Xinzhuo Liu – SEC01 (NUID 2197134)

Big Data System Engineering with Scala
Spring 2022
Assignment No. 7

Task

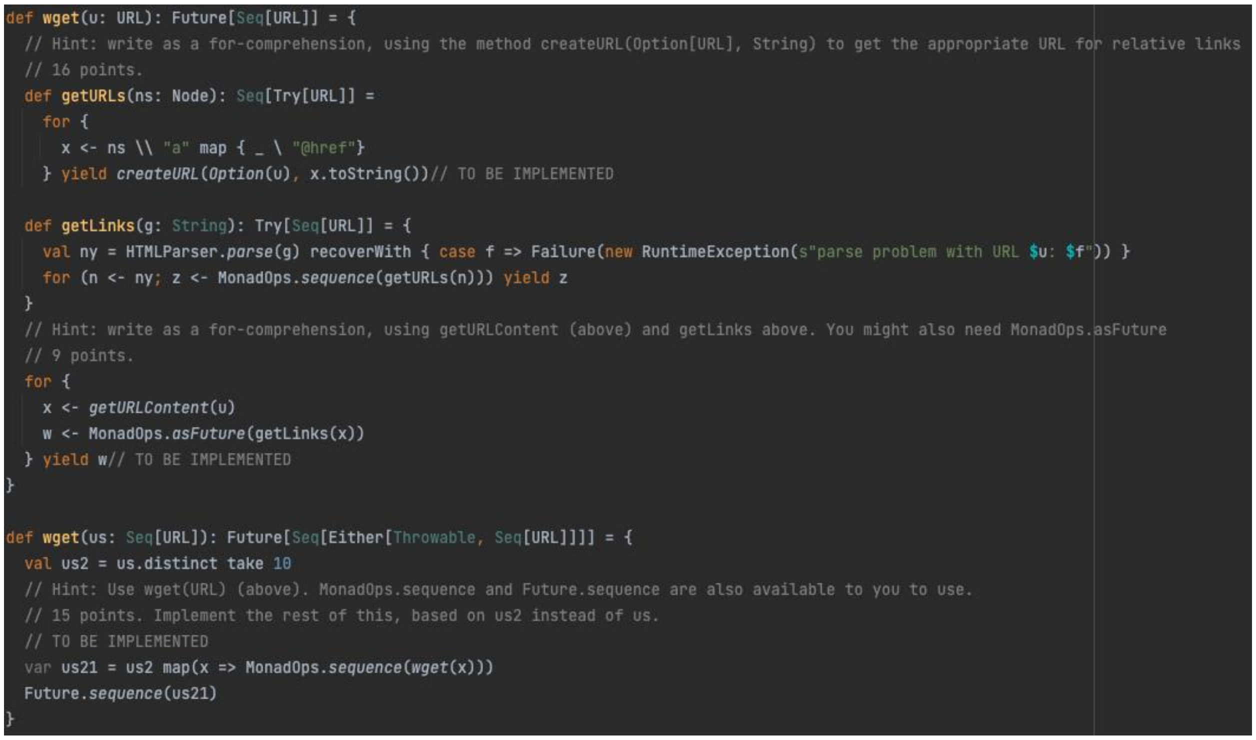
- There are five TO BE IMPLEMENTED to complete, with a total point value of 53. You may also earn up to 10 bonus points for suggestions on how to improve the web crawler (detailed code not required but you do need to explain in words what you would do). Three of these are in WebCrawler.scala. The other two are in MonadOps.scala as follows:

```scala
def sequence[X](xfs: Seq[Future[X]])(implicit executor: ExecutionContext): Seq[Future[Either[Throwable,
X]]] = ??? // TO BE IMPLEMENTED
def sequence[X](xe: Either[Throwable, X]): Option[X] = ??? // TO BE IMPLEMENTED
```

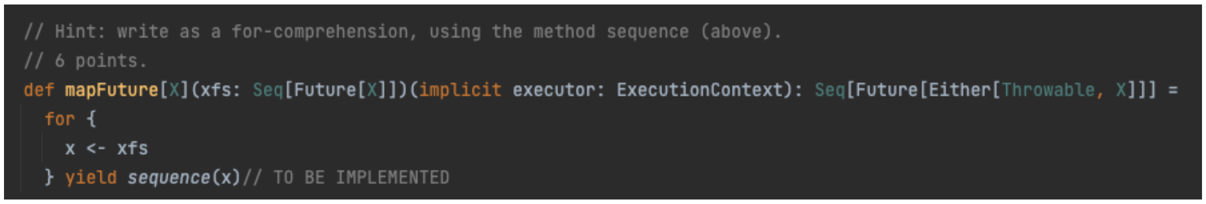
Solution/ Unit test (screenshot)

Suggestions on how to improve:

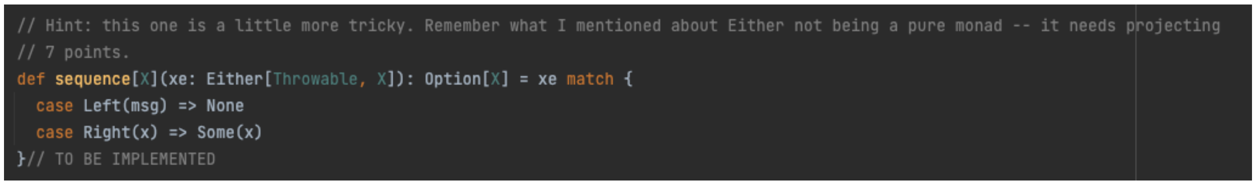
In WebCrawler.scala , I found many functions given a Try type. In my humble opinion, these functions can be replaced by Future type, I think this can save a little bit space and time for project to run more faster.



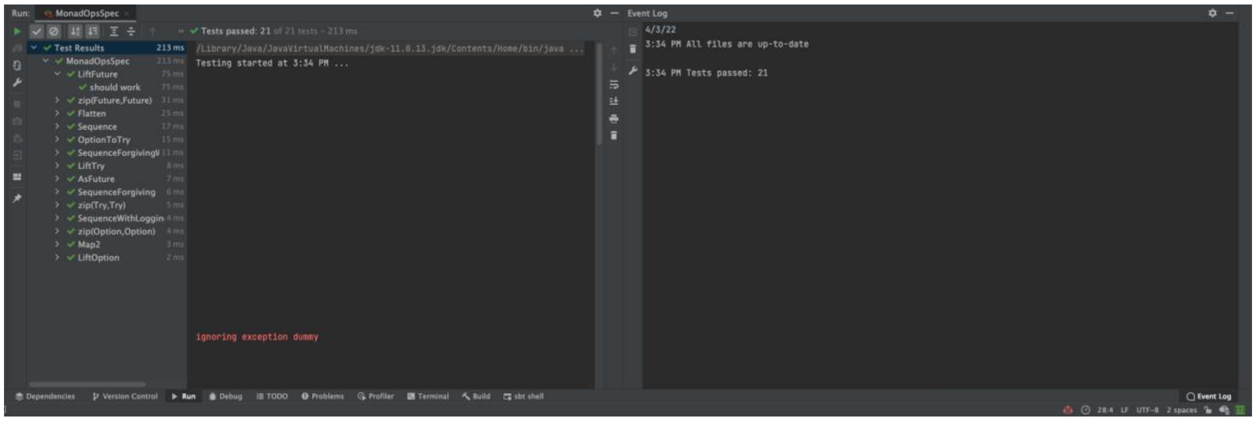
```scala
def wget(u: URL): Future[Seq[URL]] = {
  // Hint: write as a for-comprehension, using the method createURL(Option[URL], String) to get the appropriate URL for relative links
  // 16 points.
  def getURLs(ns: Node): Seq[Try[URL]] =
    for {
      x <- ns \\ "a" map { _ \ "@href"}
    } yield createURL(Option(u), x.toString())// TO BE IMPLEMENTED

  def getLinks(g: String): Try[Seq[URL]] = {
    val ny = HTMLParser.parse(g) recoverWith { case f => Failure(new RuntimeException(s"parse problem with URL $u: $f")) }
    for (n <- ny; z <- MonadOps.sequence(getURLs(n))) yield z
  }
  // Hint: write as a for-comprehension, using getURLContent (above) and getLinks above. You might also need MonadOps.asFuture
  // 9 points.
  for {
    x <- getURLContent(u)
    w <- MonadOps.asFuture(getLinks(x))
  } yield w// TO BE IMPLEMENTED
}

def wget(us: Seq[URL]): Future[Seq[Either[Throwable, Seq[URL]]]] = {
  val us2 = us.distinct take 10
  // Hint: Use wget(URL) (above). MonadOps.sequence and Future.sequence are also available to you to use.
  // 15 points. Implement the rest of this, based on us2 instead of us.
  // TO BE IMPLEMENTED
  var us21 = us2 map(x => MonadOps.sequence(wget(x)))
  Future.sequence(us21)
}
```



```scala
  // Hint: write as a for-comprehension, using the method sequence (above).
  // 6 points.
  def mapFuture[X](xfs: Seq[Future[X]])(implicit executor: ExecutionContext): Seq[Future[Either[Throwable, X]]] =
    for {
      x <- xfs
    } yield sequence(x)// TO BE IMPLEMENTED
```

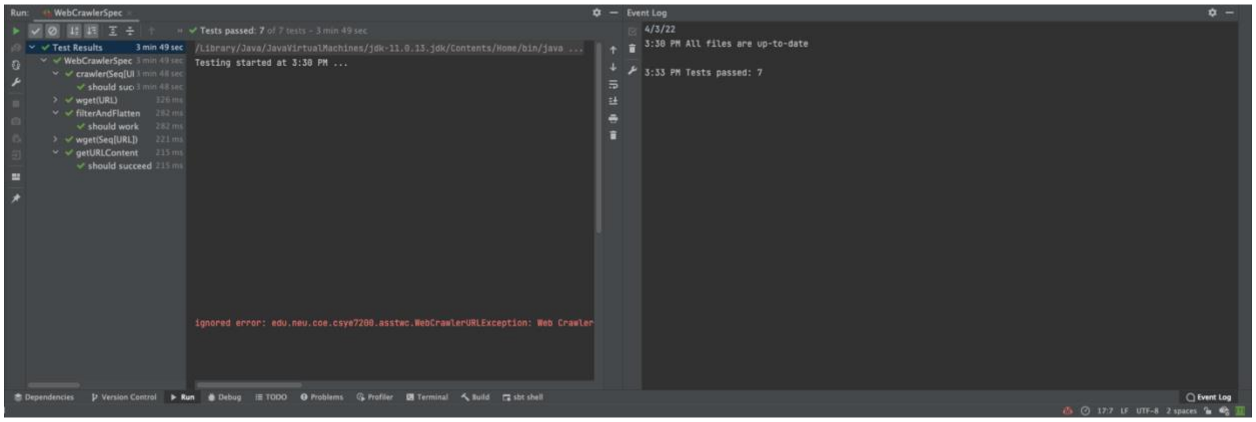


```scala
// Hint: this one is a little more tricky. Remember what I mentioned about Either not being a pure monad -- it needs projecting
// 7 points.
def sequence[X](xe: Either[Throwable, X]): Option[X] = xe match {
  case Left(msg) => None
  case Right(x) => Some(x)
}// TO BE IMPLEMENTED
```



```
Run: MonadOpsSpec
Tests passed: 21 of 21 tests - 213 ms

Test Results                 213 ms
  MonadOpsSpec               213 ms
    LiftFuture                75 ms
      should work             75 ms
    zip(Future,Future)        31 ms
    Flatten                   25 ms
    Sequence                  17 ms
    OptionToTry               15 ms
    SequenceForgivingW        11 ms
    LiftTry                    8 ms
    AsFuture                   7 ms
    SequenceForgiving          6 ms
    zip(Try,Try)               5 ms
    SequenceWithLoggin         4 ms
    zip(Option,Option)         4 ms
    Map2                       3 ms
    LiftOption                 2 ms

/Library/Java/JavaVirtualMachines/jdk-11.0.13.jdk/Contents/Home/bin/java ...
Testing started at 3:34 PM ...

ignoring exception dummy
```

```
Event Log
4/3/22
3:34 PM All files are up-to-date

3:34 PM Tests passed: 21
```



```
Run: WebCrawlerSpec
Tests passed: 7 of 7 tests - 3 min 49 sec

Test Results                 3 min 49 sec
  WebCrawlerSpec             3 min 49 sec
    crawler(Seq[UI           3 min 48 sec
      should suc             3 min 48 sec
    wget(URL)                326 ms
    filterAndFlatten         282 ms
      should work            282 ms
    wget(Seq[URL])           221 ms
    getURLContent            215 ms
      should succeed         215 ms

/Library/Java/JavaVirtualMachines/jdk-11.0.13.jdk/Contents/Home/bin/java ...
Testing started at 3:30 PM ...

ignored error: edu.neu.coe.csye7200.asstwc.WebCrawlerURLException: Web Crawler
```

```
Event Log
4/3/22
3:30 PM All files are up-to-date

3:33 PM Tests passed: 7
```

Project Source (github link to specific module relate to assignment)

Link: https://github.com/ZhongLBuL/Scala/tree/main/Assign7