CS/SE 6356 Software Maintenance, Evolution & Re-engineering
Spring 2024
Assignment 5: Mining issue trackers and version control data

Assignment due date: **April 15**

1. Goal of the assignment

In this assignment, you will learn how to mine the issue tracker and the change history of a software system. This is a team assignment. As before, you are NOT to discuss with other teams the results, but you can discuss technical issues.

2. Software system

For this and the next assignment we will use **Apache PDFBox** (https://pdfbox.apache.org/) - an open-source Java library for working with PDF documents.

Repository at: https://github.com/apache/pdfbox

Clone the PDFBox repository on your local machine. Checkout the latest stable version. Instructions for building PDFBox from source are available at: https://pdfbox.apache.org/building.html, although you will not need to build and run the software for this assignment.

3. Analyzing issue tracker activity

Answer the next questions:

1. How many issues have been reported in the project?
2. How many of these issues are bugs, improvements, new features, tasks, wishes, sub-tasks, and tests? What are their percentages? How many issues of each of these types have been resolved or closed? What are their percentages? Fill in the following table:

Type of issue	**Total**		**Resolved/Closed**	
Bug	**#**	**%**	**#**	**%**
Improvement				
Task				
Sub-task				
Wish				
Test				
Total				

Note that Apache PDFBox uses Jira as issue tracker (https://issues.apache.org/jira/projects/PDFBOX).

4. Finding co-changed code files

Co-changed code files are those that developers frequently change at the same time. For example, in the next table, which shows the files changed in each changeset (i.e., commit), files A and H are co-changed files because they are frequently changed together.

Changeset	Changed files
1	A, H
2	G
…	…
10	H, A, F
11	F
…	…
20	J, A, B, H
…	…

For each project, you must find at least

1. 4 sets of 2 co-changed Java files
2. 3 sets of 3 co-changed Java files
3. 2 sets of 4 co-changed Java files
4. 1 sets of 5 co-changed Java files

These sets must occur no less than 3 times (i.e., the number of commits that contain the co-changed files must be greater than 2). Include the list of (minimum 3) commits (i.e., their hashes) in which the files appear. Fill in the following table

List of co-changed files	List of commits
(Example) 1. preflight/src/main/java/org/apache/pdfbox/preflight/annotation/AnnotationValidatorFactory.java 2. xmpbox/src/main/java/org/apache/xmpbox/schema/XMPSchemaFactory.java	1. 1f63fb1 2. 0d6c25b 3. 8768dd8 4. e36b4ea
…	

In this example from PDFBox, the files "…/annotation/AnnotationValidatorFactory.java" and "…/schema/XMPSchemaFactory.java" co-change and appear in the commits 1f63fb1, 0d6c25b, 8768dd8, and e36b4ea. **Note:** this example is not valid as an answer.

Select the set of 3 co-changed files that are most frequently changing together and select the set of 2 co-changed files that are most frequently changed together. Analyze the code and provide an explanation why these sets files change together so frequently. Feel free to measure the coupling between the classes in these files or even run a bad smell detector.

Note: you can use the Github mirror repository for this part, or the SVN repository.

5. Linking changes and issues

Link the data from issue trackers and versioning control systems to answer the following questions:

1. What are the commits addressing each issue in the issue tracker? Note that multiple commits can be made to address one issue.
2. How many source files are added to address each issue?
3. How many source files are modified to address each issue?
4. How many source files are deleted to address each issue?
5. How many source files on average are added/modified/deleted to address issues?

You can answer these questions by filling in the following table (feel free to move the table to a spreadsheet, if the table is too long):

Issue ID	Commits	Number of source code files		
		Added	Modified	Deleted
XXX-NNN	befd8eb, fe08f0d, ced8090	5	2	1
…				
Average	--	2.5	4.3	1.1

6. Hints

1. Apache PDFBox[1] uses Jira as issue tracker. Use the issue repositories to complete this assignment.
2. Clone the project repository (or pull the last changes if you already cloned the repository) and extract a customized change log according to your data needs.
3. Use simple text matching to link issues and commits, by finding the unique identifier of the issues in the commit descriptions.
4. Organize the information using tables and graphs in a spreadsheet or a CSV file.
5. Process the extracted log automatically via a script or program. Consider using R[2] to compute the statistics on the data. However, any other package or even Excell will do.

7. Deliverables

There are two deliverables for this assignment, which must be submitted on eLearning:

1. **Results report.** A PDF document with your answers and explanations, figures, and tables that support your answers. Include the team members in the document. You can use spreadsheets to report your results. If so, reference them in the document and submit them to eLearning. Describe what each team member did for the assignment, including the effort distribution.
2. **Code**. A ZIP file with the source code you used to process the data and compute the answers. If you use a spreadsheet, include all the macros and/or formulas in it.

8. Grading

Item	Points
Answers to section 3	30
Answers to section 4	30
Answers to section 5	30
Code to process the data	10
Total	**100**

You will receive points off if:

- Your document contains typos and writing errors.

[1] https://issues.apache.org/jira/projects/PDFBOX
[2] https://www.r-project.org/