Just Enough Scala for Spark

Just Enough Scala for Spark

Dean Wampler, Ph.D. (mailto:deanwampler@gmail.com)
@deanwampler (http://twitter.com/deanwampler)
Lightbend Fast Data Platform (http://lightbend.com/fast-data-platform)

Welcome. This notebook teaches you the core concepts of Scala (http://scala-lang.org) necessary to use Apache Spark's (http://spark.apache.org) Scala API effectively. Spark does a nice job exploiting the nicest features of Scala, while avoiding most of the more difficult and obscure features.

Introduction: Why Scala?

Spark lets you use Scala, Java, Python, R, and SQL to do your work. Scala and Java appeal to *data engineers*, who do the heavy lifting of building resilient and scalable infrastructures for *Big Data*. Python, R, and SQL appeal to *data scientists*, who build models for analyzing data, including machine learning, as well as explore data interactively, where SQL is very convenient.

These aren't hard boundaries. Many people do both roles. Many data engineers like Python and may use SQL and R. Many data scientists have decided to use Scala with Spark.

Briefly, some of the advantages of using Scala include the following:

- **Performance:** Since Spark is written in Scala, when you use the RDD (http://spark.apache.org/docs/latest/programming-guide.html#resilient-distributed-datasets-rdds) API, you get the best performance and the most complete API coverage when you use Scala. However, with DataFrames (http://spark.apache.org/docs/latest/sql-programming-guide.html), code written in all five languages performs about the same.
- **Debugging:** When runtime problems occur, understanding the exception stack trace and other debug information is easiest if you know Scala. Unfortunately, the abstractions provided by the different language APIs "leak" when problems occur.
- **Concise, Expressive Code:** Compared to Java, Scala code is much more concise and several features of Scala make your code even more concise. This elevates your productivity and makes it easier to imagine a design approach and then write it down without having to translate the idea to a less flexible API that reflects idiomatic language constraints. (You'll see this in action as we go.)
- **Type Safety:** Compared to Python and R, Scala code benefits from *static typing* with *type inference*. *Static typing* means that the Scala parser finds more errors in your expressions at compile time, when they don't match expected types, rather than discovering the problem later at run time. However, *type inference* means you don't have to add a lot of explicit type information to you code. In most cases, Scala will infer the correct types for you.

Why Not Scala?

Scala isn't perfect. There are two disadvantages compared to Python and R:

- **Libraries:** Python and R have a rich ecosystem of data analytics libraries. While the picture is improving for Scala, Python and R are still well ahead.
- **Advanced Language Features:** Mastering advanced language features gives you a lot of power to exploit, but if you don't understand those features, they can get in your way when you're just trying to get work done. Scala has some sophisticated constructs, especially in its *type system*. Fortunately, Spark mostly hides the advanced constructs.

For More on Scala

I can only scratch the surface of Scala here. We'll "sketch" many concepts without too much depth. You'll drink from a firehose, but learn enough to make use of the concepts. Eventually, when you're ready to deepen your understanding, consider these resources:

- Programming Scala, Second Edition (http://shop.oreilly.com/product/0636920033073.do): My comprehensive introduction to Scala.
- Scala Language Website (http://scala-lang.org/): Where to download Scala, find documentation (e.g., the Scaladocs (http://www.scala-lang.org/api/current/#package): Scala library documentation, like Javadocs (https://docs.oracle.com/javase/8/docs/api/)), and other information.
- Lightbend (http://www.lightbend.com/services/) training, consulting, and support for Scala, Big Data tools like Spark, and the Lightbend Reactive Platform (http://www.lightbend.com/products/lightbend-reactive-platform).

For now, I recommend that you open the Scaladocs for Scala and for Spark's Scala API. Clicking these two links will open them in new browser tabs:

- Scaladocs for Scala (http://www.scala-lang.org/api/2.11.8/#package).
- Scaladocs for Spark (http://spark.apache.org/docs/latest/api/scala/index.html#package).

Tips for using Scaladocs:

- Use the search bar in the upper-left-hand side to find a particular *type*. (For example, try "RDD" in the Spark Scaladocs.)
- To search for a particular *method*, click the character under the search box for the method name's first letter, then scroll to it.

About Notebooks

You're using the Spark Notebook (http://spark-notebook.io/) environment, a Scala-centric fork of iPython (https://ipython.org/) configured for Apache Spark (http://spark.apache.org).

Notebooks let you mix documentation, like this Markdown (https://daringfireball.net/projects/markdown/) "cell", with cells that contain code, graphs of results, etc. The metaphor is a physical notebook a scientist or student might use while working in a laboratory.

The menus and toolbar at the top provide options for evaluating a cell, adding and deleting cells, etc. You'll want to learn keyboard shortcuts if you use notebooks a lot.

Let's write our first Scala program:

```
println("Hello World!")
```

```
Took: 2 seconds 344 milliseconds, at 2017-3-12 11:23
```

Tips:

Invoke the *Help > Keyboard Shortcuts* menu item, then capture the page as an image (it's a modal dialog, unfortunately). Learn a few shortcuts each day.

For now, just know that you can click into any cell to move the focus. When you're in a cell, `shift+return` evaluates the cell (parses and renders the Markdown or runs the code), then moves to the next cell. Try it for a few cells. I'll wait...

Finally, there is a right-hand sidebar with useful information. If you want to view or hide the sidebar, use the *View > Toggle sidebar* menu item. In the sidebar you'll see a link *open SparkUI* to see Spark's own web UI. Use it to see more information about what your Spark jobs are doing.

If You Are Using Docker...

Unfortunately, this notebook will disappear, along with any edits you make to it, once the Docker image is shutdown. I recommend that you periodically use *File > Download as > Spark Notebook (.snb)* to save your work!

Okay. It's particularly nice that you can edit a cell you've already evaluated and rerun it. This is great when you're experimenting with code.

`SparkSession` and `SparkContext`

When you start this notebook, Spark Notebook creates a SparkSession (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.SparkSession), the entry point (http://spark.apache.org/docs/latest/programming-guide.html#initializing-spark) for Spark 2.X programs. Spark Notebook also creates a variable pointing to the older, Spark 1.X entry point, SparkContext (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.SparkContext). If you are using Spark 1.X at work, this is the entry point you would use for your programs.

`SparkSession` and `SparkContext` know how to connect to your cluster (or run locally in the same JVM, which is what we are doing today), how to configure properties, etc. They also run a Web UI that lets you monitor your running jobs.

The instance of `SparkSession` is called `sparkSession`. The instance of `SparkContext` is called `sc`.

The next two cells simply confirm that they exist.

```
sparkSession
```

```
org.apache.spark.sql.SparkSession@25c67ca9          Took: 1 second 356 milliseconds, at 2017-3-12 11:23
```

```
sc
```

```
org.apache.spark.SparkContext@2f90e645          Took: 1 second 211 milliseconds, at 2017-3-12 11:23
```

Actually, the `SparkContext` instance is really just a "field" in the `SparkSession` instance. Note that `sc` and `sparkSession.sparkContext` point to the same object:

```
sparkSession.sparkContext
```

```
org.apache.spark.SparkContext@2f90e645          Took: 1 second 217 milliseconds, at 2017-3-12 11:24
```

Here are a few other bits of information we can get from the `SparkContext`:

```
println("Spark version:      " + sc.version)
println("Spark master:       " + sc.master)
println("Running 'locally'?: " + sc.isLocal) // Again, "locally" means running everythi
```

```
          Took: 1 second 584 milliseconds, at 2017-3-12 11:24
```

Let's Download Some Data (and Start Learning Scala)

We're going to write real Spark programs and use them as vehicles for learning Scala and how to use it with Spark.

But first, we need to download some text files we'll use, which contain some of the plays of Shakespeare. The next few cells define some helper methods (functions) to do this and then perform the download. We'll start learning Scala concepts as we go.

> **Note:** "method" vs. "function"
>
> Scala follows a common object-oriented convention where the term *method* is used for a function that's attached to a class or instance. Scala also has *functions* that are not associated with a particular class or instance. Java 8 introduced this idea of functions into Java, too, where they are usually called *lambdas* (for historical reasons...)
>
> In our next code example, we'll define a few helper *methods* for printing information, but you won't see a class definition here. So, what class is associated with these methods? When you use Scala in a notebook, you're actually using the Scala interpreter, which wraps any expressions and definitions we write into a hidden, generated class. The interpreter has to do this in order to generate valid JVM byte code.

Unfortunately, it can be a bit confusing when to use a method vs. a function, reflecting Scala's hybrid nature as an object-oriented and a functional language. Fortunately, in many cases, we can use methods and functions interchangably, so we won't worry about the distinction too much from now on.

Again, we're defining *methods* now. We'll see what a real *function* looks like soon.

Okay, here are two convenience methods for printing either an error message or a simple "information" message. We'll explain all the syntax in a moment.

```
/*
 * "info" takes a single String argument, prints it on a line,
 * and returns it.
 */
def info(message: String): String = {
    println(message)

    // The last expression in the block, message, is the return value.
    // "return" keyword not required.
    // Do no additional formatting for the return string.
    message  // No additional formatting
}
```

Took: 1 second 296 milliseconds, at 2017-3-12 11:24

Note how the signature of the method is written in the result, which is what the Scala interpreter returns. The return type of the method is the `String` at the end.

```
/*
 * "error" takes a single String argument, prints a formatted error message,
 * and returns the message.
 */
def error(message: String): String = {

    // Print the string passed to "println" and add a linefeed ("ln"):
    // See the next cell for an explanation of how the string is constructed.
    val fullMessage = s"""
        |**********************************************************************
        |
        |  ERROR: $message
        |
        |**********************************************************************
        |""".stripMargin
    println(fullMessage)

    fullMessage
}
```

Took: 1 second 37 milliseconds, at 2017-3-12 11:24

Let's try them:

```
val infoString = info("All is well.")
```

Took: 926 milliseconds, at 2017-3-12 11:24

Why is the string shown twice? The first string is the output of `println` ("print line"). The second string is the value returned from `info` and assigned to the *immutable value* (`val` keyword) named `infoString`. Note that if you didn't know what type of object was returned by `info`, Spark Notebook is showing you the type here, using the Scala syntax, `name: Type`.

Check out the right-hand side bar, too (using the *View* > *Toggle Sidebar* menu item above). It shows the values we've defined so far and their types.

```
val errorString = error("Uh oh!")
```

Took: 1 second 16 milliseconds, at 2017-3-12 11:24

We see the same multiline string twice for the same reason. If you have defined a variable previously in the notebook, you can see its value (actually, the result of calling `toString` on the value), by putting it in a cell by itself.

```
infoString
```

```
All is well.
```

Took: 1 second 199 milliseconds, at 2017-3-12 11:24

Let's explain the details of method definitions. Here is `info` again, with the comments removed for clarity:

```
def info(message: String): String = {
    println(message)
    message
}
```

Okay, here are the gory details. Don't worry about remembering all of them now, but return to this list when you need a reminder. In order, a method definition has the following elements:

- The `def` keyword.
- The method's name (`info` in this case).
- The argument list in parentheses. If there are two or more arguments, the `name:Type` items are comma-separated. If there are no arguments, the empty parentheses can be omitted. This is common for "getter"-like methods that simply return a field in an instance or something new, like `toString`. Note that the method arguments must include the `:Type`; Scala can't infer these types!
- A colon followed by the type of the value returned by the method. This return type *can* be inferred by Scala in most cases, so it's optional. However, I recommend always putting in the return type for readibility by users (and to avoid occasionally, subtle mistakes in type inference...).
- An = (equals) sign that separates the method *signature* from the *body*.
- The body in braces `{ ... }`, although if the body consists of a single expression, the braces are optional.

- The last expression in the body is used as the return value. The `return` keyword is optional and rarely used.
- Semicolons (`;`) are inferred at the end of lines (in most cases) and rarely used.

Look again at the argument list for `info`. It is `(message: String)`, where `message` is the argument name and its type is `String`. This convention for *type annotations*, `name: Type`, is also used for the return type, `error(...): String`. Again, type annotations are required by Scala for method arguments. They are optional in most cases for the return type. Scala can infer the types of most expressions and variable declarations, too.

Scala uses the same comment conventions as Java, `// ...` for a single line, and `/* ... */` for a comment block.

Note: Expression vs. Statement

An *expression* has a value, while a *statement* does not. Hence, when we assign an expression to a variable, the value the expression returns is assigned to the variable. Most "constructs" in Scala are actually expressions, even `if` conditionals and `for` loops. In Java, those are statements.

Inside `error`, we used a combination *interpolated* and *triple-quoted* string with the syntax `s"""..."""`:

- **Triple-quoted string:** `"""..."""`. Useful for embedding newlines, like we did inside `error`. (We'll see another benefit later.)
- **String interpolation:** Invoked by putting `s` before the string, e.g., `s"..."` or `s"""..."""`. Lets us embed variable references and expressions, where the string conversion will be inserted automatically. For example:

```
s"""Use braces for expressions: ${sc.version}.
You can omit the braces when just using a variable: $sc
However, watch for ambiguities like ${sc}andextrastuff"""
```

```
Use braces for expressions: 2.0.2. You can omit the braces when just using a variable:
org.apache.spark.SparkContext@2f90e645 However, watch for ambiguities like
org.apache.spark.SparkContext@2f90e645andextrastuff
```
Took: 1 second 339 milliseconds, at 2017-3-12 11:24

Another feature we used in our triple-quoted string is the ability to strip the leading whitespace off each line. The `stripMargin` method removes all whitespace before and including the `|`. This lets you indent those lines for proper code formatting, but not have that whitespace remain in the string. In the following example, the resulting string has blank lines at the beginning and end. Note what happens with whitespace before `line2`:

```
s"""
    |line 1
    |  line 2
    |""".stripMargin
```

```
line 1 line 2
```
Took: 1 second 685 milliseconds, at 2017-3-12 11:24

Like Java, character "literals" are specified single quotes, '/', while strings use double quotes, "/". Note the return types for the next two cells:

```
'/'
```

```
/                                                   Took: 1 second 33 milliseconds, at 2017-3-12 11:24
```

```
"/"
```

```
/                                                   Took: 1 second 94 milliseconds, at 2017-3-12 11:24
```

Mutable Variables vs. Immutable Values

We've already seen how to declare an *immutable* value, using the `val` keyword. Let's explore this a bit more:

- `val immutableValue = ...`: Once initialized, we can't assign a *different* value to `immutableValue`.
- `var mutableVariable = ...`: We can assign new values to `mutableVariable` as often as we want.

It's *highly recommended* that you only use `vals` unless you have a good reason for using mutability, which is a very common source of bugs!!

> **Note:** `val` is a *shallow* declaration of immutability. Immutability doesn't affect the whole object graph. Specifically, we can't change what a `val` points to, but if the object itself is mutable, we can change it! For example, a Scala Array (http://www.scala-lang.org/api/current/#scala.Array) is just a thin wrapper around Java arrays, which are mutable. Observe the following:

```
val a = Array(1, 2, 3)
```

```
                                                    Took: 731 milliseconds, at 2017-3-12 11:24
```

```
a = Array(4, 5, 6)  // not allowed
```

```
a(1) = 20  // allowed!
a
```

3 entries total

Show: 10

_1	_2
0	1
1	20
2	3

Showing 4 of 4 records

Download the Files

Now let's define a method that works like the popularity *NIX curl (http://linux.die.net/man/1/curl) utility. It's a bit long and you don't need to understand all the details, but we'll use it to download data we need for the notebook.

Most of the types used here are from Java's library (JDK). Because Scala compiles to JVM byte code, you can use any Java library you want from Scala:

- java.net.URL (https://docs.oracle.com/javase/8/docs/api/java/net/URL.html): Handles URL formatting and connections.
- java.io.File (https://docs.oracle.com/javase/8/docs/api/java/io/File.html): Working with files and directories.
- java.io.BufferedInputStream (https://docs.oracle.com/javase/8/docs/api/java/io/BufferedInputStream.html): Buffered input from an underlying stream.
- java.io.BufferedOutputStream (https://docs.oracle.com/javase/8/docs/api/java/io/BufferedOutputStream.html): Buffered output to an underlying stream.
- java.io.FileOutputStream (https://docs.oracle.com/javase/8/docs/api/java/io/FileOutputStream.html): Output to a file, specifically.

As before, we'll use comments to explain a new Scala constructs as we go.

```
// Import this utility for working with URLs. Unlike Java the semicolon ';' is not requ
import java.net.URL

// Use {...} to provide a list of things to import, when you don't want to import every
// in a package and you don't want to write a separate line for each type.
import java.io.{File, BufferedInputStream, BufferedOutputStream, FileOutputStream}

/**
 * Download a file at a URL and write it to a target directory.
 * @return the java.io.File for the downloaded file.
 */
def curl(sourceURLString: String, targetDirectoryString: String): File = {

    // The path separator on your platform: "/" on Linux and MacOS, "\" on Windows.
    val pathSeparator = File.separator

    // Use the name of the remote file as the file name in the target directory.
    // We split on the URL path elements using the separator, which is ALWAYS "/"
    // on all platforms for URLs. This gives us an array of path elements; the
    // name will be the last one.
    val sourceFileName = sourceURLString.split("/").last
    val outFileName = targetDirectoryString + pathSeparator + sourceFileName

    // Set up a connection and buffered input stream for the source file.
    println(s"Downloading $sourceURLString to $outFileName")
    val sourceURL = new URL(sourceURLString)
    val connection = sourceURL.openConnection()
    val in = new BufferedInputStream(connection.getInputStream())

    // If here, the connection was successfully opened (i.e., no exceptions thrown).
    // Now create the target directory (nothing happens if it already exists).
    val targetDirectory = new File(targetDirectoryString)
    targetDirectory.mkdirs()

    // Setup the output file and a stream to write to it.
    val outFile = new File(outFileName)
    val out = new BufferedOutputStream(new FileOutputStream(outFile))

    // Create a buffer to hold the in-flight bytes.
    val hundredK = 100*1024
    val bytes = Array.fill[Byte](hundredK)(0)   // Create byte buffer, elements set to
                                                // Array elements are _mutable_.
    // Loop until we've read everything.
    var loops = 0                               // A counter for progress feedback.
    var count = in.read(bytes, 0, hundredK)     // Read up to "hundredK" bytes at a tim
    while (count != -1) {                       // Haven't hit the end of input yet?
        if (loops % 10 == 0) print(".")         // Print occasional feedback.
        loops += 1                              // increment the counter.
        out.write(bytes, 0, count)              // Write to the new file.
        count = in.read(bytes, 0, hundredK)     // Read the next chunk and loop...
    }
    println("\nFinished!")
    in.close()                                  // Clean up! Close file & stream handle
    out.flush()
    out.close()
    outFile                                     // Returned file (if we got this far)
}
```

Took: 1 second 299 milliseconds, at 2017-3-12 11:24

Okay, before we actually use `curl`, let's create the target directory. (This is also done in `curl`, but we're using the success or failure for other purposes here.)

```
// The target directory, which we'll now create, if necessary.
val shakespeare = new File("data/shakespeare")
```

Took: 798 milliseconds, at 2017-3-12 11:24

Scala's `if` construct is actually an expression (in Java they are *statements*). The `if` expression will return `true` or `false` and assign it to `success`, which we'll use in a moment.

```
val success = if (shakespeare.exists == false) {   // doesn't exist already?
    if (shakespeare.mkdirs() == false) {           // did the attempt fail??
        error(s"Failed to create directory path: $shakespeare")  // ignore returned str
        false
    } else {                                       // successful
        info(s"Created $shakespeare")
        true
    }
} else {
    info(s"$shakespeare already exists")
    true
}
println("success = " + success)
```

Took: 1 second 253 milliseconds, at 2017-3-12 11:24

If we successfully created the output directory (or it already existed), let's download a handful of files, each with one play of Shakespeare, from http://www.cs.usyd.edu.au/~matty/Shakespeare/ (http://www.cs.usyd.edu.au/~matty/Shakespeare/).

```
val pathSeparator = File.separator
val targetDirName = shakespeare.toString
val urlRoot = "http://www.cs.usyd.edu.au/~matty/Shakespeare/texts/comedies/"
val plays = Seq(
    "tamingoftheshrew", "comedyoferrors", "loveslabourslost", "midsummersnightsdream",
    "merrywivesofwindsor", "muchadoaboutnothing", "asyoulikeit", "twelfthnight")

if (success) {
    println(s"Downloading plays from $urlRoot.")
  val successes = for {
        play <- plays
        playFileName = targetDirName + pathSeparator + play
        playFile = new File(playFileName)
        if (playFile.exists == false)
        file = curl(urlRoot + play, targetDirName)
    } yield {
        info(s"Downloaded $play and wrote $file")
        s"$playFileName:\tSuccess!"
    }

    println("Finished!")

    successes.foreach(println)
}
```

Took: 1 second 703 milliseconds, at 2017-3-12 11:25

I'm using a so-called `for` *comprehension*. They are *expressions*, not *statements* like Java's `for` loops. They have the form:

```
for {
  play <- plays
  ...
} yield { block_of_final_expressions }
```

We iterate through a collection, `plays`, and assign each one to the `play` variable (actually an immutable value for each pass through the loop).

After assigning to `play`, subsequent steps in the `for` comprehension use it. First, a java.io.File (https://docs.oracle.com/javase/8/docs/api/java/io/File.html) instance, `playFile`, is created. Then, `playFile` is used to evaluate a conditional - does the file already exist (i.e., have we already downloaded this file)?

If the file already exists, the conditional returns `false`, which short-circuits the loop and goes to the next `play` in the list. If the file doesn't exit, the final expression uses `curl` to download it.

The `yield` keyword tells Scala that I want to construct a new collection, using the expression that follows to construct each element, an *interpolated* string.

Functions

Notice that `collection.foreach(println)` in the previous code cell. This is our first example of a *function* being used. Normally functions are passed as arguments to *methods*. This is an incredibly powerful technique; we write a method that does some "general" work, then pass it a function to control the details in a particular context.

In this case, `collection.foreach(...)` iterates through `collection` (so we don't have to write this boilerplate ourselves), then we pass a function to it that will be applied to each element in the collection. The `foreach` returns nothing (a special type `Unit` actually, which is like `void` in Java). So, we use it here to simply print the values, by passing `println` as the function.

> **Note:** Actually `println` is itself a *method*, but when we use it in a function context like this, Scala treats it like a *function*. This is why in many cases we can ignore the distinction between *methods* and *functions*.

Let's look at more examples of functions. We'll use `plays` instead of `successes`, because the latter will be empty if you run that code cell more than once, since the files will already be downloaded!

```
println("Pass println as the function to use for each element:")
plays.foreach(println)

println("\nUsing an anonymous function that calls println: `str => println(str)`")
println("(Note that the type of the argument `str` is inferred to be String.)")
plays.foreach(str => println(str))
```

Took: 1 second 187 milliseconds, at 2017-3-12 11:25

So *anonymous* functions (i.e., those that don't have a name) are written `argument_list => body`.

> **Note:** Recall that *method* arguments have to be declared with types. That's usually *not* required for *function* arguments, as shown here.

```
println("\nAdding the argument type explicitly. Note that the parentheses are required.
plays.foreach((str: String) => println(str))

println("\nFor longer functions, you can use {...} instead of (...).")
println("Why? Because it gives you the familiar multiline block syntax with {...}")
plays.foreach {
  (str: String) => println(str)
}

println("\nWhy do we need to name this argument? Scala lets us use _ as a placeholder."
plays.foreach(println(_))

println("\nThe _ placeholder can be used *once* for each argument in the list.")
println("As an assume, use `reduceLeft` to sum some integers.")
val integers = 0 to 10   // Return a "range" from 0 to 10, inclusive
integers.reduceLeft((i,j) => i+j)
integers.reduceLeft(_+_)
```

```
55                                    Took: 1 second 497 milliseconds, at 2017-3-12 11:25
```

Our First Spark Program

Whew! We've learned a lot of Scala already while doing typical data science chores (i.e., fetching data).

Now let's implement a real algorithm using Spark, *Inverted Index*.

Inverted Index - When You're Tired of Counting Words...

You'll want to use *Inverted Index* when you create your next "Google killer". It takes in a corpus of documents (e.g., web pages), tokenizes the words, and outputs for each word a list of the documents that contain it, along with the corresponding counts.

This is a slightly more interesting algorithm than *Word Count*, the classic "hello world" program everyone implements when they learn Spark.

The term *inverted* here means we start with the words as part of the input *values*, while the *keys* are the document identifiers, and we'll switch ("invert") to using the words as keys and the document identifiers as values.

Here's our first version, all at once. This is *one, long expression*. Note the periods `.` at the end of the subexpressions.

```
val iiFirstPass1 = sc.wholeTextFiles(shakespeare.toString).
    flatMap { location_contents_tuple2 =>
        val words = location_contents_tuple2._2.split("""\W+""")
        val fileName = location_contents_tuple2._1.split(pathSeparator).last
        words.map(word => ((word, fileName), 1))
    }.
    reduceByKey((count1, count2) => count1 + count2).
    map { word_file_count_tup3 =>
        (word_file_count_tup3._1._1, (word_file_count_tup3._1._2, word_file_count_tup3.
    }.
    groupByKey.
    sortByKey(ascending = true).
    mapValues { iterable =>
        val vect = iterable.toVector.sortBy { file_count_tup2 =>
            (-file_count_tup2._2, file_count_tup2._1)
        }
        vect.mkString(",")
    }
```

Took: 5 seconds 763 milliseconds, at 2017-3-12 11:26

```
iiFirstPass1.take(50).foreach(println)
```

Took: 1 second 592 milliseconds, at 2017-3-12 11:31

Now let's break it down into steps, assigning each step to a variable. This extra verbosity let's us see what Scala infers for the type returned by each expression, helping us learn.

This is one of the nice features of Scala. We don't have to put in the type information ourselves, most of the time, like we would have to do for Java code. Instead, we let the compiler give us feedback about what we just created. This is especially useful when you're learning a new API, like Spark's.

```
val fileContents = sc.wholeTextFiles(shakespeare.toString)
```

Took: 911 milliseconds, at 2017-3-12 11:32

The output is telling us that `fileContents` has the type RDD[(String,String)] (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.rdd.RDD), but `RDD` is a base class and the actual instance is a `MapPartitionsRDD`, which is a "private" implementation subclass of `RDD`.

A name followed by square brackets, `[...]`, means that `RDD[...]` requires one or more type parameters in the brackets. In this case, a single type parameter, which represents the type of the records held by the `RDD`.

The single type parameter is given by `(String,String)`, which is a convenient shorthand for Tuple2[String,String] (http://www.scala-lang.org/api/current/index.html#scala.Tuple2). That is, we have two-element *tuples* as records, where the first element is a `String` for a file's fully-qualified path and the second element is a `String` for the contents of that file. This is what `SparkContext.wholeTextFiles` returns for us. We'll use the path to remember where we found words, while the contents contains the words themselves (of course).

To recap, the following two types are equivalent:

- `RDD[(String,String)]` - Note parentheses nested in brackets, `[(...)]`.
- `RDD[Tuple2[String,String]]` - Note nested brackets `[...[...]]`, not `[(...)]`.

> **NOTE:** The number in brackets `[N]` in `MapPartitionsRDD[N]` is actually an internal identifier used by Spark. In the output for `fileContents`, we're actually calling `toString` on the instance. It's confusing that `MapPartitionsRDD.toString` uses `[...]`, which looks like a type signature.

We'll see shortly that you can also write *instances* of Tuple2[T1,T2] (http://www.scala-lang.org/api/current/index.html#scala.Tuple2) with the same syntax, e.g., `("foo", 101)`, for a `(String,Int)` tuple, and similarly for *higher-arity* tuples (up to 22 elements...), e.g., `("foo", 101, 3.14159, ("bar", 202L))`. Run the next cell to see the type signature for this last tuple.

```
val footuple = ("foo", 101L, 3.14159F, ("bar", 202L))
```

Took: 773 milliseconds, at 2017-3-12 11:32

```
val (name, age, pi, (otherName, otherAge)) = footuple
```

Took: 776 milliseconds, at 2017-3-12 11:32

Do you understand it? Do you see that it's a four-element tuple and not a five-element tuple? This is because the `("bar", 202L)` is a nested tuple. It's the fourth element of the outer tuple.

Exercise: Using the next cell, try creating some more tuples with elements of different types. Trying writing assignment statements that decompose the tuple, as in the previous code cell.

```
(1,2)
```

How many `fileContents` records do we have? Not many. It should be the same number as the number of files we downloaded above.

```
fileContents.count
```

```
8
```

Took: 1 second 328 milliseconds, at 2017-3-12 11:35

> **NOTE:** We called the `RDD.count` method, whereas most Scala collections have a `size` method.

Let's look at the data, but recall that each record contains the entire contents of one play, so let's not just print the whole record, but the first 200 characters of the play.

```
fileContents.take(1).foreach(name_play => println(s"${name_play._1}:\n${name_play._2.ta
```

Took: 1 second 243 milliseconds, at 2017-3-12 11:35

Now for our next step in the calculation. First, "tokenize" the contents into words by splitting on non-alphanumeric characters, meaning all runs of whitespace (including the newlines), punctuation, etc.

Next, the fully-qualified path is verbose and the same prefix is repeated for all the files, so let's extract just the last element of it, the unique file name.

Then form new tuples with the words and file names.

> **Note:** This "tokenization" approach is very crude. It improperly handles contractions, like `it's` and hyphenated words like `world-changing`. When you kill Google, be sure to use a real *natural language processing* (NLP) tokenization technique.

```
val wordFileNameOnes = fileContents.flatMap { location_contents_tuple2 => // i.e., (fil
    val words = location_contents_tuple2._2.split("""\W+""")              // mytuple._2
    val fileName = location_contents_tuple2._1.split(pathSeparator).last  // mytuple._1
    words.map(word => ((word, fileName), 1))        // create a new tuple to return. No
}
wordFileNameOnes
```

```
MapPartitionsRDD[12] at flatMap at <console>:75
```

Took: 903 milliseconds, at 2017-3-12 11:36

I find this hard to read and shortly I'll show you a much more elegant, alternative syntax.

Let's understand the difference between `map` and `flatMap`. If I called `fileContents.map`, it would return exactly *one* new record for each record in *fileContents*. What I actually want instead are new records for each word-fileName combination, a significantly larger number (but the data in each record will be much smaller).

Using `fileContents.flatMap` gives me what I want. Instead of returning one output record for each input record, a `flatMap` returns a *collection* of new records, zero or more, for *each* input record. These collections are then *flattened* into one big collection, another `RDD` in this case.

```
wordFileNameOnes.count
```

```
173336
```

Took: 1 second 202 milliseconds, at 2017-3-12 11:36

```
wordFileNameOnes.take(10).foreach(println)
```

Took: 870 milliseconds, at 2017-3-12 11:36

What should `flatMap` actually do with each record? I pass a *function* to define what to do. I'm using an unnamed or *anonymous* function. The syntax is `argument_list => body`:

```
location_contents_tuple2 =>
    val words = ...
    ...
}
```

I have a single argument, the record, which I named `location_contents_tuple2`, a verbose way to say that it's a two-element tuple with an input file's location and contents. I don't require a type parameter after `location_contents_tuple2`, because it's inferred by Scala. The => "arrow" separates the argument list from the body, which appears on the next few lines.

When a function takes more than one argument or you add explicit type *annotations* (e.g., `: (String,Int,Double)`), then you need parentheses. Here are three examples:

```
(some_tuple3: (String,Int,Double)) => ...
(arg1, arg2, arg3) => ...
(arg1: String, arg2: Int, arg3: Double) => ...
```

We're letting Scala infer the argument type in our case, `(String,String)`.

Wait, I said we're passing a function as an argument to `flatMap`. If so, why am I using braces `{...}` around this function argument instead of parentheses `(...)` like you would normally expect when passing arguments to a method like `flatMap`?

It's because Scala lets us substitute braces instead of parentheses so we have the familiar block-like syntax `{...}` we know and love for `if` and `for` expressions. I could use either braces or parentheses here. The convention in the Scala community is to use braces for a multi-line anonymous function and to use parentheses for a single expression when it fits on the same line.

Now, for each `location_contents_tuple2`, I access the *first* element using the `_1` method and the *second* element using `_2`.

The file `contents` is in the second element. I split it by calling Java's `String.split` method, which takes a *regular expression* string. Here I specify a regular expression for one or more, non-alphanumeric characters. `String.split` returns an `Array[String]` of the words.

```
val words = location_contents_tuple2._2.split("""\W+""")
```

For the first tuple element, I extract the file name at the end of the location path. This isn't necessary, but it makes the output more readable if I remove the long, common prefix from the path.

```
val fileName = location_contents_tuple2._1.split(pathSeparator).last
```

Finally, still inside the anonymous function passed to `flatMap`, I use Scala's `Array.map` (*not* `RDD.map`) to transform each `word` into a tuple of the form `((word, fileName), 1)`.

```
words.map(word => ((word, fileName), 1))
```

Why did I embed a tuple of `(word, fileName)` inside the "outer" tuple with a `1` as the second element? Why not just write a three-element tuple, `(word, fileName, 1)`? It's because I'll use the `(word, fileName)` as a *key* in the next step, where I'll find all unique word-fileName combinations (using the equivalent of a `group by` statement). So, using the nested `(word, fileName)` as my *key* is most convenient. The `1` *value* is a "seed" count, which I'll use to count the occurrences of the unique `(word, fileName)` pairs.

Notes:

- For historical reasons, tuple indices start at 1, not 0. Arrays and other Scala collections index from 0 (like in Java).
- Recall that *method* arguments have to be declared with types, but that's usually *not* required for *function* arguments.
- Another benefit of triple-quoted strings that makes them nice for regular expressions is that you don't have to escape regular expression metacharacters, like `\W`. If I used a single-quoted string, I would have to write it as `"\\W+"`. Your choice...

Let's count the number of records we have and look at a few of the lines. We'll use the `RDD.take` method to grab the first 10 lines, then loop over them and print them.

```
wordFileNameOnes.count
```

```
173336
```

Took: 1 second 31 milliseconds, at 2017-3-12 11:40

```
wordFileNameOnes.take(10).foreach(println)
```

Took: 951 milliseconds, at 2017-3-12 11:40

We asked for results, so we forced Spark to run a job to compute results. Spark pipelines, like `iiFirstPass1` are *lazy*; nothing is computed until we ask for results.

When you're learning, it's useful to print some data to better understand what's happening. Just be aware of the extra overhead of running lots of Spark jobs.

The first record shown has `""` (blank) as the word:

```
((,asyoulikeit),1)
```

Also, some words have all capital letters:

```
((DRAMATIS,asyoulikeit),1)
```

(You can see where these capitalized words occur if you look in the original files.) Later on, We'll filter out the blank-word records and use lower case for all words.

Now, let's join all the unique `(word,fileName)` pairs together.

```
val uniques = wordFileNameOnes.reduceByKey((count1, count2) => count1 + count2)
uniques
```

```
ShuffledRDD[13] at reduceByKey at <console>:77
```

Took: 814 milliseconds, at 2017-3-12 11:40

In SQL you would use `GROUP BY` for this (including SQL queries you might write with Spark's Dataset/DataFrame (http://spark.apache.org/docs/latest/sql-programming-guide.html) API). However, in the `RDD` API, this is too expensive for our needs, because we don't care about the groups themselves, the long list of repeated `(word,fileName)` pairs. We only care about how many elements are in each group, that is their *size*. That's the purpose of the `1` in the tuples and the use of `RDD.reduceByKey`. It brings together all records with the same key, the unique `(word,fileName)` pairs, and then applies the anonymous function to "reduce" the values, the `1`s. I simply sum them up to compute the group counts.

Note that the anonymous function `reduceByKey` expects must take two arguments, so I need parentheses around the argument list. Since this function fits on the same line, I used parentheses for `reduceByKey`, instead of braces.

> **Note:** All the `*ByKey` methods operate on two-element tuples and treat the first element as the key, by default.

How many are there? Let's see a few:

```
uniques.count
```

```
27276
```

Took: 1 second 367 milliseconds, at 2017-3-12 11:40

As you would expect from a `GROUP BY`-like statement, the number of records is smaller than before. There are about 1/6 as many records now, meaning that on average, each `(word,fileName)` combination appears 6 times.

```
uniques.take(30).foreach(println)
```

Took: 906 milliseconds, at 2017-3-12 11:40

For *inverted index*, we want our final keys to be the words themselves, so let's restructure the tuples from `((word,fileName),count)` to `(word,(fileName,count))`. Now, I'll still output two-element, key-value tuples, but the `word` will be the key and the `(fileName,count)` tuple will be the value.

```
val words = uniques.map { word_file_count_tup3 =>
    (word_file_count_tup3._1._1, (word_file_count_tup3._1._2, word_file_count_tup3._2))
}
```

Took: 595 milliseconds, at 2017-3-12 11:43

The nested tuple methods, e.g., `_1._2`, are hard to read, making the logic somewhat obscure. We'll see a beautiful and elegant alternative shortly.

Now I'll use an actual `group by` operation, because I now need to retain the groups. Calling `RDD.groupByKey` uses the first tuple element, now just the `words`, to bring together all occurrences of the unique words. Next, I'll sort the result by word, ascending alphabetically.

```
val wordGroups = words.groupByKey.sortByKey(ascending = true)
wordGroups
```

```
ShuffledRDD[18] at sortByKey at <console>:81                Took: 902 milliseconds, at 2017-3-12 11:43
```

Note that each group is actually a Scala Iterable (http://www.scala-lang.org/api/current/index.html#scala.collection.Iterable), i.e., an abstraction for some sort of collection. (It's actually a Spark-defined, private collection type called a `CompactBuffer`.)

```
wordGroups.count
```

```
11951                                          Took: 1 second 83 milliseconds, at 2017-3-12 11:43
```

```
wordGroups.take(30).foreach(println)
```

```
                                               Took: 1 second 180 milliseconds, at 2017-3-12 11:43
```

Finally, let's clean up these `CompactBuffers`. Let's convert each to a Scala Vector (http://www.scala-lang.org/api/current/index.html#scala.collection.immutable.Vector) (a collection with *O(1)* performance for most operations), then sort it *descending* by count, so the locations that mention the corresponding word the *most* appear *first* in the list. (Think about how you would want a search tool to work...)

Note we're using `Vector.sortBy`, not an `RDD` sorting method. It takes a function that accepts each collection element and returns something used to sort the collection. By returning `(-fileNameCountTuple2._2, fileNameCountTuple2)`, I effectively say, "sort by the counts *descending* first, then sort by the file names." Why does `-fileNameCountTuple2._2` cause counts to be sorted descending, because I'm returning the negative of the value, so larger counts will be less than smaller counts, e.g., `-3 < -2`.

Finally, I take the resulting `Vector` and make a comma-separated string with the elements, using the helper method `mkString`.

What's `RDD.mapValues`? I could use `RDD.map`, but I'm not changing the keys (the words), so rather than have to deal with the tuple with both elements, `mapValues` just passes in the value part of the tuple and reconstructs new `(key,value)` tuples with the new value that my function returns. So, `mapValues` is more convenient to use than `map` when I have two-element tuples and I'm not modifying the keys.

```
val iiFirstPass2 = wordGroups.mapValues { iterable =>
    val vect = iterable.toVector.sortBy { file_count_tup2 =>
        (-file_count_tup2._2, file_count_tup2._1)
    }
    vect.mkString(",")
}
```

Took: 774 milliseconds, at 2017-3-12 11:46

We're done! The number of records is the same as for `wordGroups` (do you understand why?), so let's just see see some of the records.

```
iiFirstPass2.take(30).foreach(println)
```

Took: 973 milliseconds, at 2017-3-12 11:46

Okay. Looks reasonable.

Next, I'll refine the code using a very powerful feature, *pattern matching*, which both makes the code more concise and easier to understand. It's my *favorite* feature of Scala.

Before I do that, try a few refinements on your own.

Exercises:

Try the following exercises in the next cell, which is a copy of `iiFirstPast`:

- Convert all words to lower case. Calling `toLowerCase` on a string is all you need. Where's a good place to insert this logic?
- Add a filter statement to remove the first entry for the blank word `""`. You could do this one of two ways, using another "step" with RDD.filter (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.rdd.RDD) (search the Scaladoc page for the `filter` method), *or* using the similar Scala collections method, scala.collection.Seq.filter (http://www.scala-lang.org/api/current/index.html#scala.collection.Seq). Both versions take a *predicate* function, one that returns `true` if the record should be *retained* and `false` otherwise. Do you think one choice is better than the other? Why? Or, are they basically the same? Reasons might include code comprehension and performance of one over the other.

I'll implement both changes in subsequent refinements below.

```
val iiFirstPass2 = sc.wholeTextFiles(shakespeare.toString).
    flatMap { location_contents_tuple2 =>
        val words = location_contents_tuple2._2.split("""\W+""")
        val fileName = location_contents_tuple2._1.split(pathSeparator).last
        words.filter(_.length != 0).map(word => ((word.toLowerCase, fileName), 1))
    }.
    reduceByKey((count1, count2) => count1 + count2).
    map { word_file_count_tup3 =>
        (word_file_count_tup3._1._1, (word_file_count_tup3._1._2, word_file_count_tup3.
    }.
    groupByKey.  // (word, iterable(...))
    sortByKey(ascending = true).

    mapValues { iterable =>
        val vect = iterable.toVector.sortBy { file_count_tup2 =>
            (-file_count_tup2._2, file_count_tup2._1)
        }
        vect.mkString(",")
    }
```

Took: 1 second 724 milliseconds, at 2017-3-12 11:48

Let's try it!

```
iiFirstPass2.take(30).foreach(println)
```

Took: 1 second 101 milliseconds, at 2017-3-12 11:49

Pattern Matching

We've studied a real program and we've learned quite a bit of Scala. Let's improve it with my favorite Scala feature, *pattern matching*.

Here's the "first pass" version again for easy reference.

```
val iiFirstPass3 = sc.wholeTextFiles(shakespeare.toString).
    flatMap { location_contents_tuple2 =>
        val words = location_contents_tuple2._2.split("""\W+""")
        val fileName = location_contents_tuple2._1.split(pathSeparator).last
        words.map(word => ((word, fileName), 1))
    }.
    reduceByKey((count1, count2) => count1 + count2).
    map { word_file_count_tup3 =>
        (word_file_count_tup3._1._1, (word_file_count_tup3._1._2, word_file_count_tup3.
    }.
    groupByKey.
    sortByKey(ascending = true).
    mapValues { iterable =>
        val vect = iterable.toVector.sortBy { file_count_tup2 =>
            (-file_count_tup2._2, file_count_tup2._1)
        }
        vect.mkString(",")
    }
```

Took: 1 second 709 milliseconds, at 2017-3-12 11:51

Now here is a new implementation that uses *pattern matching*.

I've also made two other additions, the solutions to the last exercises, which remove empty words "" and fix mixed capitalization, using the following additions:

- `filter(word => word.size > 0)` to remove the empty words. (In Spark and Scala collections, `filter` has the positive sense; what should be retained?) It's indicated by the comment `// #1`.
- `word.toLowerCase` to convert all words to lower case uniformly, so that words like HAMLET, Hamlet, and hamlet in the original texts are treated as the same, since we're counting word occurrences. See comment `// #2`.

```
val iiPatternMatching = sc.wholeTextFiles(shakespeare.toString).
  flatMap {
        case (location, contents) =>
            val words = contents.split("""\W+""").
                filter(word => word.size > 0)                      // #1
            val fileName = location.split(pathSeparator).last
            words.map(word => ((word.toLowerCase, fileName), 1))   // #2
    }.
    reduceByKey((count1, count2) => count1 + count2).
    map {
        case ((word, fileName), count) => (word, (fileName, count))
    }.
    groupByKey.
    sortByKey(ascending = true).
    mapValues { iterable =>
        val vect = iterable.toVector.sortBy {
            case (fileName, count) => (-count, fileName)
        }
        vect.mkString(",")
    }
```

Took: 1 second 554 milliseconds, at 2017-3-12 11:51

```
iiPatternMatching.take(10).foreach(println)
```

Took: 1 second 246 milliseconds, at 2017-3-12 11:53

There are more things we can do with pattern matching. We'll return to it below.

Let's understand this code. First, compare how I did the filtering for empty words and conversion to lower case with your exercise solutions above. Inside the function passed to `flatMap`, I filtered for empty words immediately after splitting the contents into words and I converted to lower case as I constructed the new tuples. My choice reduces the number of output records from `flatMap` by at most one record per input line, which shouldn't have a significant impact on performance. Filtering itself adds some extra overhead.

Also, the way Spark implements steps like `map`, `flatMap`, `filter`, it would incur about the same overhead if I added an `RDD.filter` step instead. Note that we could also do the filtering later in the pipeline, after `groupByKey`, for example. So, whichever approach you implemented above is probably fine. You could do performance profiling of the different approaches, but you may not notice a significance difference until you use very large input data sets.

The function I pass to `flatMap` now looks like this:

```
flatMap {
    case (location, contents) =>
        val words = contents.split("""\W+""").
            filter(word => word.size > 0)                      // #1
        val fileName = location.split(pathSeparator).last
        words.map(word => ((word.toLowerCase, fileName), 1))   // #2
}.
```

Compare it to the previous version (ignoring the enhancements for blank words and capitalization, marked with the #1 and #2 comments):

```
flatMap { location_contents_tuple2 =>
    val words = location_contents_tuple2._2.split("""\W+""")
    val fileName = location_contents_tuple2._1.split(pathSeparator).last
    words.map(word => ((word, fileName), 1))
}.
```

Instead of `location_contents_tuple2` a variable name for the whole tuple, I wrote `case (location, contents)`. The `case` keyword says I want to *pattern match* on the object passed to the function. If it's a two-element tuple (and I know it always will be in this case), then *extract* the first element and assign it to a variable named `location` and extract the second element and assign it to a variable named `contents`.

Now, instead of accessing the location and content with the slighly obscure and verbose `location_contents_tuple2._1` and `location_contents_tuple2._2`, respectively, I use meaningful names, `location` and `contents`. The code becomes more concise and more readable.

This special function syntax is called a *partial function*. It's partial in the sense that it might only handle some inputs, not all possible inputs, which would make it *total*. Partial functions have the form:

```
{                              // Opening curly brace (you can't use parenthese
s here)
  case pattern1 => body1       // First pattern and body. For multi-line bodie
s, put on next lines.
  case pattern2 => body2       // Second pattern and body (optional).
  ...                          // Optional additional patterns.
}                              // Closing curly brace.
```

Each record will be tested against the patterns, in order, until a match is found. In our case, we know the format will be a two-element tuple, so `case (location, contents)` matches perfectly and only one `case` clause is required. If I fail to match any of the patterns, an exception is thrown.

The `reduceByKey` step is unchanged:

```
reduceByKey((count1, count2) => count1 + count2).
```

To be clear, this isn't a pattern-matching expression; there is no `case` keyword. It's just a "regular" function that takes two arguments, for the two things I'm adding.

My favorite improvement is the next line:

```
map {
    case ((word, fileName), count) => (word, (fileName, count))
}.
```

Compare it to the previous, obscure version:

```
map { word_file_count_tup3 =>
    (word_file_count_tup3._1._1, (word_file_count_tup3._1._2, word_file_count_
tup3._2))
}.
```

The new implementation makes it clear what I'm doing; just shifting parentheses! That's all it takes to go from the `(word, fileName)` keys with `count` values to `word` keys and `(fileName, count)` values. Note that pattern matching works just fine with nested structures, like `((word, fileName), count)`.

I hope you can appreciate how elegant and concise this expression is! Note how I thought of the next transformation I needed to do in preparation for the final group-by, to switch from `((word, fileName), count)` to `(word, (fileName, count))` and *I just wrote it down exactly as I pictured it!*

Code like this makes writing Scala Spark code a sublime experience for me. I hope it will for you, too ;)

The next two expressions are unchanged:

```
groupByKey.
sortByKey(ascending = true).
```

The final `mapValues` now uses pattern matching to sort the `Vector` in each record:

```
mapValues { iterable =>
    val vect = iterable.toVector.sortBy {
        case (fileName, count) => (-count, fileName)
    }
    vect.mkString(",")
}
```

Compared to the original version, it's again easier to read:

```
mapValues { iterable =>
    val vect = iterable.toVector.sortBy { file_count_tup2 =>
        (-file_count_tup2._2, file_count_tup2._1)
    }
    vect.mkString(",")
}
```

The function I pass to `sortBy` returns a tuple used for sorting, with `-count` to force *descending* numerical sort (biggest first) and `fileName` to secondarily sort by the file name, for equivalent counts. I could ignore file name order and just return `-count` (not a tuple). However, if you need more repeatable output in a distributed system like Spark, say for example to use in unit test validation, then the secondary sorting by file name is handy.

Compare this code to `iiFirstPass3`. It is much clearer to write just the constituents and the literal representations of their nested structures in parentheses, rather than using tuple variable with accessors, like `_1` and `_2`.

Final Implementation and Using DataFrames

Let's verify we still get reasonable results. Let's also transition to Spark's DataFrame (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrame) API for its convenient display options. `DataFrames` are part of Spark SQL (http://spark.apache.org/docs/latest/sql-programming-guide.html), as is SparkSession (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.SparkSession). We can create a `DataFrame` from our `RDD` using the `SparkSession`. In Spark 1.X, you would create a SQLContext (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.SQLContext) instead. `SQLContext` is still available, but deprecated.

Now convert the `RDD` to a `DataFrame` with `spark.createDataFrame`, then use `toDF` (convert to another `DataFrame`?) with new names for each "column". (If you want to back port this code to Spark 1.6.X, either rename `spark` to `sqlContext`, the conventional name for a `SQLContext` instance, or just use the name `spark` for the instance.)

```
val iiPatternMatchingDF = sparkSession.createDataFrame(iiPatternMatching).toDF("word",
```

Took: 2 seconds 869 milliseconds, at 2017-3-12 11:57

If we ask for the value of a DataFrame, Spark Notebook renders it with a paged, table view.

```
iiPatternMatchingDF
```

1 >>

Took: 2 seconds 888 milliseconds, at 2017-3-12 11:57

word	locations_counts

You could also use the `show` command, especially when using the `spark-shell`. It has two optional arguments:

```
iiPatternMatchingDF.show(truncate=false, numRows=5)
```

Took: 1 second 150 milliseconds, at 2017-3-12 12:0

```
iiPatternMatchingDF.printSchema
```

Took: 833 milliseconds, at 2017-3-12 12:0

Our Final Version: Supporting SQL Queries

To play with some more Spark, let's write SQL queries to explore the resulting data.

To do this, let's first refine the output. Instead of creating a string for the list of `(location,count)` pairs, which is opaque to our SQL schema (i.e., just a string), let's "unzip" the collection into two arrays, one for the `locations` and one for the `counts`. That way, if we ask for the first element of each array, we'll have nicely separate fields that work better with Spark SQL queries.

"Zipping" and "unzipping" work like a mechanical zipper. If I have a collection of tuples, say `List[(String, Int)]`, I convert this single collection of "zippered" values into two collections (in a tuple) of single values, `(List[String], List[Int])`. Zipping is the inverse operation.

```
val x = Seq("one", "two", "three")
val y = Seq(1, 2, 3)
val xy = x.zip(y)
```

Took: 623 milliseconds, at 2017-3-12 12:3

Note the type signature of `xy` above. Now note the type and result of calling `unzip`.

```
xy.unzip
```

```
(List(one, two, three),List(1, 2, 3))
```

Took: 669 milliseconds, at 2017-3-12 12:3

Here is our final implementation, `ii1` rewritten with this change.

```
val ii = sc.wholeTextFiles(shakespeare.toString).
    flatMap {
        case (location, contents) =>
            val words = contents.split("""\W+""").
                filter(word => word.size > 0)                      // #1
            val fileName = location.split(pathSeparator).last
            words.map(word => ((word.toLowerCase, fileName), 1))   // #2
    }.
    reduceByKey((count1, count2) => count1 + count2).
    map {
        case ((word, fileName), count) => (word, (fileName, count))
    }.
    groupByKey.
    sortByKey(ascending = true).
    map {                         // Must use map now, because we'll format new records
      case (word, iterable) =>    // Hence, pattern match on the whole input record.

        val vect = iterable.toVector.sortBy {
            case (fileName, count) => (-count, fileName)
        }

        // Use `Vector.unzip`, which returns a single, two element tuple, where each
        // element is a collection, one for the locations and one for the counts.
        // I use pattern matching to extract these two collections into variables.
        val (locations, counts) = vect.unzip

        // Lastly, I'll compute the total count across all locations and return
        // a new record with all four fields. The `reduceLeft` method takes a function
        // that knows how to "reduce" the collection down to a final value, working
        // from the left.
        val totalCount = counts.reduceLeft((n1,n2) => n1+n2)

        (word, totalCount, locations, counts)
    }
```

Took: 1 second 815 milliseconds, at 2017-3-12 12:3

We've changed the ending `mapValues` call to a `map` call, because we'll construct entirely new records, not just new values with the same keys. Hence the full records, two-element tuples are passed in, rather than just the values, so we'll pattern match on the tuple:

```
    map {                         // Must use map now, because we'll format ne
w records.
      case (word, iterable) =>    // Hence, pattern match on the whole input r
ecord.

        val vect = iterable.toVector.sortBy {
            case (fileName, count) => (-count, fileName)
        }
```

We have a `Vector[String, Int]` of two-element tuples `(fileName, count)`. We use `Vector.unzip` to create a single, two element tuple, where each element is now a collection, one for the locations and one for the counts. The type is `(Vector[String], Vector[Int])`.

We can also use pattern matching with assignment! We immediately decompose the two-element tuple:

```
        // I use pattern matching to extract these two collections into variab
les.
        val (locations, counts) = vect.unzip
```

Finally, it's convenient to know how many locations and counts we have, so we'll compute another new column for the their count and format a four-element tuple as the final output.

```
        // Lastly, I'll compute the total count across all locations and retur
n
        // a new record with all four fields. The `reduceLeft` method takes a
 function
        // that knows how to "reduce" the collection down to a final value, wo
rking
        // from the left.
        val totalCount = counts.reduceLeft((n1,n2) => n1+n2)

        (word, totalCount, locations, counts)
    }
```

Okay! Now let's create a DataFrame (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrame) with this data. The `toDF` method just returns the same `DataFrame`, but with appropriate names for the columns, instead of the synthesized names that `createDataFrame` generates (e.g., `_c1`, `_c2`, etc.)

Caching the `DataFrame` in memory prevents Spark from recomputing `ii` from the input files *every time* I write a query!

Finally, to use SQL, I need to "register" a temporary table.

```
val iiDF = sparkSession.createDataFrame(ii).toDF("word", "total_count", "locations", "c
iiDF.cache
iiDF.createOrReplaceTempView ("inverted_index")  // Use registerTempTable for Spark 1.6
```

Took: 1 second 759 milliseconds, at 2017-3-12 12:3

Let's remind ourselves of the schema:

```
iiDF.printSchema
```

Took: 779 milliseconds, at 2017-3-12 12:4

The following SQL query extracts the top location by count for each word, as well as the total count across all locations for the word. The Spark SQL dialect supports Hive SQL syntax (plus extensions in Spark 2.X) for extracting elements from arrays, maps, and structs (details (https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF#LanguageManualUDF-CollectionFunctions)). Here I access the first element (index zero) from each array, using bracket syntax, `[n]`.

```
val topLocations = sparkSession.sql("""
SELECT word, total_count, locations[0] AS top_location, counts[0] AS top_count
FROM inverted_index
""")
topLocations
```

1 >> Took: 2 seconds 30 milliseconds, at 2017-3-12 12:4

word	total_count	top_location	top_count

Exercises: The next cell copies the previous query, as a starting point.

- Change the query to return the top two locations and counts.
- Try other queries. For example, find all occurrences of words like `love`, `hate`, etc.

See the Appendix for some solutions.

```
val topTwoLocations = sparkSession.sql("""
SELECT word, total_count, locations[0] AS top_location, counts[0] AS top_count
FROM inverted_index
""")
topTwoLocations
```

Removing the "Stop Words"

Did you notice that one record we saw above was for the word "a". Not very useful if you're using this data for text searching, *sentiment mining*, etc. So called *stop words*, like *a*, *an*, *the*, *he*, *she*, *it*, etc., could also be removed.

Recall the `filter` logic I added to remove "", `word => word.size > 0`. I could replace it with `word => keep(word)`, where `keep` is a method that does any additional filtering I want, like removing stop words.

Exercise:

- Complete the implementation the following `keep(word: String):Boolean` method. It currently just returns false if the word is zero length. Change it to add a check to see if the word appears in a list of *stop words*. If so, also return false. Hard code a short list of your own words (e.g., `the`, `a`, `he`, `she`, etc.). (See the Appendix for the solution.)
- "Extra Credit": Use a Spark [broadcast variable] for the `Set` of stop words.

Question: why is a Set (http://www.scala-lang.org/api/2.11.8/#scala.collection.Set) a good data structure to use for stop words?

```
def keep(word: String): Boolean = word.size > 0
```

Took: 488 milliseconds, at 2017-3-12 12:6

We call `keep` in the revised version of the program. See the `// here...` comment.

```
val iistop = sc.wholeTextFiles(shakespeare.toString).
    flatMap {
        case (location, contents) =>
            val words = contents.split("""\W+""").
                filter(keep)                        // here. Recall you could also writ
            val fileName = location.split(pathSeparator).last
            words.map(word => ((word.toLowerCase, fileName), 1))   // #2
    }.
    reduceByKey((count1, count2) => count1 + count2).
    map {
        case ((word, fileName), count) => (word, (fileName, count))
    }.
    groupByKey.
    sortByKey(ascending = true).
    map {                            // Must use map now, because we'll format new records
      case (word, iterable) =>       // Hence, pattern match on the whole input record.

        val vect = iterable.toVector.sortBy {
            case (fileName, count) => (-count, fileName)
        }

        // Use `Vector.unzip`, which returns a single, two element tuple, where each
        // element is a collection, one for the locations and one for the counts.
        // I use pattern matching to extract these two collections into variables.
        val (locations, counts) = vect.unzip

        // Lastly, I'll compute the total count across all locations and return
        // a new record with all four fields. The `reduceLeft` method takes a function
        // that knows how to "reduce" the collection down to a final value, working
        // from the left.
        val totalCount = counts.reduceLeft((n1,n2) => n1+n2)

        (word, totalCount, locations, counts)
    }
```

Took: 1 second 710 milliseconds, at 2017-3-12 12:6

```
iistop.take(50).foreach(println)
```

Took: 944 milliseconds, at 2017-3-12 12:6

More on Pattern Matching Syntax

We've only scratched the surface of pattern matching. Let's explore it some more.

Here's another anonymous function using pattern matching that extends the previous function we passed to `flatMap`:

```
    {
        case (location, "") =>
            Array.empty[((String, String), Int)]  // Return an empty array
        case (location, contents) =>
            val words = contents.split("""\W+""")
            val fileName = location.split(pathSep).last
            words.map(word => ((word, fileName), 1))
    }.
```

You can have multiple `case` clauses, some of which might match on specific literal values ("" in this case) and others which are more general. The first case clause handles files with no content. The second clause is the same as before.

Pattern matching is *eager*. The first successful match in the order as written will win. If you reversed the order here, the `case (location, "")` would never match and the compiler would throw an "unreachable code" warning for it.

Note that you don't have to put the lines after the => inside braces, `{...}` (although you can). The => and `case` keywords (or the final `}`) are sufficient to mark these blocks. Also, for a single-expression block, like the one for the first case clause, you can put the expression on the same line after the => if you want (and it fits).

Finally, if none of the case clauses matches, then a MatchError (http://www.scala-lang.org/api/current/index.html#scala.MatchError) exception is thrown. In our case, we *always* know we'll have two-element tuples, so the examples so far are fine.

Here's a final contrived example to illustrate what's possible, using a sequence of objects of different types. First, we define a class for `Person`. We'll explain `case class` syntax later, but for now, just think of it as a simple Java class, where the argument list is automatically converted to fields of the class.

```
case class Person(name: String, age: Int)
```

Took: 779 milliseconds, at 2017-3-12 12:10

```
val stuff = Seq(1, 3.14159, 2L, 10, 4.4F, ("one", 1), (404F, "boo"), ((11, 12), 21, 31)
               "hello", new Person("Dean", 100))

stuff.foreach {
    case 10                   => println(s"Found a 10")
    case i: Int               => println(s"Found an Int:   $i")
    case l: Long              => println(s"Found a Long:   $l")
    case f: Float             => println(s"Found a Float:  $f")
    case d: Double            => println(s"Found a Double: $d")
    case Person(name2, age2)  => println(s"A person: $name2, $age2")
    case (x1, x2) =>
        println(s"Found a two-element tuple with elements of arbitrary type: ($x1, $x2)
    case ((x1a, x1b), _, x3) =>   // The "_" means match on this field, but then ignore
        println(s"Found a three-element tuple with 1st and 3th elements: ($x1a, $x1b),
    case default              => println(s"Found something else: $default")
}
```

Took: 1 second 214 milliseconds, at 2017-3-12 12:10

A few notes.

- A literal like `1` is inferred to be `Int`, while `3.14159` is inferred to be `Double`. Add `L` or `F`, respectively, to infer `Long` or `Float` instead.
- Note how we mixed specific type checking, e.g., `i: Int`, with more loosely-typed expressions, e.g., `(x1, x2)`, which expects a two-element tuple, but the element types are unconstrained.
- All the words `i`, `l`, `f`, `d`, `x1`, `x2`, `x3`, and `default` are arbitrary variable names. Yes `default` is not a keyword, but an arbitrary choice for a variable name. We could use anything we want.
- The last `default` clause specifies a variable with no type information. Hence, it matches *anything*, which is why this clause must appear last. This is the idiom to use when you aren't sure about the types of things you're matching against and you want to avoid a possible MatchError (http://www.scala-lang.org/api/current/index.html#scala.MatchError).
- If you want to match that something *exists*, but you don't need to bind it to a variable, then use _, as in the three-element tuple example.
- The three-element tuple example also demonstrates that arbitrary nesting of expressions is supported, where the first element is expected to be a two-element tuple.

All the anonymous functions we've seen that use these pattern matching clauses have this format:

```
{
    case firstCase => ...
    case secondCase => ...
    ...
}
```

This format has a special name. It's called a *partial function*. All that means is that we only "promise" to accept arguments that match at least one of our `case` clauses, not any possible input.

The other kind of anonymous function we've seen is a *total function*, to be precise.

Recall we said that for total functions you can use either `(...)` or `{...}` around them, depending on the "look" you want. For *partial functions*, you *must* use `{...}`.

Also, recall that we used pattern matching with assignment, even with our case class `Person`!

```
val (a, (b, (c1, _), d)) = ("A", ("B", ("C1", "C2"), "D"))
println(s" $a, $b, $c1, $c2, $d")
val Person(name, _) = Person("Dean", 29)
```

Try adding an `"E"` element to the tuple on the right-hand side, without changing the left-hand side. What happens? Try removing the `"D"` and `"E"` elements. What happens now?

We'll come back to one last example of pattern matching when we discuss *case classes*.

"Scala for Spark 102"

We've covered a lot already in this notebook, focusing on the most important topics you need to know about Scala for daily use. Let's call them the "Scala for Spark 101" material.

At this point, I suggest you create a new notebook and play with Spark using what you've learned so far, then come back to this point if you run into something we didn't cover already. Chances are you're ready to learn the next bits of useful Scala, the "102" material.

What Runs Where?

Consider Spark code like the following:

```
myRDD.map {
  case ((word, fileName), count) => (word, (fileName, count))
}
```

It's completely unobvious where this code actually *runs*. In a typical compiled Scala program or the Scala interpreter, it would all run in the process of the program or interpreter. That's also true when you use local mode for Spark, but *not* when you run in a cluster.

Locally and in a cluster situation, your *driver* program (including the Spark shell) runs the `myRDD.map` code immediately, which constructs a directed, acyclic graph (DAG) of processing steps to run *later*, when you ask for results. In other words, Spark is lazy and only runs the computation on demand. So, what happens to the function passed to `map` in this case?

Spark serializes the function so it can be transmitted to nodes on the cluster for execution. Even in local mode it serializes the code, even though it will be executed in the same JVM process as your driver.

Usually, all this is transparent to you, but sometimes you'll write a function that can't be serialized and you'll get a `NotSerializable` exception. Spark tries to tell you what part of the function isn't serializable, but usually your function references something outside, like a field in an enclosing class, and that pulls in data structures that can't be serialized. I won't discuss this issue further here, but you can search for longer explanations and workarounds.

Scala's Object Model

Scala is a *hybrid*, object-oriented and functional programming language. The philosophy of Scala is that you exploit object orientation for encapsulation of details, i.e., *modularity*, but use functional programming for its logical precision when implementing those details. Most of what we've seen so far falls into the functional programming camp. Much of data manipulation and analysis is really Mathematics. Functional programming tries to stay close to how functions and values work in Mathematics.

However, when writing non-trivial Spark programs, it's occasionally useful to exploit the object-oriented features.

Classes vs. Instances

Scala uses the same distinction between classes and instances that you find in Java. Classes are like *templates* used to create instances.

We've talked about the *types* of things, like `word` is a `String` and `totalCount` is an `Int`. A class defines a *type* in the same sense.

Here is an example class that we might use to represent the inverted index records we just created:

```
class IIRecord1(
    word: String,
    total_count: Int,
    locations: Array[String],
    counts: Array[Int]) {

    /** CSV formatted string, but use [a,b,c] for the arrays */
    override def toString: String = {
        val locStr = locations.mkString("[", ",", "]")  // i.e., "[a,b,c]"
        val cntStr = counts.mkString("[", ",", "]")  // i.e., "[1,2,3]"
        s"$word,$total_count,$locStr,$cntStr"
    }
}

new IIRecord1("hello", 3, Array("one", "two"), Array(1, 2))
```

```
hello,3,[one,two],[1,2]                    Took: 640 milliseconds, at 2017-3-12 12:13
```

When defining a class, the argument list after the class name is the argument list for the *primary constructor*. You can define secondary constructors, too, but it's not very common, in part for reasons we'll see shortly.

Note that when you override a method that's defined in a parent class, like Java's `Object.toString`, Scala requires you to add the `override` keyword.

We created an *instance* of `IIRecord1` using `new`, just like in Java.

Finally, as a side note, we've been using `Ints` (integers) all along for the various counts, but really for "big data", we should probably use `Longs`.

Objects

I've been careful to use the word *instance* for things we create from classes. That's because Scala has built-in support for the Singleton Design Pattern (https://en.wikipedia.org/wiki/Singleton_pattern), i.e., when we only want one instance of a class. We use the `object` keyword.

For example, in Java, you define a class with a `static void main(String[] arguments)` method as your entry point into your program. In Scala, you use an `object` to hold `main`, as follows:

```
object MySparkJob {

    val greeting = "Hello Spark!"

    def main(arguments: Array[String]) = {
        println(greeting)

        // Create your SparkContext or SparkSession, etc., etc.
    }
}
```

Took: 641 milliseconds, at 2017-3-12 12:13

Just as for classes, the name of the object can be anything you want. There is no `static` keyword in Scala. Instead of adding `static` methods and fields to classes as in Java, you put them in objects instead, as here.

> **NOTE:** Because the Scala compiler must generate valid JVM byte code, these definitions are converted into the equivalent, Java-like static definitions in the output byte code.

Case Classes

Tuples are handy for representing records and for decomposing them with pattern matching. However, it would be nice if the fields were *named*, as well as *typed*. A good use for a class, like our `IIRecord1` above, us to represent this structure and give us named fields. Let's now refine that class definition to exploit some extra, very useful features in Scala.

Consider the following definition of a *case class* that represents our final record type.

```
case class IIRecord(
    word: String,
    total_count: Int = 0,
    locations: Array[String] = Array.empty,
    counts: Array[Int] = Array.empty) {

    /**
     * Different than our CSV output above, but see toCSV.
     * Array.toString is useless, so format these ourselves.
     */
    override def toString: String =
        s"""IIRecord($word, $total_count, $locStr, $cntStr)"""

    /** CSV-formatted string, but use [a,b,c] for the arrays */
    def toCSV: String =
        s"$word,$total_count,$locStr,$cntStr"

    /** Return a JSON-formatted string for the instance. */
    def toJSONString: String =
        s"""{
        |  "word":        "$word",
        |  "total_count": $total_count,
        |  "locations":   ${toJSONArrayString(locations)},
        |  "counts"       ${toArrayString(counts, ", ")}
        |}
        |""".stripMargin

    private def locStr = toArrayString(locations)
    private def cntStr = toArrayString(counts)

    // "[_]" means we don't care what type of elements; we're just
    // calling toString on them!
    private def toArrayString(array: Array[_], delim: String = ","): String =
        array.mkString("[", delim, "]")  // i.e., "[a,b,c]"

    private def toJSONArrayString(array: Array[String]): String =
        toArrayString(array.map(quote), ", ")

    private def quote(word: String): String = "\"" + word + "\""
}
```

Took: 983 milliseconds, at 2017-3-12 12:13

I said that defining secondary constructors is not very common. In part, it's because I used a convenient feature, the ability to define default values for arguments to methods, including the primary constructor. The default values mean that I can create instances without providing all the arguments explicitly, as long as there is a default value defined, and similarly for calling methods. Consider these two examples:

```
val hello = new IIRecord("hello")
val world = new IIRecord("world!", 3, Array("one", "two"), Array(1, 2))

println("\n`toString` output:")
println(hello)
println(world)

println("\n`toJSONString` output:")
println(hello.toJSONString)
println(world.toJSONString)

println("\n`toCSV` output:")
println(hello.toCSV)
println(world.toCSV)
```

Took: 1 second 44 milliseconds, at 2017-3-12 12:13

I added `toJSONString` to illustrate adding *public* methods, the default visibility, and *private* methods to a class definition. By the way, when there are no methods or non-field variables to define, I can omit the body complete; no empty `{}` required.

Recall that the `override` keyword is required when redefining `toString`.

Okay, what about that `case` keyword? It tells the compiler to do several useful things for us, eliminating a lot of boilerplate that we would have to write for ourselves with other languages, especially Java:

1. Treat each constructor argument as an immutable (`val`) private field of the instance.
2. Generate a public reader method for the field with the same name (e.g., `word`).
3. Generate *correct* implementations of the `equals` and `hashCode` methods, which people often implement incorrectly, as well as a default `toString` method. You can use your own definitions by adding them explicitly to the body. We did this for `toString`, to format the arrays in a nicer way than the default `Array[_].toString` method.
4. Generate an `object IIRecord`, i.e., with the same name. The object is called the *companion object*.
5. Generate a "factory" method in the companion object that takes the same argument list and instantiates an instance.
6. Generate helper methods in the companion object that support pattern matching.

Points 1 and 2 make each argument behave as if they are public, read-only fields of the instance, but they are actually implemented as described.

Point 3 is important for correct behavior. Case class instances are often used as keys in Maps (http://www.scala-lang.org/api/current/index.html#scala.collection.Map) and Sets (http://www.scala-lang.org/api/current/index.html#scala.collection.Set), Spark RDD and DataFrame methods, etc. In fact, you should *only* use your case classes or Scala's built-in types with well-defined `hashCode` and `equals` methods (like `Int` and other number types, `String`, tuples, etc.) as keys.

For point 4, the *companion object* is generated automatically by the compiler. It adds the "factory" method discussed in point 5, and methods that support pattern matching, point 6. You can explicitly define these methods and others yourself, as well as fields to hold state. The compiler will still insert these other methods. However, see Ambiguities with Companion Objects. The bottom line is that you shouldn't define case classes in notebooks like this with extra methods in the companion object, due to parsing ambiguities.

Point 5 means you actually rarely use `new` when creating instances. That is, the following are effectively equivalent:

```scala
val hello1 = new IIRecord("hello1")
val hello2 = IIRecord("hello2")
```

Took: 532 milliseconds, at 2017-3-12 12:13

What actually happens in the second case, without `new`? The "factory" method is actually called `apply`. In Scala, whenever you put an argument list after any *instance*, including these `objects`, as in the `hello2` case, Scala looks for an `apply` method to call. The arguments have to match the argument list for apply (number of arguments, types of arguments, accounting for default argument values, etc.). Hence, the `hello2` declaration is really this:

```scala
val hello2b = IIRecord.apply("hello2b")
```

Took: 683 milliseconds, at 2017-3-12 12:13

You can exploit this feature, too, in your other classes. We talked about word stemming above. Suppose you write a stemming library and declare an object for as the entry point. Here, I'll just do something simple; assume a trailing "s" means the word is a plural and remove it (a bad assumption...):

```scala
object stem {
    def apply(word: String): String = word.replaceFirst("s$", "") // insert real implem
}

println(stem("dog"))
println(stem("dogs"))
```

Took: 828 milliseconds, at 2017-3-12 12:13

Note how it looks like I'm calling a function or method named `stem`. Scala allows object and class names to start with a lower case letter.

Finally, point 6 means we can use our custom case classes in pattern matching expressions, as we saw previously with a `Person` case class. I won't go into the methods actually implemented in the companion object and how they support pattern matching. I'll just use the "magic" in the following example that "parses" or previously-defined `hello` and `world` instances.

```
Seq(hello, world).map {
    case IIRecord(word, 0, _, _) => s"$word with no occurrences."
    case IIRecord(word, cnt, locs, cnts) =>
        s"$word occurs $cnt times: ${locs.zip(cnts).mkString(", ")}"
}
```

2 entries total

Show: 10

Search:

string value
hello with no occurrences.
world! occurs 3 times: (one,1), (two,2)

Showing 2 of 2 records

Pages: Previous 1 Next

Took: 1 second 117 milliseconds, at 2017-3-12 12:14

The first case clause ignores the locations and counts, because I know they will be empty arrays if the total count is 0!

The second case clause uses the `zip` method to put the locations and counts back together. Recall we used `unzip` to create the separate collections.

Datasets and DataFrames

We've mostly used Spark's RDD API, where it's common to use case classes to represent the "schema" of records when working with RDDs, but also with a new type, Dataset[T] (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.Dataset), analogous to `RDD[T]`, where the `T` represents the type of records.

A problem with DataFrames (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrame) is the fact that the fields are untyped until you try to access them. `Datasets` restore the type safety of `RDDs` by using a case class as the definition of the schema.

`Datasets` were introduced in Spark 1.6.0, but they were somewhat incomplete until the 2.0.0 release, where `Dataset[T]` is the real type you work with when using SparkSQL. `DataFrame` is still around, but now it's a *type alias* for `Dataset[Row]`, where Row (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.Row) is the loosely-typed representation of the rows in the `Dataset` and its columns (or fields).

A *type alias* is another Scala feature where you can give a new name to a type expression. In this case: `type DataFrame = Dataset[Row]`. You use `DataFrame` as if it's declared as a regular type, but you are actually working with a `Dataset[Row]`.

Importing Everything in a Package

In Java, `import foo.bar.*;` means import everything in the `bar` package.

In Scala, `*` is actually a legal method name; think of defining multiplication for custom numeric types, like `Matrix`. Hence, this import statement in Scala would be ambigious. Therefore, Scala uses _ instead of `*`, `import foo.bar._` (with the semicolon inferred).

Incidentally, what would that `*` method definition look like? Something like this:

```
case class Matrix(rows: Array[Array[Double]]) {  // Each row is an Array[Doubl
e]

    /** Multiple this matrix by another. */
    def *(other: Matrix): Matrix = ...

    /** Add this matrix by another. */
    def +(other: Matrix): Matrix = ...

    ...
}

val row1: Array[Array[Double]] = ...
val row2: Array[Array[Double]] = ...
val m1 = Matrix(rows1)
val m2 = Matrix(rows2)
val m1_times_m2 = m1 * m2
val m1_plus_m2 = m1 + m2
```

Operator Syntax

Wait!! What's this `m1 * m2` stuff?? Shouldn't it be `m1.*(m2)`. It would be really convenient to use "operator syntax", more precisely called *infix operator notation* for many methods like `*` and + here. The Scala parser supports this with a simple relaxation of the rules; when a method takes a single argument, you can omit the period `.` and parentheses `(...)`. Hence the following really is equivalent:

```
val m1_times_m2 = m1.*(m2)
val m1_times_m2 = m1 * m2
```

This convenience can lead to confusing code, especially for beginners to Scala, so use it cautiously.

Traits

Traits are similar to Java 8 *interfaces*, used to define abstractions, but with the ability to provide "default" implementations of the methods declared. Unlike Java 8 interfaces, traits can also have fields representing "state" information about instances. There is a blury line between traits and *abstract classes*, again where some member methods or fields are not defined. In both cases, a subtype of a trait and/or an abstract class must define any undefined members if you want to construct instances of it.

So, why have both traits and abstract classes? It's because Java only allows *single inheritance*; there can be only one *parent* type, which is normally where you would use an abstract class, but Scala lets you "mix in" one or more additional traits (or use a trait as the parent class - yes, confusing). A great example "mix in" trait is one that implements logging. Any "service" type can mix in the logging trait to get "instant" access to this reusable functionality. Schematically, it looks like the following:

```
// Assume severity `Level` and `Logger` types defined elsewhere...
trait Logging {

    def log(level: Level, message: String): Unit = logger.log(level, message)

    private logger: Logger = ...
}

abstract class Service {
    def run(): Unit   // No body, so abstract!
}

class MyService extends Service with Logging {
    def run(): Unit = {
        log(INFO, "Staring MyService...")
        ...
        log(INFO, "Finished MyService")
    }
}
```

`Unit` is Scala's equivalent to Java's `void`. It actually is a true type with a single return value, unlike `void`, but we use it in the same sense of "nothing useful will be returned".

Ranges

What if you want some numbers between a start and end value? Use a Range (http://www.scala-lang.org/api/current/index.html#scala.collection.immutable.Range), which has a nice literal syntax, e.g., `1 until 100`, `2 to 200 by 3`.

The `Range` always includes the lower bound. Using `to` in a `Range` makes it *inclusive* at the upper bound. Using `until` makes it *exclusive* at the upper bound. Use `by` to specify a delta, which defaults to `1`.

```
1 until 10
```

9 entries total

Show: 10

_1	_2
0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9

Showing 9 of 9 records

```
1 to 10
```

10 entries total

Show:
10

_1	_2
0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10

Showing 10 of 10 records

```
1 to 10 by 3
```

4 entries total

Show:
10

_1	_2
0	1
1	4
2	7
3	10

Showing 4 of 4 records

When you need a small test data set to play with Spark, ranges can be convenient.

```
val rdd7 = sc.parallelize(1 to 50).
    map(i => (i, i%7)).
    groupBy{ case (i, seven) => seven }.
    sortByKey()
rdd7.take(7).foreach(println)
```

Took: 1 second 121 milliseconds, at 2017-3-12 12:14

SparkContext (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.SparkContext) also has a `range` method that effectively does the same thing as `sc.parallelize(some_range)`.

Scala Interpreter (REPL) vs. Notebooks vs. Scala Compiler

This notebook has been using a running Scala interpreter, a.k.a. *REPL* ("read, eval, print, loop") to parse the Scala code. The Spark distribution comes with a `spark-shell` script that also lets you use the interpreter from the command line, but without the nice notebook UI.

If you use `spark-shell`, there are a few other behavior changes you should know about.

Using :paste Mode

By default the Scala interpreter treats *each line* you enter separately. This can cause surprises compared to how the Scala *compiler* works, where it treats all the code in the same file in the same context.

For example, the following code, where the expression continues on the second line, is handled successfully by the compiler, but not by the interpreter.

```
(1 to 100)
.map(i => i*i)
```

the Interpreter thinks it finished parsing the expression when it hit the new line after the literal Range (http://www.scala-lang.org/api/current/index.html#scala.collection.immutable.Range), `1 to 100`. It then throws an error on the opening `.` on the next line. On the other hand, the compiler keeps compiling, ignoring the new line in this case.

This notebook also does the same thing as the "raw" interpreter, but in some cases, notebooks will use an interpeter command, `:paste` that tells the parser to parse all of the lines that follow together, just like the compiler would parse them, until the "end of input", which you indicate with `CTRL-D`.

You can't experiment with it through this notebook, but your session would look something like this:

```
scala> :paste
// Entering paste mode (ctrl-D to finish)

(1 to 10)
.map(i => i*i)
<CTRL-D>

// Exiting paste mode, now interpreting.

res0: scala.collection.immutable.IndexedSeq[Int] = Vector(1, 4, 9, 16, 25, 36,
 49, 64, 81, 100)

scala>
```

Ambiguities with Companion Objects

As I wrote this notebook, I *wanted* to demonstrate using the companion object `IIRecord` to define a method explicitly, but this leads to an ambiguity later on in the notebook if you attempt to use this method. The notebook gets confused between the case class and the object.

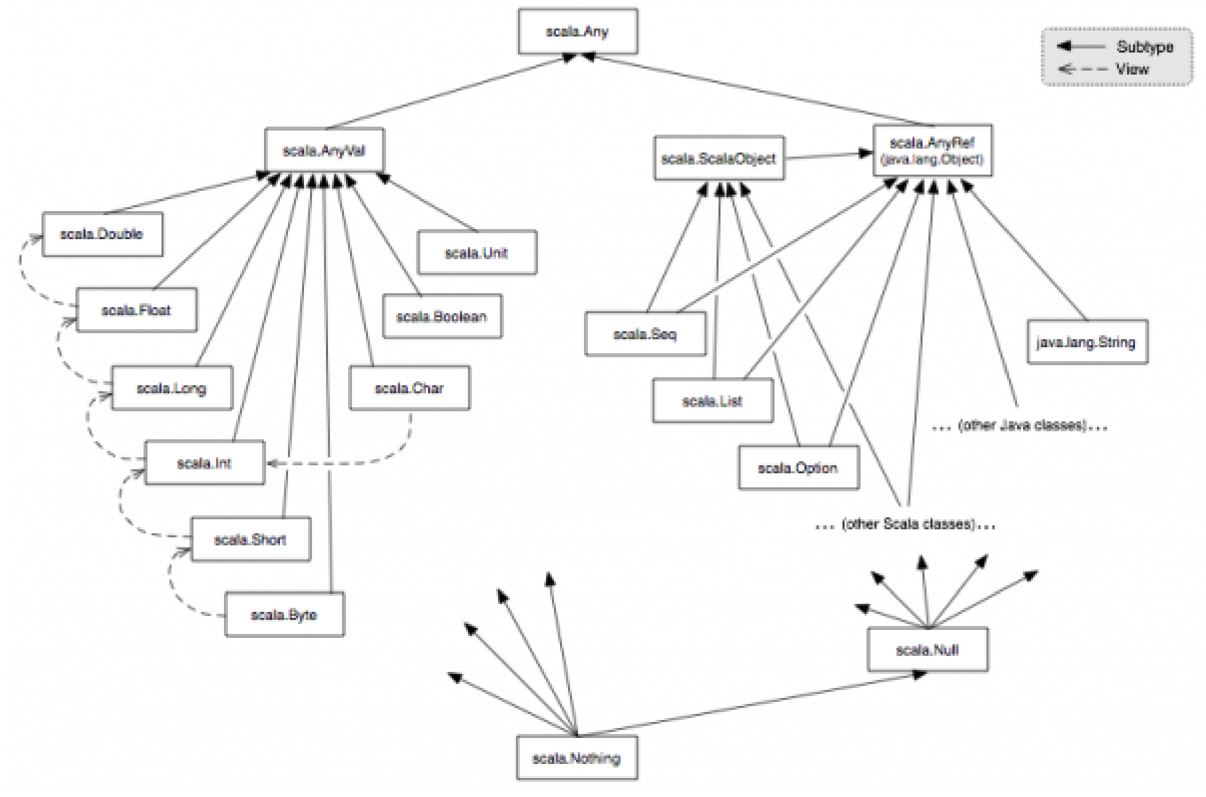
While unfortunate, it's also true that once you start defining more involved case classes, with more than trivial methods and explicit additions to the default companion object, you should really define these types outside the notebook in a compiled library that you use within the notebook.

The details are beyond our scope here, but basically, you set up a project with your Scala code and build it using your favorite build tool. SBT (http://www.scala-sbt.org/) is a popular choice for Scala, but Maven, Gradle, etc. can be used.

You want to generate a *jar* file with the compiled artifacts, then when you start `spark-shell`, submit a Spark job with `spark-submit` or use a notebook environment like this one, you specify the jar for inclusion. For `spark-shell` and `spark-submit`, invoke it with the `--jars myproject.jar` option. For Toree with Jupyter, see the discussion on the FAQ page (https://toree.incubator.apache.org/documentation/user/faq.html).

Scala's Type Hierarchy

Scala's type hierarchy is similar to Java's, but with some interesting differences.



In Java, all *reference types* are descended from java.lang.Object (https://docs.oracle.com/javase/8/docs/api/java/lang/Object.html). The name *reference type* reflects the fact that the instances for all these types are allocated on the *heap* and program variables are references to those heap locations.

The primitives types, `int`, `long`, etc. are not considered part of the type hierarchy and are treated specially. This is in part a performance optimization, as instances of these types fit in CPU registers and the values are pushed onto stack frames. However, they have wrapper or "boxed" types, `Integer`, `Long`, etc., that are part of the type hierarchy, which you must use with Java's collections, for example (with the exception of arrays).

Instead, Scala treats the primitives at the code level as basically the same as the reference types. You don't use `new Int(100)` for example, but you can call methods on `Int` instances. The code generated, in most cases, uses the optimized JVM primitives.

Hence, the Scala type hierarchy defines a type Any (http://www.scala-lang.org/api/current/#scala.Any) to be the a parent type of *both* reference types and "value" types (for the primitives). Each of those subhierarchies have parent types, AnyRef (http://www.scala-lang.org/api/current/#scala.AnyRef) is effectively the same as java.lang.Object (https://docs.oracle.com/javase/8/docs/api/java/lang/Object.html), and AnyVal (http://www.scala-lang.org/api/current/#scala.AnyVal) is the parent of the value types.

Finally, for better "soundness", the Scala type system defines a real type to represent Null (http://www.scala-lang.org/api/current/#scala.Null) and Nothing (http://www.scala-lang.org/api/current/#scala.Nothing). By defining `Null` to be the subtype of all reference types `AnyRefs` (but not `AnyVals`), it supports at the type level the (unfortunate) practice of using `null` for a reference value.

However, `null` is not allowed for an `AnyVal`, so the true "bottom type" of the hierarchy is `Nothing`. Why is that useful. I'll explain in the next section.

Try vs. Option vs. null

Recall the signature of our `curl` method near the beginning of this notebook:

```
def curl(sourceURLString: String, targetDirectoryString: String): File = ...
```

It returns a `File` when everything goes well, but it could throw an exception. An alternative is return a `Try[File]`, where the Try (http://www.scala-lang.org/api/current/index.html#scala.util.Try) encapsulates both cases in the return value, as we'll discuss next. We'll also discuss an alternative, Option (http://www.scala-lang.org/api/current/index.html#scala.Option).

Suppose instead that we declared `curl` to return util.Try[File] (http://www.scala-lang.org/api/current/index.html#scala.util.Try). The only change to the body would be to simply add `Try` before the opening bracket:

```
def curl(sourceURLString: String, targetDirectoryString: String): Try[File] =
 Try {...}
```

Now, the reader knows from the method signature that it might fail somehow. If a call fails, the relevant exception will be returned wrapped in a subclass of `Try`, called util.Failure[File] (http://www.scala-lang.org/api/current/index.html#scala.util.Failure). However, if `curl` succeeds, the `File` will be returned wrapped in the other subclass of `Try`, util.Success[File] (http://www.scala-lang.org/api/current/index.html#scala.util.Success).

Because of Scala's type safety, the caller of `curl` would have to determine if a `Success` or `Failure` was returned and handle it appropriately.

Scala does not have methods declare the exceptions they will throw using the `throws` clause, like Java. So, looking at the signature of our original version, there's no obvious way to know if it throws an exception *or* returns `null` on failure:

```
def curl(sourceURLString: String, targetDirectoryString: String): File = {...}
```

If we choose to catch exceptions internally and return `null`, the caller has to remember to check for `null`. Otherwise, the infamous NullPointerException (https://docs.oracle.com/javase/8/docs/api/java/lang/NullPointerException.html) might happen occasionally if the caller assumes a non-`null` value is returned. So, using `Try[File]` prevents us from this loophole. *It helps the user do the right thing!*

Also, using `Try` rather than simply throwing an exception, means that `curl` always returns "normally", so the caller maintains full control of the call stack and special exception-catching logic isn't required (although you can do that in Scala, if you want to).

What are all the possible valid subclasses of `Try`? Really, there are only two, `Success` and `Failure`. It would be a mistake to allow a user to define other subtypes, like `MaybeCouldFailButWhoKnows`, because users of `Try` in pattern matching will always want to know that there are only two possibilities. Scala adds a keyword to enforce this logical behavior. `Try` is actually declared as follows:

```
sealed abstract class Try[+T] extends AnyRef
```

(`AnyRef` is the same as Java's `Object` supertype.) The `sealed` keyword says that *no* subclasses of `Try` can be declared, *except* in the same source file (which the library author wrote). Hence, users of `Try` can't declare their own subclasses, subverting the logical structure of this type hierarchy and other user's code that relies on this structure.

What if we have a situation where it makes no sense to involve an exception, but we want similar logic to handle the case where I have something or I don't (put another way, I have either 0 or 1 elements)? This is where Option[T] (http://www.scala-lang.org/api/current/index.html#scala.Option) comes in.

`Option` is analogous to `Try`, it is a `sealed` abstract type with two possible subtypes:

- Some[T] (http://www.scala-lang.org/api/current/index.html#scala.None): I have a an instance of `T` for you, inside the `Some[T]`.
- None (http://www.scala-lang.org/api/current/index.html#scala.None): I don't have a value for you, sorry.

Note that a hash map is a great example where I either have a value for a given key or I don't. Therefore, for Scala's Map[K,V] (http://www.scala-lang.org/api/current/index.html#scala.collection.Map) abstraction, where `K` is the key type and `V` is the value type, the `get` method has this signature:

```
def get(key: K): Option[V]
```

Once again, you know from the type signature that you may or may not get a value instance for the input key, *and* you **must** determine whether you got a `Some[V]` or a `None` as the result. By never returning a `null` value, we remove the risk of a `NullPointerException` if the caller forgets to check for it!

So, how do we determine which `Option[T]` was returned? Let's look a few examples using `Option`. Can you guess what they are doing? Check the Option Scaladocs (http://www.scala-lang.org/api/current/#scala.Option) to confirm. `Try` can be used similarly, with a few other ways available that we won't discuss here (but see the Try Scaladocs (http://www.scala-lang.org/api/current/#scala.util.Try)).

```
val options = Seq(None, Some(2), Some(3), None, Some(5))

val numbers1 = options.map { o =>
    o.getOrElse(-1)
}
```

Took: 588 milliseconds, at 2017-3-12 12:15

```
val numbers2 = options.map {
    case None    => -1
    case Some(i) => i  // Note how we extract the enclosed value.
}
```

Took: 566 milliseconds, at 2017-3-12 12:15

If you just want to ignore the `None` values, use a *for comprehension*. We could print them as before, but this time we'll `yield` each value, constructing a new `Seq` (sequence) of numbers.

```
val numbers = for {
    opt   <- options  // loop through the options, assign each to "option"
    value <- opt      // extract the value from the Some, or if None, skip to the next
} yield value
```

Took: 603 milliseconds, at 2017-3-12 12:15

Finally, you might wonder how `None` is declared. Consider this example:

```
val opts: Seq[Option[String]] = Seq(Some("hello"), None, Some("world!"))
opts.foreach(println)
```

Took: 776 milliseconds, at 2017-3-12 12:15

This works, so it must mean that `None` is a valid subclass of `Option[String]`. That's actually true for all `Option[T]`. How can a single object be a valid subtype for *all* of them? Here is how it's declared (omitting some details):

```
object None extends Option[Nothing] {...}
```

`None` carries no "state" information, because it doesn't wrap an instance like `Some[T]` does. Hence, we only need one instance for all uses, so it's declared as an object. Recall we mentioned above that the type system has a Nothing (http://www.scala-lang.org/api/current/#scala.Nothing) type, which is a subtype of all other types. Without diving into too many details, if a variable is of type `Option[String]`, then you can use an `Option[Nothing]` for it (i.e., the latter is a subtype of the former). This is why `Nothing` is useful, for cases like `None`, so we can have one instance of it, but still obey the rules of Scala's object-oriented type system.

Implicits

Scala has a powerful mechanism known as *implicits* that is used in the Spark Scala API. Implicits are a big topic, so we'll focus just on the uses of it that are most important to understand.

Type Conversions

We used `RDD` methods like `reduceByKey` above, but if you search for this method in the RDD Scaladoc page (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.rdd.RDD), you won't find it. Instead it's defined in the PairRDDFunctions (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.rdd.PairRDDFunctions) type (along with all the other `*ByKey` methods). So, how can we use these methods as if they are defined for `RDD`??

When the Scala compiler sees code calling a method that doesn't exist on the type, it looks for an *implicit conversion* in the current scope, which can transform the instance into another type (i.e., by wrapping it), where the other type provides the needed method. The full signature inferred for the method as it's used must match the definition in the wrapping class.

> **Note:** If you don't find a method in the Spark Scaladocs (http://spark.apache.org/docs/latest/api/scala/index.html#package) for a type where you think it should be defined, look for related helper types with the method.

Here's a small Scala example of how this works:

```
// A sample class. Note it doesn't define a `toJSON` method:
case class Person(name: String, age: Int = 0)
```

Took: 577 milliseconds, at 2017-3-12 12:15

```
// To scope them, define implicit conversions within an object
object implicits {

    // `implicit` keyword tells the compiler to consider this conversion.
    // It takes a `Person`, returning a new instance of `PersonToJSONString`,
    // then resolves the invocation of `toJSON`.
    implicit class PersonToJSONString(person: Person) {
        def toJSON: String = s"""{"name": ${person.name}, "age": ${person.age}}"""
    }
}

import implicits._        // Now it is visible in the current scope.

val p = Person("Dean Wampler", 39)

// Magic conversion to `PersonToJSONString`, then `toJSON` is called.
p.toJSON
```

Took: 1 second 208 milliseconds, at 2017-3-12 12:15

For `RDDs`, the implicit conversions to `PairRDDFunctions` and other support types are handled for you. Similarly, some conversions are used by Spark SQL.

```
val wtc = iiDF.select($"word", $"total_count")
wtc.show
```

Took: 1 second 5 milliseconds, at 2017-3-12 12:15

The column-reference syntax `$"name"` is implemented using the same mechanism in the Scala library that implements interpolated strings, `s"$foo"`. The `import sqlc.implicits._` makes it available.

Note we imported something from an *instance*, rather than a package or type, as allowed in Java. This can be a useful feature in Scala, but it's also fragile, If you try `import sqlContext.implicits._`, you'll get a compiler error that a "stable identifier" is required. It turns out that doing the value assignment, `val sqlc = sqlContext` first meets this requirement. This is unique to the notebook environment. You normally won't see this problem if you use the `spark-shell` that comes with a Spark distribution or you write a Spark program and compile it with the Scala compiler.

However, it would be better if Spark defined this `implicits` object on the `SQLContext` companion object instead of on instances of it!

For completeness, but unrelated to implicits, the `DataFrame` API lets you write SQL-like queries with a programmatic API. If you want to use built in functions like `min`, `max`, etc. on columns, you need the following `import` statement:

```
import org.apache.spark.sql.functions._
```

Took: 460 milliseconds, at 2017-3-12 12:17

Now we can use `min`, `max`, `avg`, etc.

```
val mma = iiDF.select(min("total_count"), max("total_count"), avg("total_count"))
mma.show
```

Took: 1 second 480 milliseconds, at 2017-3-12 12:17

Implicit Method Arguments

One other use of implicits worth understanding is *implicit arguments* to methods. You will encounter this mechanism used when you read the Spark Scaladocs, even though you might never realize you're actually using it in your code!

Recall I mentioned previously that you can define default values for method arguments. I just used it for the `age` argument for `Person`:

```
case class Person(name: String, age: Int = 0)
```

Sometimes we need something more sophisticated. For example, our library might have a group of methods that need a special argument passed to them that provides useful "context" information, but you don't want the user to be required to explicitly pass this argument every time. Other times you might use implicit arguments to make the API "cleaner", but still have some control over what's allowed.

Here's an example, that's partly inspired by Scala's Seq.sum (http://www.scala-lang.org/api/current/#scala.collection.Seq) method. Wouldn't it be great if I happen to have a collection of things I can "add" together, if I could just call `sum` on the collection? Let's do this in a slightly different way, with a helper `sum` method outside of `Seq`.

```
trait Add[T] {
    def add(t1: T, t2: T): T
}

// Nested implicits so they don't conflict with the previous object implicits.
object Adder {
    object implicits {
        implicit val intAdd = new Add[Int] {
            def add(i1: Int, i2: Int): Int = i1+i2
        }
        implicit val doubleAdd = new Add[Double] {
            def add(d1: Double, d2: Double): Double = d1+d2
        }
        implicit val stringAdd = new Add[String] {
            def add(s1: String, s2: String): String = s1+s2
        }
        // etc...
    }
}

import Adder.implicits._

// NOTE: TWO argument lists!
def sum[T](ts: Seq[T])(implicit adder: Add[T]): T = {
    ts.reduceLeft((t1, t2) => adder.add(t1, t2))
}
```

Took: 1 second 455 milliseconds, at 2017-3-12 12:18

```
sum(0 to 10)
```

```
55
```

Took: 1 second 40 milliseconds, at 2017-3-12 12:18

```
sum(0.0 to 5.5 by 0.3)
```

```
51.29999999999999
```

Took: 880 milliseconds, at 2017-3-12 12:18

```
sum(Seq("one", "two", "three"))
```

```
onetwothree
```

Took: 757 milliseconds, at 2017-3-12 12:18

```
// Will fail, because there's no Add[Char] in scope:
sum(Seq('a', 'b', 'c'))   // Characters
```

So, the implicit values `intAdd`, `doubleAdd`, and `stringAdd`, were used by the Scala interpreter for the `adder` argument in the second *argument list* for `sum`. Note that you have to use a second argument list and all arguments there must be implicit.

We could have avoided using implicit arguments if we defined custom `sum` methods for every type. That would have been simpler in this trivial case, but for nontrivial methods, the duplication is worth avoiding. Another advantage of this mechanism is that the user can define her own implicit `Add[T]` instances for domain types (say for example, `Money`) and they would "just work".

The Scala collections API uses this mechanism to know how to construct a new collection of the same kind as the input collection when you use `map`, `flatMap`, `reduceLeft`, etc.

Spark uses this pattern for Encoders (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.Encoder) in Spark SQL. Encoders are used to serialize values into the new, compact memory encoding introduced in the *Tungsten* project (see for example, here (https://spark-summit.org/2015/events/deep-dive-into-project-tungsten-bringing-spark-closer-to-bare-metal/)). Here's an example of creating a Dataset (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.Dataset), where the `toDS` method is first "added" to a Scala Seq (http://www.scala-lang.org/api/current/#scala.collection.Seq) through an implicit conversion (specifically SQLImplicits.localSeqToDatasetHolder (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.SQLImplicits), which is brought into scope by the `import sqlc.implicits._` statement earlier) and then `toDS` uses `Encoders` internally.

```
(0 to 10).toDS()
```

```
1                         Took: 1 second 238 milliseconds, at 2017-3-12 12:18
value
```

Conclusions

I appreciate the effort you put into studying this notebook. I hope you enjoyed it as much as I enjoyed writing it. Please post issues on how I can improve it to the GitHub repo (https://github.com/deanwampler/JustEnoughScalaForSpark).

Now you know the core elements of Scala that you need for using the Spark Scala API. I hope you can appreciate the power and elegance of Scala. I hope you will choose to use it for all of your data engineering tasks, not just for Spark.

What about data science? There are many people who use Scala for data science in Spark, but today Python and R have much richer libraries for Mathematics and Machine Learning. That will change over time, but for now, you'll need to decide which language best fits your needs.

As you use Scala, there will be more things you'll want to understand that we haven't covered, including common idioms, conventions, and tools used in the Scala community. The references at the beginning of the notebook will give you the information you need.

Best wishes.

Dean Wampler (mailto:deanwampler@gmail.com)
@deanwampler (http://twitter.com/deanwampler)

Appendix: Exercise Solutions

Let's discuss the solutions to exercises that weren't already solved earlier in the notebook.

Filter for Plays that Have "of" in the Name

You can add the condition (comment `// <== here`) immediate after defining `play`. You could do it later, after either of the subsequent two expressions, but then you're doing needless computation. Change `true` to `false` to print plays that don't contain "of".

```
val list2 = for {
    play <- plays
    if (play.contains("of") == true)                            // <== here
    playFileString = targetDirName + pathSeparator + play
    playFile = new File(playFileString)
} yield {
    val successString = if (playFile.exists) "Success!" else "NOT FOUND!!"
    s"$playFileString\t$successString"
}
list2.foreach(println)
```

Took: 1 second 44 milliseconds, at 2017-3-12 12:19

More Queries

```
val topLocationsLoveHate = sparkSession.sql("""
    SELECT word, total_count, locations[0] AS top_location, counts[0] AS top_count
    FROM inverted_index
    WHERE word LIKE 'love%' OR word LIKE 'unlove%' OR word LIKE 'hate%'
""")
topLocationsLoveHate.show(40)
```

Took: 1 second 281 milliseconds, at 2017-3-12 12:20

Return the Top Two Locations and Counts

We used the `DataFrame` API to write a SQL query that returned the top location and count. Adding the next one is straightforward. What do you observe is returned when there isn't a second location and count?

```
val topTwoLocations = sparkSession.sql("""
    SELECT word, total_count,
        locations[0] AS first_location,  counts[0] AS first_count,
        locations[1] AS second_location, counts[1] AS second_count
    FROM inverted_index
    WHERE word LIKE '%love%' OR word LIKE '%hate%'
""")
topTwoLocations.show(100)
```

Took: 921 milliseconds, at 2017-3-12 12:20

Removing Stop Words

Recall you were asked to implement a `keep(word: String):Boolean` method that filters stop words.

First, let's implement `keep`. You can find lists of stop words on the web. One such list for English can be found here. It includes many words that you might not consider stop words. Nevertheless, I'll just use a smaller list here.

Note that I'll use a Scala Set (http://www.scala-lang.org/api/current/index.html#scala.collection.immutable.Set) to hold the stop words. We want *O(1)* look-up performance. We just want to know if the word is in the set or not.

I'll also add `""`, so I can remove the explicit test for it.

Finally, we'll embed the whole thing in a new Scala `object`. This extra encapsulation is a way to work around occasional problems with "task not serializable" errors.

```
object StopWords {
  val stopWords = sc.broadcast(Set("", "a", "an", "and", "I", "he", "she", "it", "the")

  def keep(word: String): Boolean = {
    word.size > 0 && (stopWords.value.contains(word) == false)
  }

  def iiStop(sc: SparkContext) = sc.wholeTextFiles(shakespeare.toString).
        flatMap {
            case (location, contents) =>
                val words = contents.split("""\W+""").
                    map(word => word.toLowerCase).  // Do this early, before keep(), be
                    filter(word => keep(word))      // <== filter here
                val fileName = location.split(java.io.File.separator).last
                words.map(word => ((word, fileName), 1))
        }.
        reduceByKey((count1, count2) => count1 + count2).
        map {
            case ((word, fileName), count) => (word, (fileName, count))
        }.
        groupByKey.
        sortByKey(ascending = true).
        map {
            case (word, iterable) =>
                val vect = iterable.toVector.sortBy {
                    case (fileName, count) => (-count, fileName)
                }
                val (locations, counts) = vect.unzip
                val totalCount = counts.reduceLeft((n1,n2) => n1+n2)
                (word, totalCount, locations, counts)
        }
}
```

Took: 982 milliseconds, at 2017-3-12 12:20

NOTE: If the following cell *still* throws a "Task not serializable" exception, it's a due to an ambiguity in the Scala interpreter. The code will work fine in a standalone Spark program and may work better in a shorter notebook!

```
StopWords.iiStop(sc).take(100).foreach(println)
```

One last thing, we now have `filter(word => keep(word))`, but note how we used `println` in the previous cell to see results. We can do something similar with `filter` and instead write `filter(keep)`.

What does this mean exactly? It tells the compiler "convert the *method* `keep` to a *function* and pass that to `filter`." This works because `keep` already does what `filter` wants, take a single string argument and return a boolean result.

Passing `keep` is actually different than passing `word => keep(word)`, which is an *anonymous* function that *calls* keep. We are using `keep` as the function itself, rather than constructing a function that uses `keep`.

Build: | **buildTime**-*Wed Nov 23 12:19:16 UTC 2016* | **formattedShaVersion**-*0.7.0-c955e71d0204599035f603109527e679aa3bd570* | **sbtVersion**-*0.13.8* | **scalaVersion**-*2.11.8* | **sparkNotebookVersion**-*0.7.0* | **hadoopVersion**-*2.7.2* | **jets3tVersion**-*0.7.1* | **jlineDef**-*(jline,2.12)* | **sparkVersion**-*2.0.2* | **withHive**-*true* |.