GitIssueFinder

GitIssueFinder finds the issue keys belonging to the Git commits that are given in the input file. The program should be given the following arguments:

```
INPUT_FILE_PATH GIT_LOCATION GIT_DIRECTORY ISSUE_PREFIXES
```

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`.

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_DIRECTORY:

- This is the directory of the Git repostiory on the machine where to program is executed on. This directory is created by manually cloning the repository that needs to be analyzed (do not clone this repository too deep in the file structure of your system, as this could break the program). For Apache Cassandra on a Windows machine this could be for example: `C:\Users\arjan\Documents\GitHub\cassandra`

ISSUE_PREFIXES:

- This is the prefix of the issues tokens in the issue tracking system. Apache Cassandra issues have the prefix `CASSANDRA`. Some projects, including Apache Hadoop, use multiple issue prefixes. These prefixes can be separated by commas and can look like `HADOOP,YARN,HDFS,MAPREDUCE`

Example sets of arguments

The following arguments are used to analyze Apache Cassandra:

```
input/commitsCassandra.txt "C:/Program Files/Git/cmd/git.exe"
"C:\Users\arjan\Documents\GitHub\cassandra" CASSANDRA
```

The following arguments are used to analyze Apache Hadoop:

```
input/commitsHadoop.txt "C:/Program Files/Git/cmd/git.exe"
"C:\Users\arjan\Documents\GitHub\hadoop" HADOOP,YARN,HDFS,MAPREDUCE
```

The following arguments are used to analyze Apache Tajo:

```
input/commitsTajo.txt "C:/Program Files/Git/cmd/git.exe"
"C:\Users\arjan\Documents\GitHub\tajo" TAJO
```

Output

The output of the program is a txt file (output.txt). It contains lines with on each line hashes of the analyzed commits. When one or more issue tokens were found for a commits, the tokens will be printed below the commit hash with an indentation. Below is a sample output:

```
3282f5ecf187ecbb56b8d73ab9a9110c010898b0
0fd8f0a52fbd69c47d073373abfe7d2437bbd9ca
    CASSANDRA-16602
f6d19512c4d79f800371da1e54dfe01cae5d894e
    CASSANDRA-16588
4bfe68717d9a419ab6a0b3a681478b39117dee80
    CASSANDRA-16601
```

Performance

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

- Apache Cassandra (26000 commits): 20 minutes
- Apache Hadoop (25000 commits): 20 minutes
- Apache Tajo (2200 commits): 2 minutes