# APIzation: Generating Reusable APIs from StackOverflow Code Snippets

[ASE’ 21] Valerio Terragni etc.

Lingyu Zhang

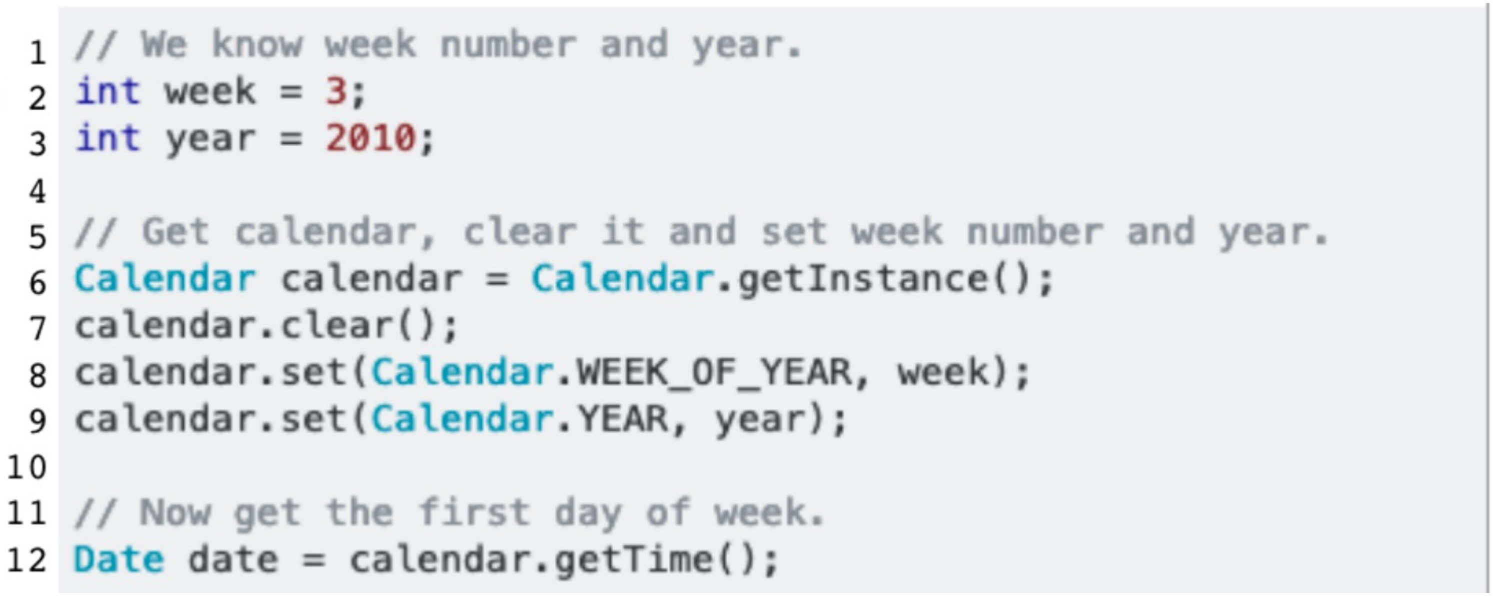
2023/03/21

# Outline

- Motivation
- Understanding Real Word APIzations
- APIZATOR: An Automated APIzation Tool for Java Code Snippets
- Evaluation

# Motivation

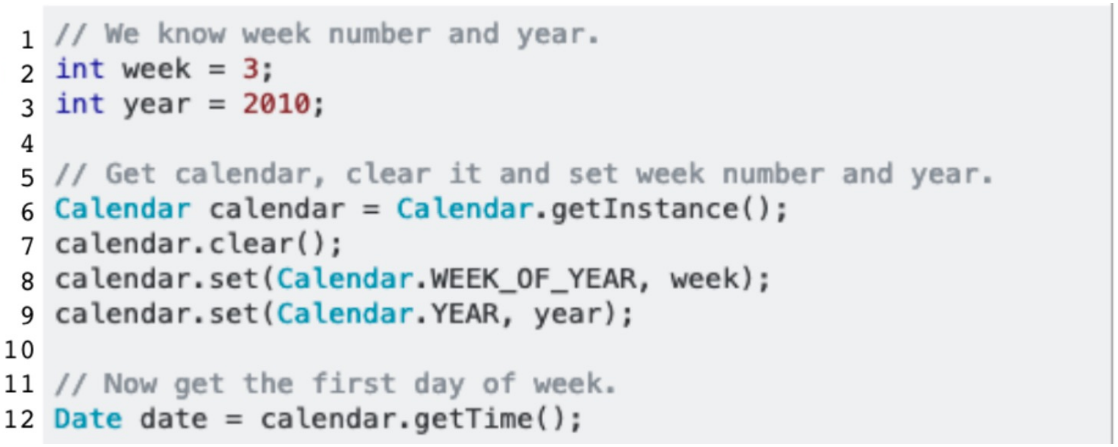
- Code snippet (CS) from StackOverflow is often incomplete for immediate reuse
    - Lack of type declarations
    - Lack of a well-formed method declaration



```
// We know week number and year.
int week = 3;
int year = 2010;

// Get calendar, clear it and set week number and year.
Calendar calendar = Calendar.getInstance();
calendar.clear();
calendar.set(Calendar.WEEK_OF_YEAR, week);
calendar.set(Calendar.YEAR, year);

// Now get the first day of week.
Date date = calendar.getTime();
```

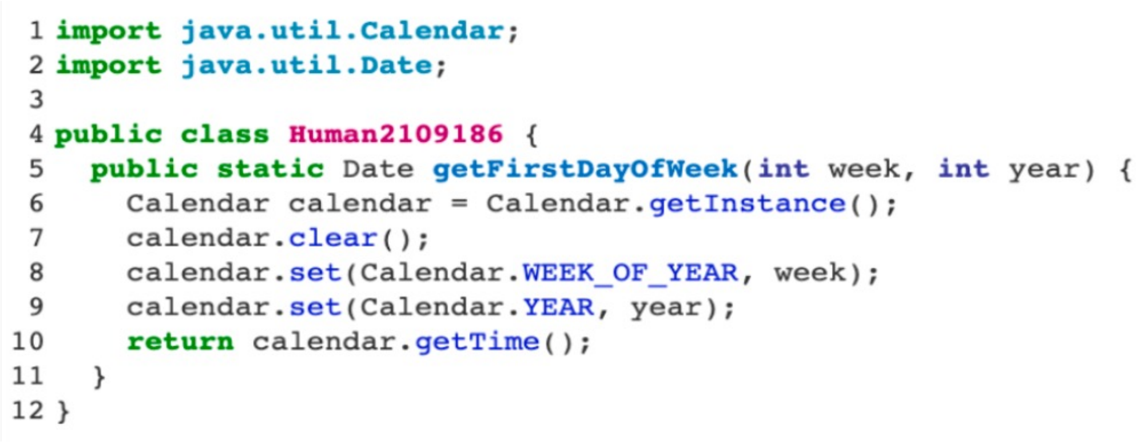
# APIzation

- Activity of creating an API for CSs without a well-formed method declaration



```java
// We know week number and year.
int week = 3;
int year = 2010;

// Get calendar, clear it and set week number and year.
Calendar calendar = Calendar.getInstance();
calendar.clear();
calendar.set(Calendar.WEEK_OF_YEAR, week);
calendar.set(Calendar.YEAR, year);

// Now get the first day of week.
Date date = calendar.getTime();
```

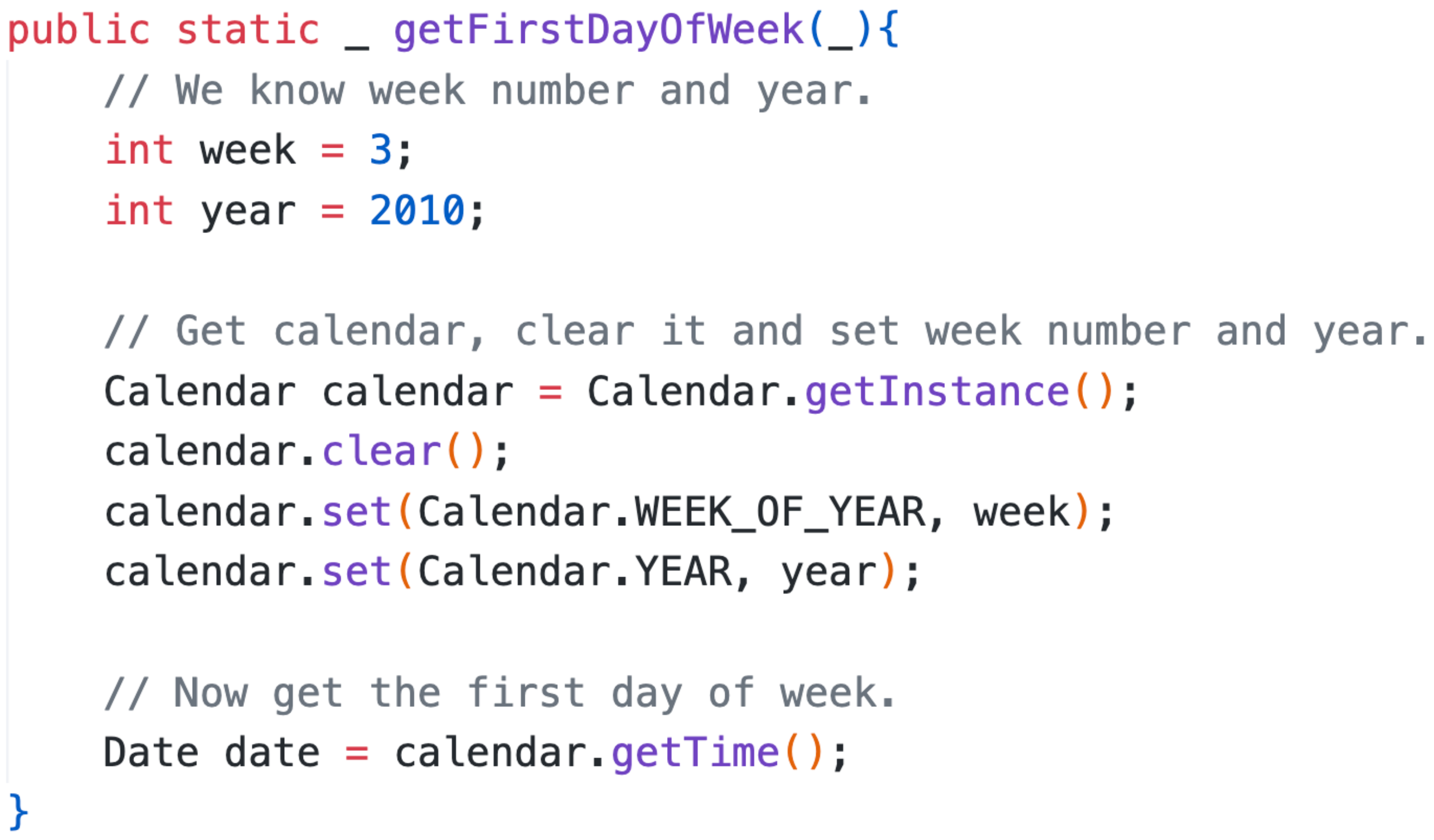
Automate?



```java
import java.util.Calendar;
import java.util.Date;

public class Human2109186 {
  public static Date getFirstDayOfWeek(int week, int year) {
    Calendar calendar = Calendar.getInstance();
    calendar.clear();
    calendar.set(Calendar.WEEK_OF_YEAR, week);
    calendar.set(Calendar.YEAR, year);
    return calendar.getTime();
  }
}
```

# Process of APIzation

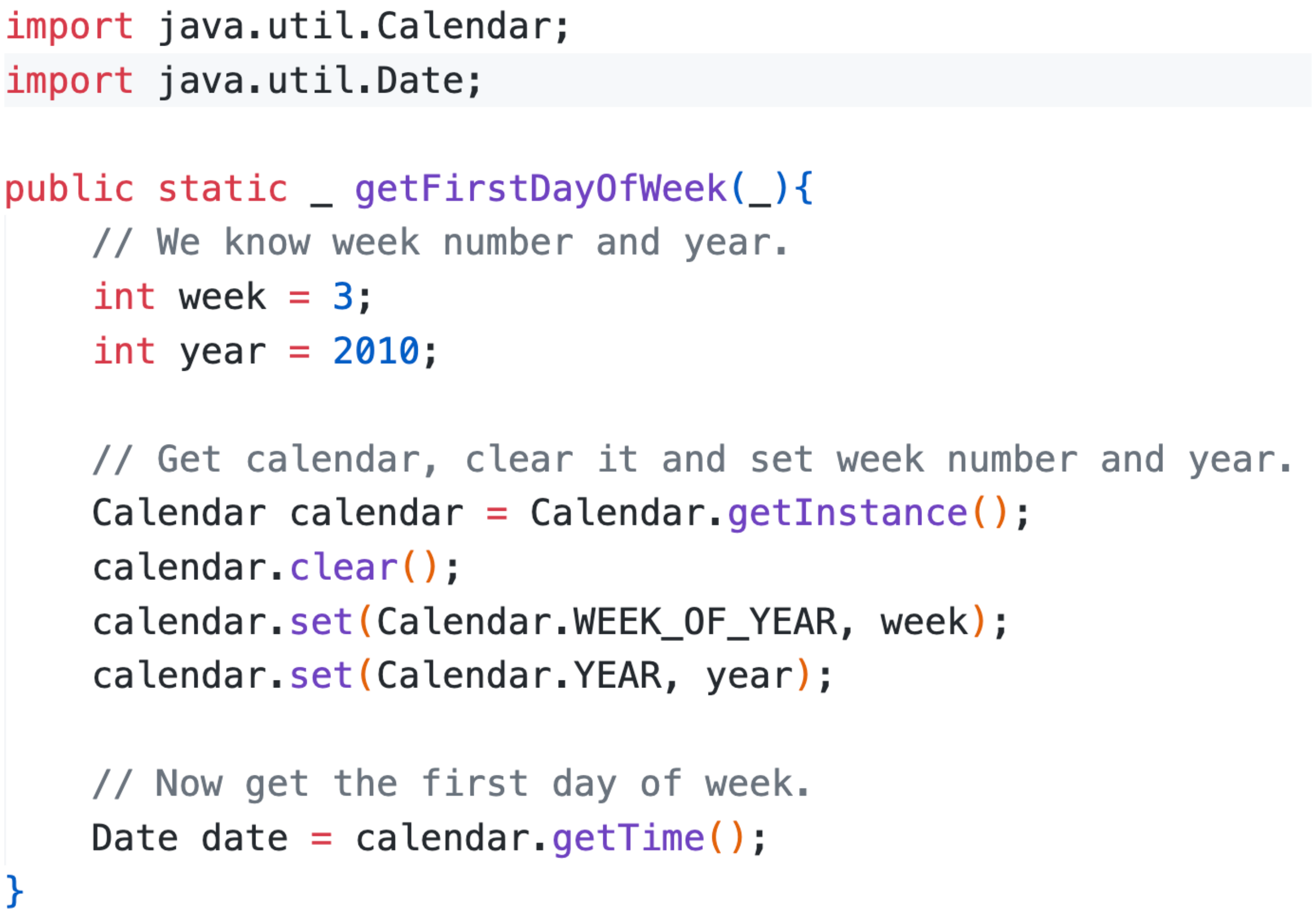
1. Choose modifiers and a method name



```
public static _ getFirstDayOfWeek(_){
    // We know week number and year.
    int week = 3;
    int year = 2010;

    // Get calendar, clear it and set week number and year.
    Calendar calendar = Calendar.getInstance();
    calendar.clear();
    calendar.set(Calendar.WEEK_OF_YEAR, week);
    calendar.set(Calendar.YEAR, year);

    // Now get the first day of week.
    Date date = calendar.getTime();
}
```

# Process of APIzation

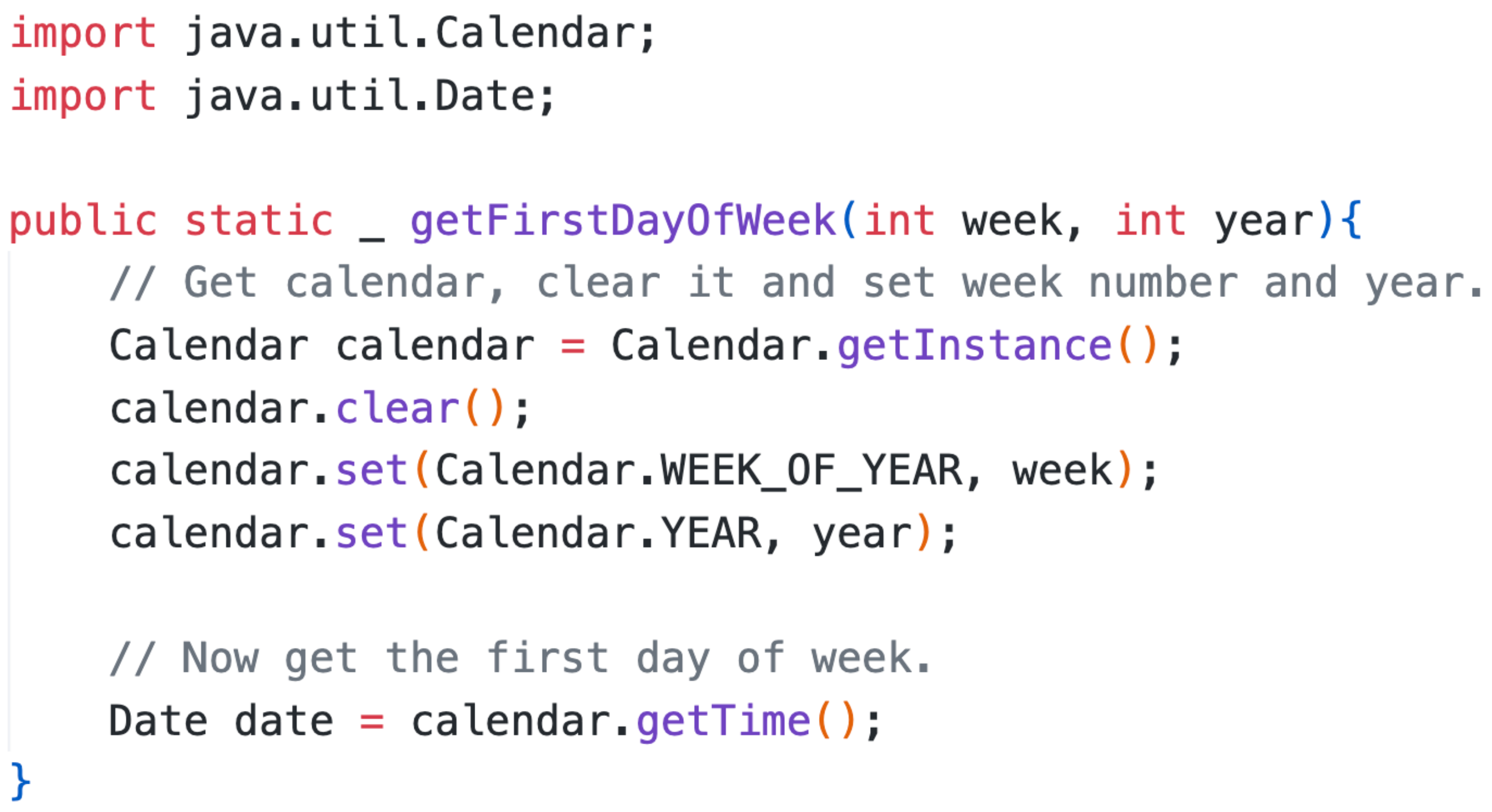
2. Recover missing declarations



```
import java.util.Calendar;
import java.util.Date;

public static _ getFirstDayOfWeek(_){
    // We know week number and year.
    int week = 3;
    int year = 2010;

    // Get calendar, clear it and set week number and year.
    Calendar calendar = Calendar.getInstance();
    calendar.clear();
    calendar.set(Calendar.WEEK_OF_YEAR, week);
    calendar.set(Calendar.YEAR, year);

    // Now get the first day of week.
    Date date = calendar.getTime();
}
```

# Process of APIzation

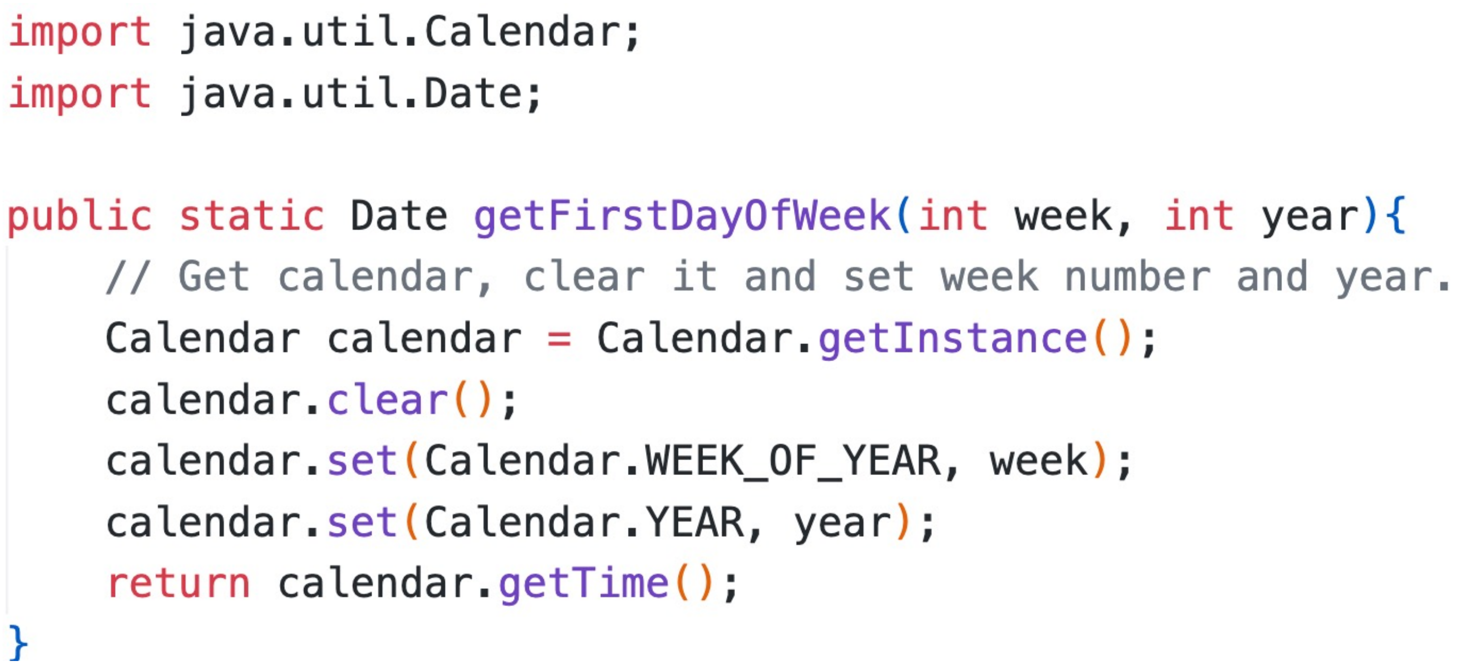
3. Extract intended input parameters



```java
import java.util.Calendar;
import java.util.Date;

public static _ getFirstDayOfWeek(int week, int year){
    // Get calendar, clear it and set week number and year.
    Calendar calendar = Calendar.getInstance();
    calendar.clear();
    calendar.set(Calendar.WEEK_OF_YEAR, week);
    calendar.set(Calendar.YEAR, year);

    // Now get the first day of week.
    Date date = calendar.getTime();
}
```

# Process of APIzation

4. Extract output



```java
import java.util.Calendar;
import java.util.Date;

public static Date getFirstDayOfWeek(int week, int year){
    // Get calendar, clear it and set week number and year.
    Calendar calendar = Calendar.getInstance();
    calendar.clear();
    calendar.set(Calendar.WEEK_OF_YEAR, week);
    calendar.set(Calendar.YEAR, year);
    return calendar.getTime();
}
```

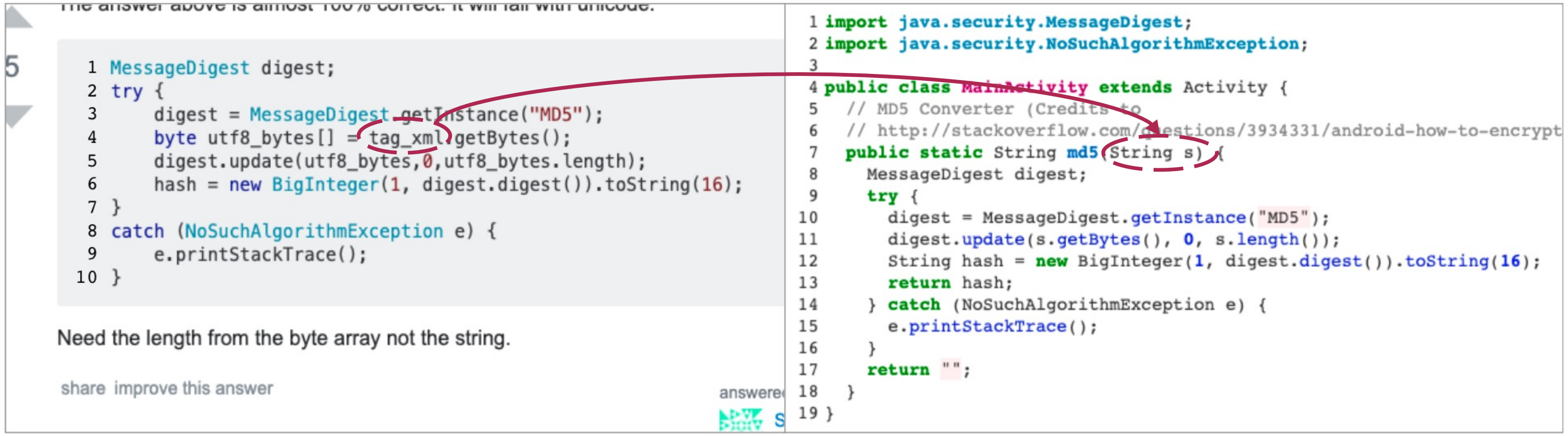
# Understanding Real Word APIzations

- Data collection approach
  - Explicit StackOverflow link
  - Type 3 code clone
  - Manual check

- 135 $< CS, API >$ pairs reference 509 variables

# Findings on method parameters

- PATT–notdecl



The answer above is almost 100% correct. It will fail with unicode.

5

```java
MessageDigest digest;
try {
    digest = MessageDigest.getInstance("MD5");
    byte utf8_bytes[] = tag_xml.getBytes();
    digest.update(utf8_bytes,0,utf8_bytes.length);
    hash = new BigInteger(1, digest.digest()).toString(16);
}
catch (NoSuchAlgorithmException e) {
    e.printStackTrace();
}
```

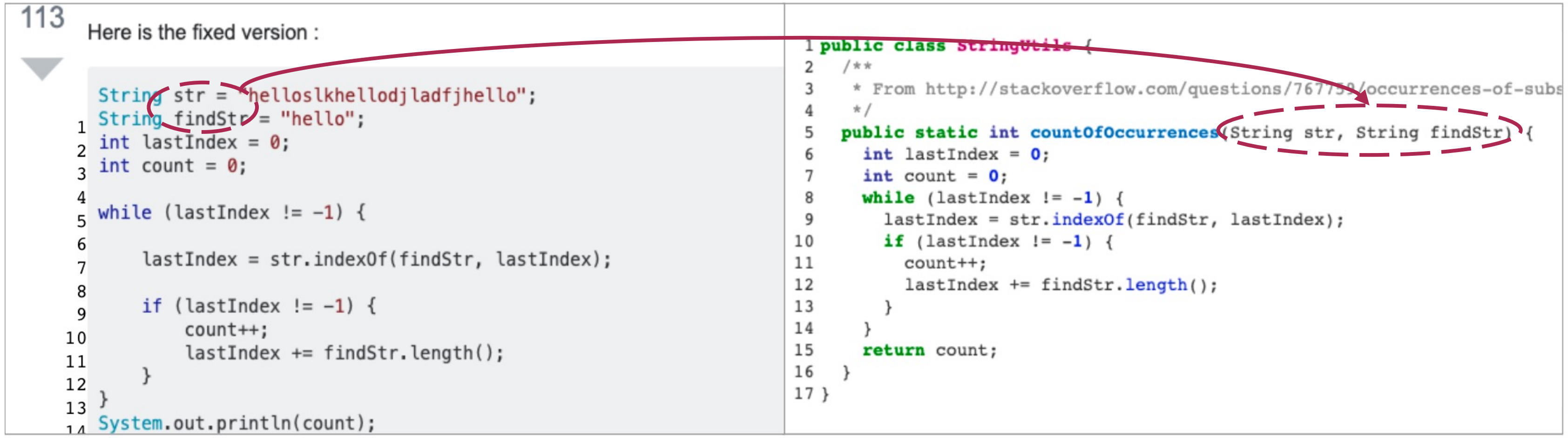
Need the length from the byte array not the string.

share improve this answer

answered

```java
import java.security.MessageDigest;
import java.security.NoSuchAlgorithmException;

public class MainActivity extends Activity {
  // MD5 Converter (Credits to
  // http://stackoverflow.com/questions/3934331/android-how-to-encrypt
  public static String md5(String s) {
    MessageDigest digest;
    try {
      digest = MessageDigest.getInstance("MD5");
      digest.update(s.getBytes(), 0, s.length());
      String hash = new BigInteger(1, digest.digest()).toString(16);
      return hash;
    } catch (NoSuchAlgorithmException e) {
      e.printStackTrace();
    }
    return "";
  }
}
```

- PATT–const



113

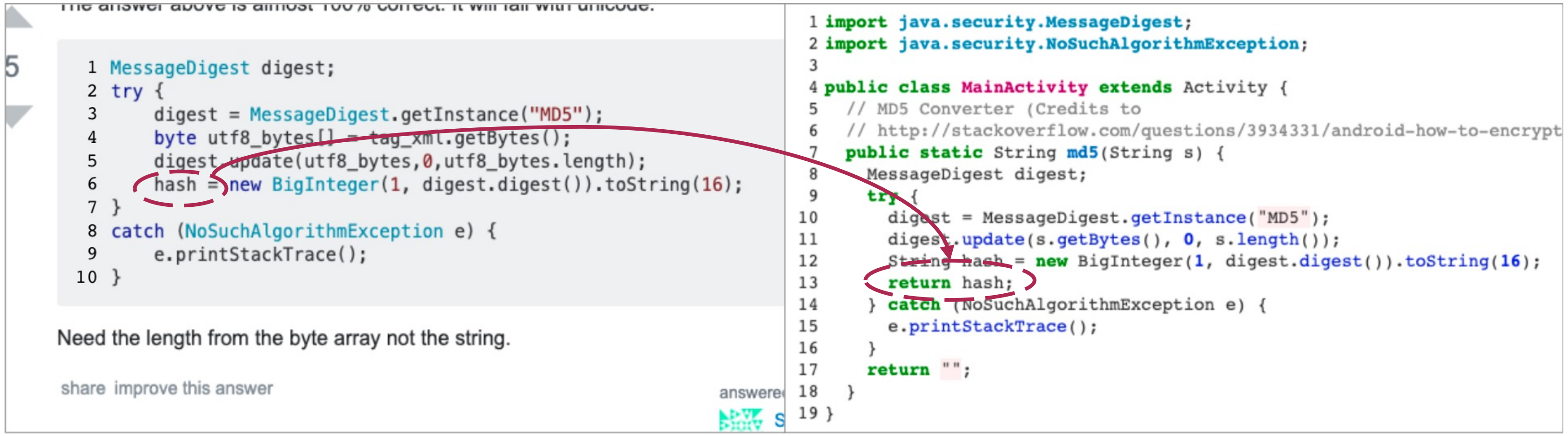
Here is the fixed version :

```java
String str = "helloslkhellodjladfjhello";
String findStr = "hello";
int lastIndex = 0;
int count = 0;

while (lastIndex != -1) {

    lastIndex = str.indexOf(findStr, lastIndex);

    if (lastIndex != -1) {
        count++;
        lastIndex += findStr.length();
    }
}
System.out.println(count);
```

```java
public class StringUtils {
  /**
   * From http://stackoverflow.com/questions/767759/occurrences-of-subs
   */
  public static int countOfOccurrences(String str, String findStr) {
    int lastIndex = 0;
    int count = 0;
    while (lastIndex != -1) {
      lastIndex = str.indexOf(findStr, lastIndex);
      if (lastIndex != -1) {
        count++;
        lastIndex += findStr.length();
      }
    }
    return count;
  }
}
```

# Findings on return statements

- PATT–latest



The answer above is almost 100% correct. It will fail with unicode.

5

```java
MessageDigest digest;
try {
    digest = MessageDigest.getInstance("MD5");
    byte utf8_bytes[] = tag_xml.getBytes();
    digest.update(utf8_bytes,0,utf8_bytes.length);
    hash = new BigInteger(1, digest.digest()).toString(16);
}
catch (NoSuchAlgorithmException e) {
    e.printStackTrace();
}
```

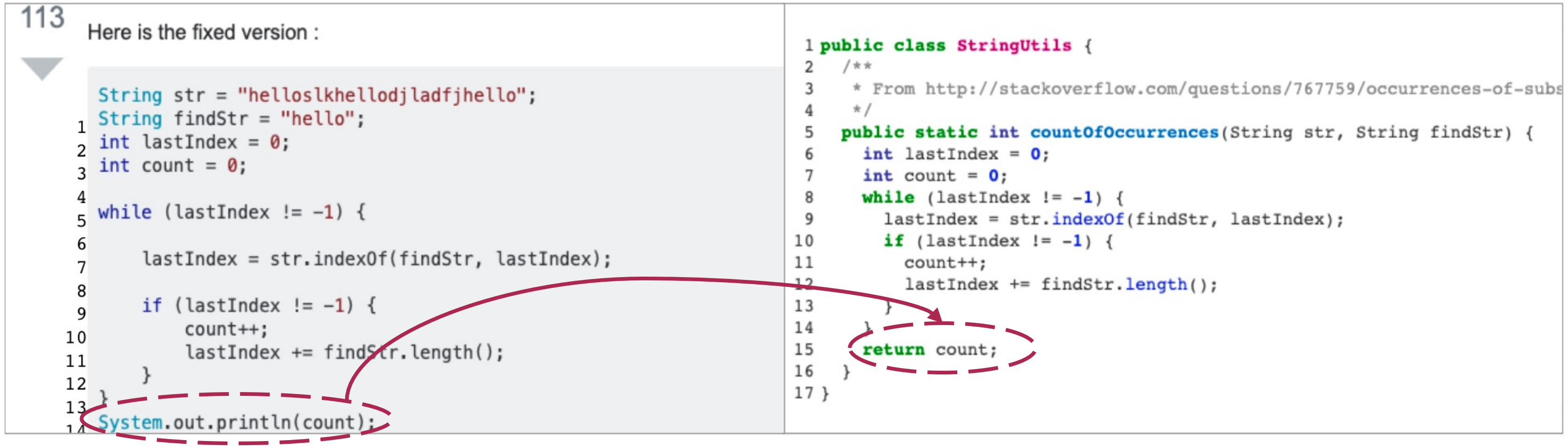
Need the length from the byte array not the string.

share improve this answer

answered

```java
import java.security.MessageDigest;
import java.security.NoSuchAlgorithmException;

public class MainActivity extends Activity {
  // MD5 Converter (Credits to
  // http://stackoverflow.com/questions/3934331/android-how-to-encrypt
  public static String md5(String s) {
    MessageDigest digest;
    try {
      digest = MessageDigest.getInstance("MD5");
      digest.update(s.getBytes(), 0, s.length());
      String hash = new BigInteger(1, digest.digest()).toString(16);
      return hash;
    } catch (NoSuchAlgorithmException e) {
      e.printStackTrace();
    }
    return "";
  }
}
```

- PATT–syso



113

Here is the fixed version :

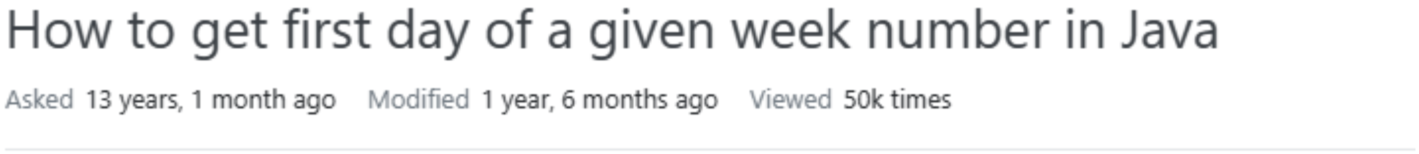
```java
String str = "helloslkhellodjladfjhello";
String findStr = "hello";
int lastIndex = 0;
int count = 0;

while (lastIndex != -1) {

    lastIndex = str.indexOf(findStr, lastIndex);

    if (lastIndex != -1) {
        count++;
        lastIndex += findStr.length();
    }
}
System.out.println(count);
```

```java
public class StringUtils {
  /**
   * From http://stackoverflow.com/questions/767759/occurrences-of-subs
   */
  public static int countOfOccurrences(String str, String findStr) {
    int lastIndex = 0;
    int count = 0;
    while (lastIndex != -1) {
      lastIndex = str.indexOf(findStr, lastIndex);
      if (lastIndex != -1) {
        count++;
        lastIndex += findStr.length();
      }
    }
    return count;
  }
}
```

# Choose Modifiers and A Method Name

- Modifiers: `public static`
  - API must be accessible by any other class
  - Avoiding instantiating objects for invoking the API

- Method name: Part-of-Speech (POS) Tagger
  - Generating from the title of the corresponding StackOverflow page
    - Assigning parts of speech to each word in the title
    - Combining the main “verb” and the corresponding “direct object”



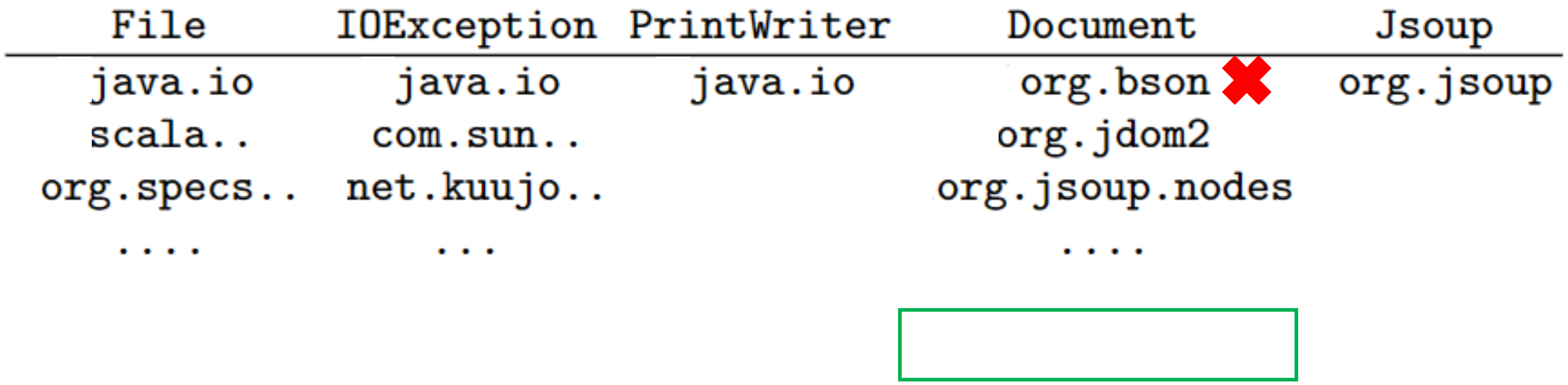
How to get first day of a given week number in Java

Asked 13 years, 1 month ago Modified 1 year, 6 months ago Viewed 50k times

→ getDay

# Recover Missing Declarations

- Type Declarations: CSNIPPEX*
  - A greedy algorithm based on the clustering hypothesis

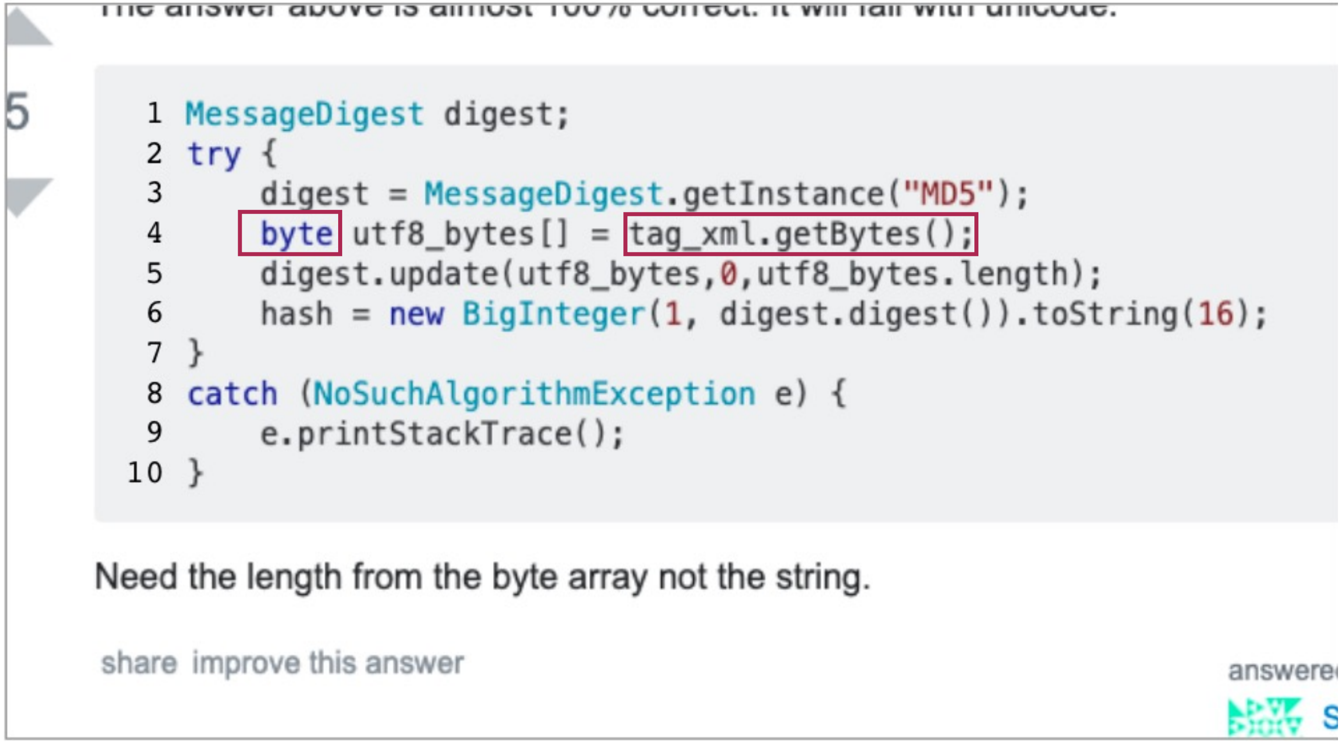


| File | IOException | PrintWriter | Document | Jsoup |
|---|---|---|---|---|
| java.io | java.io | java.io | org.bson ✖ | org.jsoup |
| scala.. | com.sun.. | | org.jdom2 | |
| org.specs.. | net.kuujo.. | | org.jsoup.nodes | |
| .... | ... | | .... | |

*: Valerio Terragni, Yepang Liu, Shing-Chi Cheung,CSNIPPEX: automated synthesis of compilable code snippets from Q&A sites. ISSTA 2016: 118-129

# Recover Missing Declarations

- Variable Declaration: Baker*
  - Identifying the most plausible type of $v$ by leveraging the usages of $v$ in the API



The answer above is almost 100% correct. It will fail with unicode.

5

```
MessageDigest digest;
try {
    digest = MessageDigest.getInstance("MD5");
    byte utf8_bytes[] = tag_xml.getBytes();
    digest.update(utf8_bytes,0,utf8_bytes.length);
    hash = new BigInteger(1, digest.digest()).toString(16);
}
catch (NoSuchAlgorithmException e) {
    e.printStackTrace();
}
```

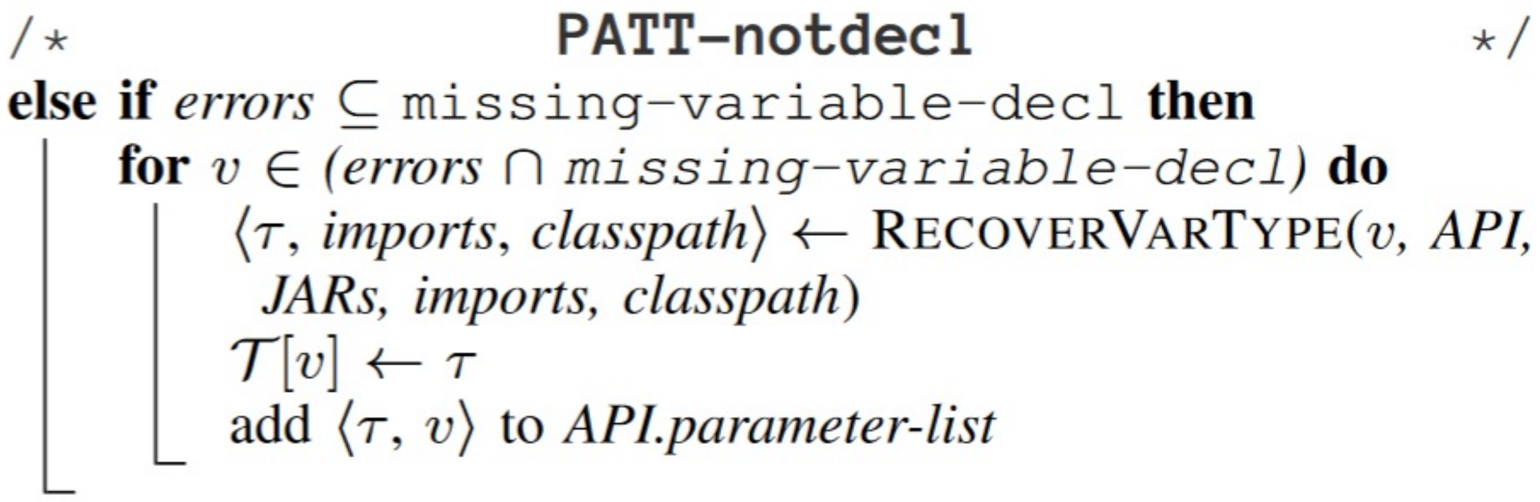
Need the length from the byte array not the string.

→ `String tax_xml`

*: Siddharth Subramanian, Laura Inozemtseva, Reid Holmes. Live API documentation. ICSE 2014: 643-652

# Extract intended input parameters

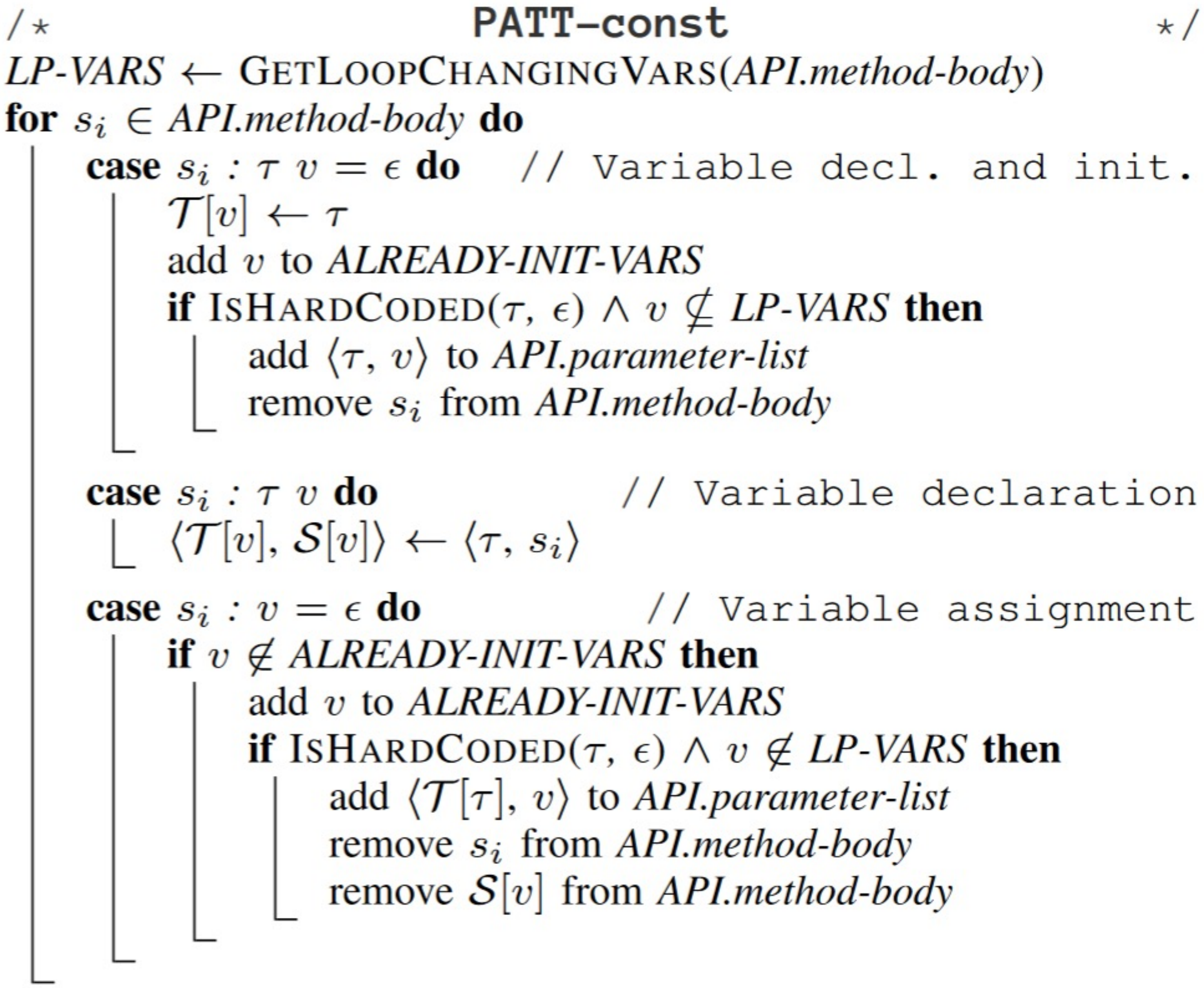
- `PATT-notdecl`
  - Undeclared variables are input parameters



```
/*                        PATT-notdecl                        */
else if $errors \subseteq$ missing-variable-decl then
    for $v \in (errors \cap missing\text{-}variable\text{-}decl)$ do
        $\langle \tau, imports, classpath \rangle \leftarrow$ RecoverVarType($v$, $API$,
          $JARs$, $imports$, $classpath$)
        $\mathcal{T}[v] \leftarrow \tau$
        add $\langle \tau, v \rangle$ to $API.parameter\text{-}list$
```

# Extract intended input parameters

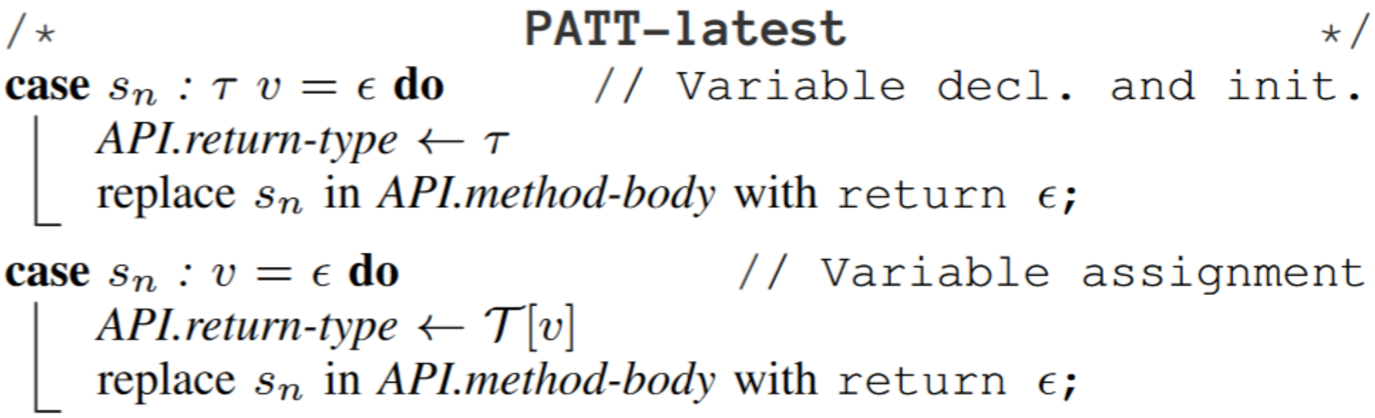
- `PATT−const`
  - Variables with const value are input parameters



```
/*                         PATT−const                          */
$LP\text{-}VARS \leftarrow$ GetLoopChangingVars($API.method\text{-}body$)
for $s_i \in API.method\text{-}body$ do
    case $s_i : \tau\ v = \epsilon$ do    // Variable decl. and init.
        $\mathcal{T}[v] \leftarrow \tau$
        add $v$ to $ALREADY\text{-}INIT\text{-}VARS$
        if IsHardCoded($\tau, \epsilon$) $\wedge\ v \not\subseteq LP\text{-}VARS$ then
            add $\langle \tau, v \rangle$ to $API.parameter\text{-}list$
            remove $s_i$ from $API.method\text{-}body$

    case $s_i : \tau\ v$ do                // Variable declaration
        $\langle \mathcal{T}[v], \mathcal{S}[v] \rangle \leftarrow \langle \tau, s_i \rangle$

    case $s_i : v = \epsilon$ do           // Variable assignment
        if $v \notin ALREADY\text{-}INIT\text{-}VARS$ then
            add $v$ to $ALREADY\text{-}INIT\text{-}VARS$
            if IsHardCoded($\tau, \epsilon$) $\wedge\ v \notin LP\text{-}VARS$ then
                add $\langle \mathcal{T}[\tau], v \rangle$ to $API.parameter\text{-}list$
                remove $s_i$ from $API.method\text{-}body$
                remove $\mathcal{S}[v]$ from $API.method\text{-}body$
```

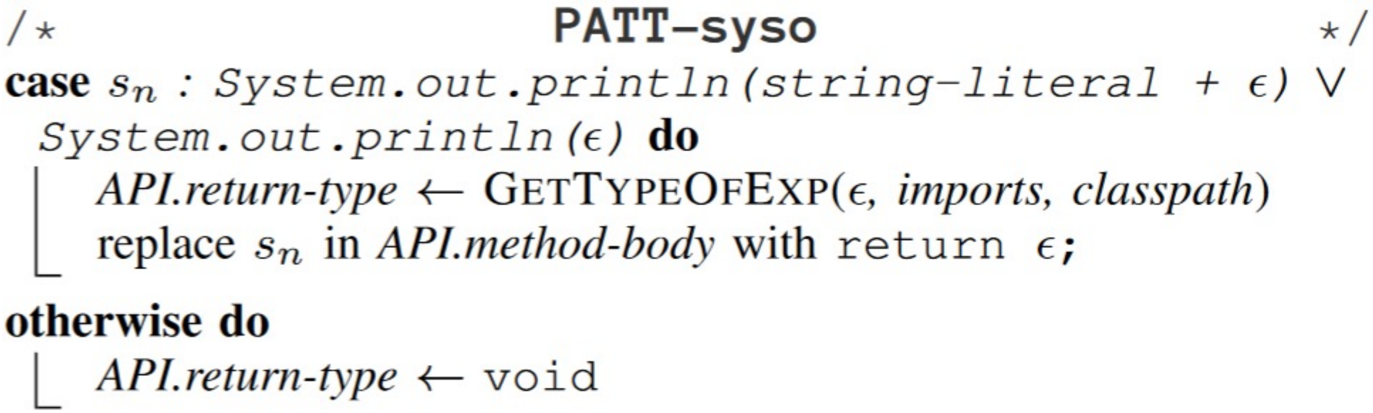
# Extract output

- PATT−latest



```
/*                    PATT−latest                    */
case $s_n : \tau\ v = \epsilon$ do        // Variable decl. and init.
    $API.return\text{-}type \leftarrow \tau$
    replace $s_n$ in $API.method\text{-}body$ with return $\epsilon$;
case $s_n : v = \epsilon$ do              // Variable assignment
    $API.return\text{-}type \leftarrow \mathcal{T}[v]$
    replace $s_n$ in $API.method\text{-}body$ with return $\epsilon$;
```

- PATT−syso



```
/*                    PATT−syso                    */
case $s_n$ : System.out.println(string−literal + $\epsilon$) $\vee$
  System.out.println($\epsilon$) do
    $API.return\text{-}type \leftarrow$ GETTYPEOFEXP($\epsilon$, imports, classpath)
    replace $s_n$ in $API.method\text{-}body$ with return $\epsilon$;
otherwise do
    $API.return\text{-}type \leftarrow$ void
```

# Evaluation

- RQ1: Identical APIs

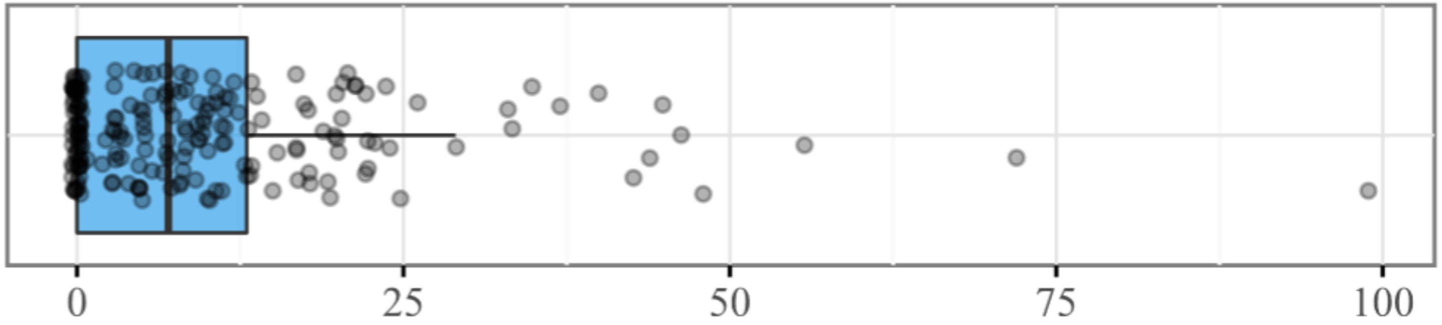
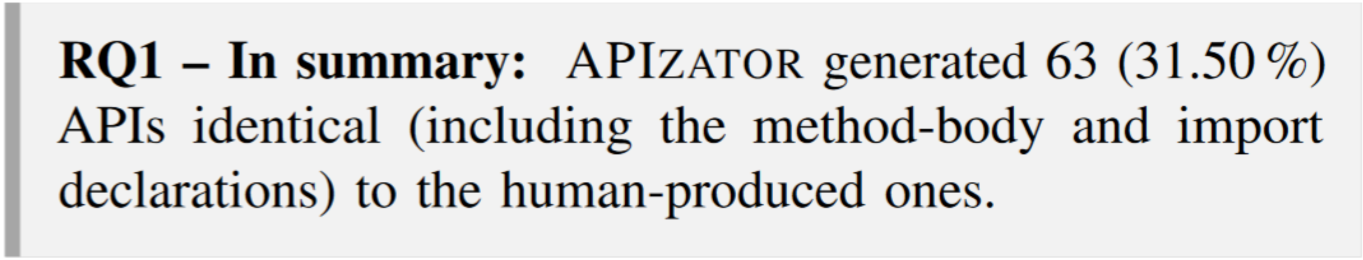


Fig. 3. Distribution of the number of AST differences.



**RQ1 – In summary:** APIZATOR generated 63 (31.50 %) APIs identical (including the method-body and import declarations) to the human-produced ones.

# Evaluation

- RQ2: Method Parameters

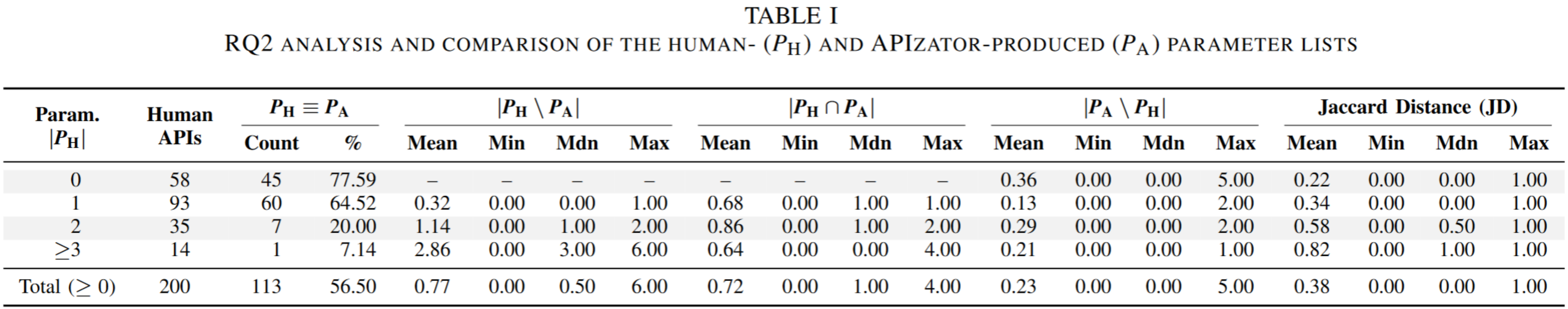
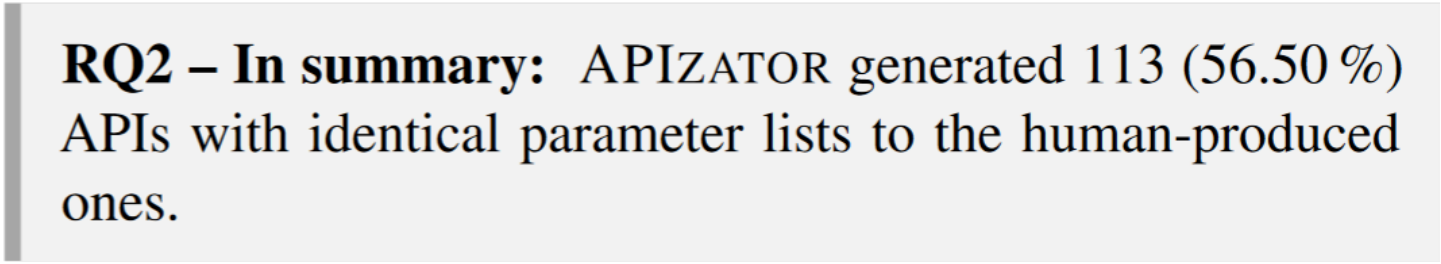


TABLE I

RQ2 ANALYSIS AND COMPARISON OF THE HUMAN- ($P_{\mathrm{H}}$) AND APIZATOR-PRODUCED ($P_{\mathrm{A}}$) PARAMETER LISTS

| Param. $\lvert P_{\mathrm{H}} \rvert$ | Human APIs | $P_{\mathrm{H}} \equiv P_{\mathrm{A}}$ | | $\lvert P_{\mathrm{H}} \setminus P_{\mathrm{A}} \rvert$ | | | | $\lvert P_{\mathrm{H}} \cap P_{\mathrm{A}} \rvert$ | | | | $\lvert P_{\mathrm{A}} \setminus P_{\mathrm{H}} \rvert$ | | | | Jaccard Distance (JD) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Count | % | Mean | Min | Mdn | Max | Mean | Min | Mdn | Max | Mean | Min | Mdn | Max | Mean | Min | Mdn | Max |
| 0 | 58 | 45 | 77.59 | – | – | – | – | – | – | – | – | 0.36 | 0.00 | 0.00 | 5.00 | 0.22 | 0.00 | 0.00 | 1.00 |
| 1 | 93 | 60 | 64.52 | 0.32 | 0.00 | 0.00 | 1.00 | 0.68 | 0.00 | 1.00 | 1.00 | 0.13 | 0.00 | 0.00 | 2.00 | 0.34 | 0.00 | 0.00 | 1.00 |
| 2 | 35 | 7 | 20.00 | 1.14 | 0.00 | 1.00 | 2.00 | 0.86 | 0.00 | 1.00 | 2.00 | 0.29 | 0.00 | 0.00 | 2.00 | 0.58 | 0.00 | 0.50 | 1.00 |
| $\geq$3 | 14 | 1 | 7.14 | 2.86 | 0.00 | 3.00 | 6.00 | 0.64 | 0.00 | 0.00 | 4.00 | 0.21 | 0.00 | 0.00 | 1.00 | 0.82 | 0.00 | 1.00 | 1.00 |
| Total ($\geq$ 0) | 200 | 113 | 56.50 | 0.77 | 0.00 | 0.50 | 6.00 | 0.72 | 0.00 | 1.00 | 4.00 | 0.23 | 0.00 | 0.00 | 5.00 | 0.38 | 0.00 | 0.00 | 1.00 |



**RQ2 – In summary:** APIZATOR generated 113 (56.50 %) APIs with identical parameter lists to the human-produced ones.

# Evaluation

- RQ3: Return Statements

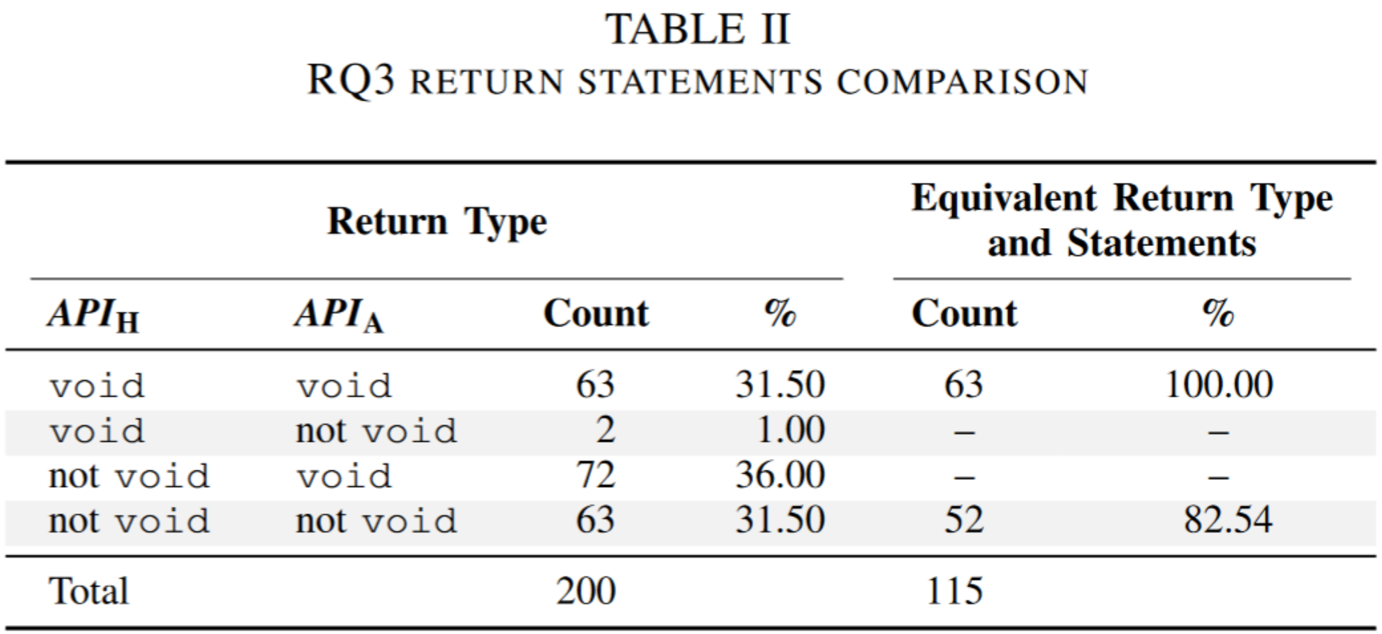
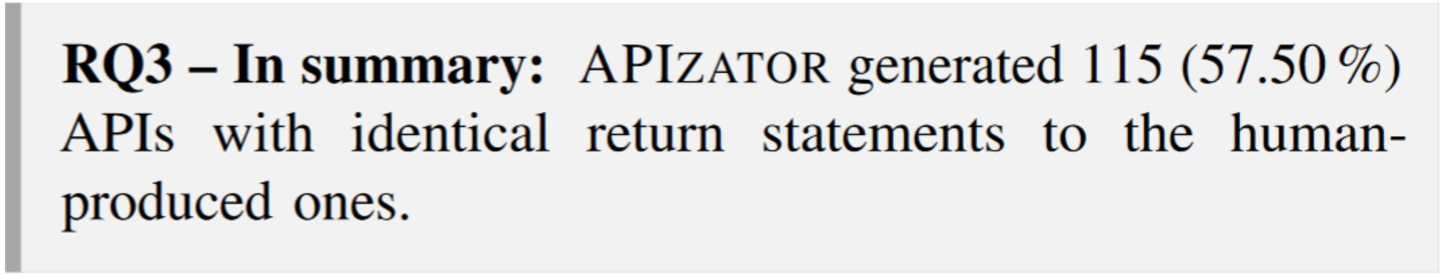


TABLE II
RQ3 RETURN STATEMENTS COMPARISON

| **Return Type** | | | | **Equivalent Return Type and Statements** | |
|---|---|---|---|---|---|
| ***API*$_{\mathrm{H}}$** | ***API*$_{\mathrm{A}}$** | **Count** | **%** | **Count** | **%** |
| `void` | `void` | 63 | 31.50 | 63 | 100.00 |
| `void` | not `void` | 2 | 1.00 | – | – |
| not `void` | `void` | 72 | 36.00 | – | – |
| not `void` | not `void` | 63 | 31.50 | 52 | 82.54 |
| Total | | 200 | | 115 | |



**RQ3 – In summary:** APIzator generated 115 (57.50 %) APIs with identical return statements to the human-produced ones.

# Thanks

Comments are welcome!