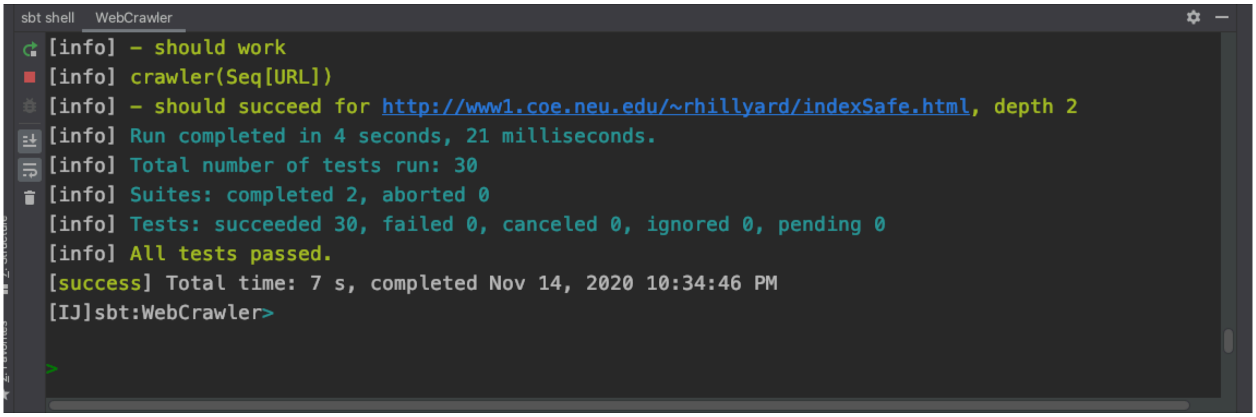
CSYE 7200 Big Data System Engineering using Scala

001306707
Ting-Kai Liu

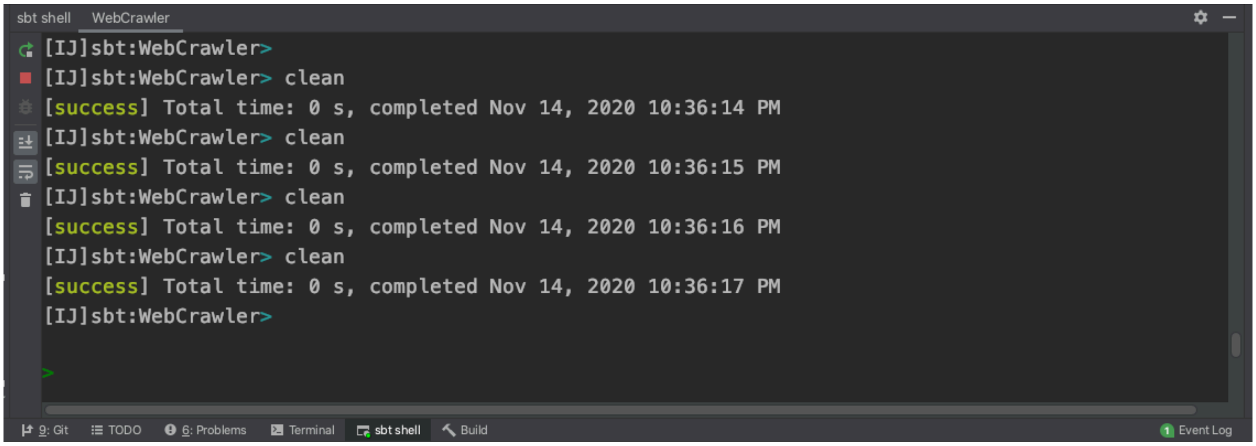
A. WebCrawler - Test result:



```
[info] - should work
[info] crawler(Seq[URL])
[info] - should succeed for http://www1.coe.neu.edu/~rhillyard/indexSafe.html, depth 2
[info] Run completed in 4 seconds, 21 milliseconds.
[info] Total number of tests run: 30
[info] Suites: completed 2, aborted 0
[info] Tests: succeeded 30, failed 0, canceled 0, ignored 0, pending 0
[info] All tests passed.
[success] Total time: 7 s, completed Nov 14, 2020 10:34:46 PM
[IJ]sbt:WebCrawler>

>
```

B. sbt clean up:



```
[IJ]sbt:WebCrawler>
[IJ]sbt:WebCrawler> clean
[success] Total time: 0 s, completed Nov 14, 2020 10:36:14 PM
[IJ]sbt:WebCrawler> clean
[success] Total time: 0 s, completed Nov 14, 2020 10:36:15 PM
[IJ]sbt:WebCrawler> clean
[success] Total time: 0 s, completed Nov 14, 2020 10:36:16 PM
[IJ]sbt:WebCrawler> clean
[success] Total time: 0 s, completed Nov 14, 2020 10:36:17 PM
[IJ]sbt:WebCrawler>

>
```

C. Suggestion to improve the web crawler:

1. Scraping the data only if the content has been updated.
2. Scraping only partial of the content for a given URL.
3. Store the scraping result.