MavenDependencyAnalyzer

MavenDependencyAnalyzer analyzes Maven dependencies of a Git repository. The child commit's dependencies are compared with the parent commit's dependencies. Changes in the dependencies printed to the output files. The program should be given the following arguments:

```
INPUT_FILE_PATH DEPENDENCY_FILENAME GIT_LOCATION GIT_URL REPOSITORY_NAME FILTER
```

Further explanation of the commands with examples

INPUT_FILE_PATH:

- This is the path to the input file. The input file is created by running the following command inside a Git repository: `git rev-list --parents MAIN_BRANCH_NAME > input.txt`. A few sample input files can be found in the folder `input_files`.

DEPENDENCY_FILENAME:

- This is the filename of the files that contain the Maven dependencies. Dependencies of Apache Cassandra are stored in a file called `build.xml`. Dependencies of Apache Hadoop and Tajo are stored in files called `pom.xml`.

GIT_LOCATION:

- This is the location of the command-line version of Git on the machine where the program is executed on. For a typical Windows machine this would be for example: `C:/ProgramFiles/Git/cmd/git.exe`

GIT_URL:

- This is the URL of the Git repository. This program uses this URL for cloning the repository. The URL of Apache Cassandra is for example: `https://github.com/apache/cassandra.git`

REPOSITORY_NAME:

- This is the name of the repository. For the three Apache project (Cassandra, Hadoop and Tajo) these names are: `cassandra`, `hadoop` and `tajo`

FILTER:

- This filter is used to skip dependencies that have a certain groupId. Providing the filter `org.apache.cassandra` would skip dependencies with `org.apache.cassandra` as the groupId. The empty string can be used when no dependencies have to be skipped.

Example sets of arguments

The following arguments are used to analyze Apache Cassandra:

```

```

```
input/commitsCassandra.txt build.xml "C:/Program Files/Git/cmd/git.exe"
https://github.com/apache/cassandra.git cassandra "org.apache.cassandra"
```

The following arguments are used to analyze Apache Hadoop:

```
input/commitsHadoop.txt pom.xml "C:/Program Files/Git/cmd/git.exe"
https://github.com/apache/hadoop.git hadoop "org.apache.hadoop"
```

The following arguments are used to analyze Apache Tajo:

```
input/commitsTajo.txt pom.xml "C:/Program Files/Git/cmd/git.exe"
https://github.com/apache/tajo.git tajo "org.apache.tajo"
```

Output

The output of the program consists of two files, diff.txt and diffAmounts.csv.

diff.txt contains on each line the hash of the commit that was analyzed. If there are Maven dependency changes found for that commit, these changes are printed below the commit hash with an indentation. A sample output:

```
3db64445f90e6fdb5ef550fc37dd7e8cd1161561
0dc5bd51f5ff36434bf7b5244242977ecbb47e39
    updated: 1
        <dependency>

file=C:\Users\arjan\Documents\GitHub\bachelor\DependencyAnalyzer\..\tmp\cassandra5
\cassandra\build.xml
            artifactPomId=parent-pom
            artifactId=snakeyaml
            groupId=org.yaml
            version=1.23
        </dependency>
        <dependency>

file=C:\Users\arjan\Documents\GitHub\bachelor\DependencyAnalyzer\..\tmp\cassandra5
\cassandra\build.xml
            artifactPomId=parent-pom
            artifactId=snakeyaml
            groupId=org.yaml
            version=1.11
        </dependency>
8b25cd58bfa646db9e6c24c51896950da02945db
d4eba9faa1b57fed205813a639fe53bbbdc06ef1
```

diffAmounts.csv contains the number of changed Maven dependencies for each commit. The first column is the commit hash, the second column is the amount of added dependencies, the third column is the amount of deleted dependencies and the fourth column is the amount of updated dependencies. A sample output:

```
1371883db3d8bf7d7c54e0baaca89c6c2d2a5abe, 4, 0, 0,
efe830e1f7e3f2b4dfb6c401326a06f2518c66b3, 0, 0, 0,
8f5b7fec711316a87f2ab37429228d7065e17c3a, 0, 0, 0,
2facbc97ea215faef1735d9a3d5697162f61bc8c, 0, 0, 0,
9b32b8a4369049aec6e0848d21f524a40d2c93f1, 0, 0, 0,
efa25fc8d10bbfcefe14fc6f2a623b6a8b73b5cd, 0, 0, 0,
d421e82ee0ffd66d3f382bfbe0b69b7b275edce3, 0, 0, 0,
d42087a63309178b96909c012dd0073fe0b6ea11, 0, 0, 0,
ec9b7b9d376e9c98cedb3ad4eae90311923bc7bd, 0, 0, 0,
a70253124681d006c865441b194d76d8e3058d64, 0, 0, 0,
24013e5c5ae538442cab083f8644563ea149ed7b, 0, 0, 0,
a1285ac92ded45ab6e9f6c7c98917daf14a4a320, 0, 0, 0,
17d379ca1be7b7b4d490cda39dcc1b30f5cb3bd3, 2, 0, 2,
cf39d031279620eb5684ad384dff72f7325104cd, 29, 39, 20,
```

Performance

Below you can find performance samples, achieved on a laptop with an Intel i7-8750h:

- Apache Cassandra (26000 commits): 2 hours
- Apache Hadoop (25000 commits): 2 hours
- Apache Tajo (2200 commits): 10-15 minutes