Weikun Zhuang – SEC01 (NUID 001537998)

Big Data System Engineering with Scala
Fall 2022
Assignment No. #6

-List of Tasks Implemented

Implemented 3 TODOs in *WebCrawler.scala*, and *MonadOps.scala*

-Code

```scala
def asOption[X](xe: Either[Throwable, X]): Option[X] = xe.toOption
```

```scala
def wget(url: URL)(implicit ec: ExecutionContext): Future[Seq[URL]] = {
    // Hint: write as a for-comprehension, using the method createURL(Option[URL], String) to get the appropriate U
    // 16 points.
    def getURLs(ns: Node): Seq[Try[URL]] = {
        // for {n <- ns \\ "a"; nh <- n \ "@href"} yield createURL(Some(url), nh.text)
        // val test = ns \\ "a" map(_ \ "@href") // .filter(t => canParse(new URL(t toString)))
        // println(test)
        // for (t <-test) println(canParse(new URL(t toString)))
        for {n <- ns \\ "a"; nh <- n \ "@href"} yield validateURL(createURL(Some(url), nh.toString))
    } // TO BE IMPLEMENTED

    def getLinks(g: String): Try[Seq[URL]] = {
        val ny: Try[Node] = HTMLParser.parse(g) recoverWith { case f => Failure(new RuntimeException(s"parse proble
        for (n <- ny; uys = getURLs(n); us <- MonadOps.sequenceForgiveSubsequent(uys) { case _: WebCrawlerProtocolE
    }
    // Hint: write as a for-comprehension, using getURLContent (above) and getLinks above. You will also need Monad
    // 9 points.
    for (sf <- getURLContent(url); us <- MonadOps.asFuture(getLinks(sf))) yield us // TO BE IMPLEMENTED
}
```

-Unit tests

[Report added retroactively after the deadline. However, code unchanged]

```
[info] WebCrawlerSpec:
[info] getURLContent
[info] - should succeed for https://www1.coe.neu.edu/~rhillyard/indexSafe.html
[info] wget(URL)
[info] - should succeed for https://www1.coe.neu.edu/~rhillyard/indexSafe.html
[info] - should not succeed for https://www1.coe.neu.edu/junk
[info] - should not succeed for https://www1.coe.neu.edu/~rhillyard/indexSafe.html
[info] wget(Seq[URL])
[info] - should succeed for https://www1.coe.neu.edu/~rhillyard/indexSafe.html, https://www.google.com/
[info] wget(Seq[URL])
[info] - should succeed for https://www1.coe.neu.edu/~rhillyard/indexSafe.html
[info] filterAndFlatten
[info] - should work
[info] crawl(Seq[URL])
[info] - should succeed for https://www1.coe.neu.edu/~rhillyard/indexSafe.html, max 4
[info] Unstring
[info] - should ignore Unstring(0)
[info] - should Unstring(1) gobble one character
[info] Run completed in 21 seconds, 400 milliseconds.
[info] Total number of tests run: 45
[info] Suites: completed 3, aborted 0
[info] Tests: succeeded 45, failed 0, canceled 0, ignored 0, pending 0
[info] All tests passed.
[success] Total time: 22 s, completed Nov 20, 2022, 3:20:20 PM
```