password-hashing-add-salt-pepper-or-is-salt-enough
http://security.stackexchange.com/questions/16809/is-a-hmac-ed-password-is-more-secure-than-a-bcrypt-ed-or-scrypt-ed-password

Testing bcrypt time
pip install py-bcrypt python-bcrypt
pip freeze | grep bcrypt

py-bcrypt==0.4
python-bcrypt==0.3.1
"""
import bcrypt

def crypt(rounds):
    salt = bcrypt.gensalt(rounds)
```