**Announcement**

SciCom version 0.2.2 has been released. SciCom (Scientific Computing) for Ruby brings the power of R to the Ruby community. SciCom is based on Renjin, a JVM-based interpreter for the R language for statistical computing.

**R on the JVM**

Over the past two decades, the R language for statistical computing has emerged as the de facto standard for analysts, statisticians, and scientists. Today, a wide range of enterprises – from pharmaceuticals to insurance – depend on R for key business uses. Renjin is a new implementation of the R language and environment for the Java Virtual Machine (JVM), whose goal is to enable transparent analysis of big data sets and seamless integration with other enterprise systems such as databases and application servers.

Renjin is still under development, but it is already being used in production for a number of client projects, and supports most CRAN packages, including some with C/Fortran dependencies.

**SciCom and Renjin**

SciCom integrates with Renjin and allows the use of R inside a Ruby script. In a sense, SciCom is similar to other solutions such as RinRuby, Rpy2, PipeR, etc. However, since SciCom and Renjin both target the JVM there is no need to integrate both solutions and there is no need to send data between Ruby and R, as it all resides in the same JVM. Further, installation of SciCom does not require the installation of GNU R; Renjin is the interpreter and comes with SciCom. Finally, although SciCom provides a basic interface to Renjin similar to RinRuby, a much tighter integration is also possible (see examples below).

**SciCom with Standard R Interface**

SciCom allows R programmers to use R commands inside a Ruby script in a way similar to RinRuby by calling method eval and passing to it an R script:



Programmers can also use here docs to integrate an R script inside a Ruby script. The next example show a model for predicting baseball wins based on runs allowed and runs scored. The data comes from Baseball-Reference.com.



The output of the program above is:



**The SciCom “language”**

SciCom also allows for implementing R scripts in a “language” that is just like Ruby, so that the developer does not need to know that she is actually writing an R script. All R methods are accessible through an R namespace.

The next script is the same baseball model done in R above using SciCom ‘language’:



We show bellow an example of calculating the correlation matrix without using the build-in functions. First this is done in an R script and then using SciCom:

Now the same code using SciCom



As another example, here is a SciCom script to print the number of days for every month is 2005:



As can be seen from these examples, R methods can be accessed through the R namespace in SciCom, so, R method ‘seq’ is called in SciCom as ‘R.seq’. R methods that are applied on objects can be called in two ways, either using the R namespace as in ‘R.factor’ or directly on the object, as in this case we did ‘cmonth.factor’.

This last example shows how SciCom allows method chaining, which is not possible in an R script.

**What´s New**

This version integrates SciCom with MDArray. MDArray is a multi dimensional array implemented for JRuby inspired by NumPy ([www.numpy.org](http://www.numpy.org)) and Masahiro Tanaka´s Narray (narray.rubyforge.org). MDArray stands on the shoulders of Java-NetCDF and Parallel Colt. At this point MDArray has libraries for linear algebra, mathematical, trigonometric and descriptive statistics methods.

NetCDF-Java Library is a Java interface to [NetCDF files](http://www.unidata.ucar.edu/software/netcdf/index.html), as well as to many other types of scientific data formats.  It is developed and distributed by Unidata (<http://www.unidata.ucar.edu>).

Parallel Colt (<https://sites.google.com/site/piotrwendykier/software/parallelcolt>) is a [multithreaded](http://en.wikipedia.org/wiki/Thread_%28computer_science%29) version of [Colt](http://dsd.lbl.gov/~hoschek/colt/) (<http://acs.lbl.gov/software/colt/>). Colt provides a set of Open Source Libraries for High Performance Scientific and Technical Computing in Java. Scientific and technical computing is characterized by demanding problem sizes and a need for high performance at reasonably small memory footprint.

**Converting MDArray to R Array (same backing store)**

An MDArray can be converted to an R array by calling method ‘R.md’.

First, let´s create an MDArray of shape [4, 3]:



This is arr1 as printed from MDArray:



Now, converting this array to an R array and printing it:



The result is:



One very important aspect of this conversion is that both the MDArray and the R array use the same backing store, and thus, this conversion does not do any copying and has very low cost. However, WITH GREAT POWER COMES GREAT RESPONSABILITIES: since MDArray and the R array have the same backing store, a change in MDArray will also change the value of the R array. Renjin assumes that the vector will never change and delays calculation of the vector to the latest possible time. If values change, the result can be unexpected, so, any changes to an MDArray should be done with care.

**Array indexing**

MDArrays are indexed starting at 0, while R arrays are indexed starting at 1. In order to facilitate the use of converted MDArrays we introduced method ‘ri’ (r-indexing) that converts an MDArray index into an R matrix index.

Comparing the content of the MDArray and R array defined above can be done with:



* We first create a byte MDArray. Byte arrays are converted to logical vectors in R;
* arr1.get\_index retrieves all indexes from arr1 in order;
* we then compare arr1[\*ct] (the array given its index) with r\_matrix.ri(\*ct) (.ri converts the given index to an R index)
* In R, indexing a vector returns a new vector. If we want to get a scalar and not a vector, SciCon provides method .gz.
* Finally, comp is converted to an logical vector in R and we call method all on this vector. Method all returns true if all elements of the vector are true. In this case, all elements are true and comp.all.gt print true.

**Multi-dimensional arrays**

Multi-dimensional arrays can also be converted into R arrays using method ‘.md’. However, multi-dimension definition for MDArray and R arrays are different. For instance, an MDArray defined with the following dimensions [3, 2, 2] indicates that there are 3 vector of 2 x 2 dimensions.

The figure bellow shows a [3, 2, 2] array in MDArray.

****

Bellow we show a [3, 2, 2] array created in R. In R this specification indicates that the user wants to build an array of 2 vectors with size [3, 2].



In order to allow for easy use of converted arrays, when multi-dimensional arrays are converted from MDArray to R array the R array is dimensioned in order to be identical to the MDArray. As such, if the MDArray above is converted to an R array, the R array dimension is [2, 2, 3].

**Dicing and Slicing MDArrays**

MDArrays can be sliced and diced in many ways. A slilced MDArray can be converted to R array as any other MDArray. From the point of view of R, this is just a normal array.

When working with two dimensional arrays, each line is viewed as a new record and there is no information encoded in the line number. Columns encode information and each column has a different type of value, for example, “name”, “age”, “phone number”, etc.

With multi-dimensional arrays, dimensions can encode information. For example, let´s suppose we are developing a system to analyze quotes from multiple stocks. Working with two dimensional arrays we would have a file for each stock, in which each row would be a new record and columns would represent, “open”, “high”, “low”, “close”, etc. In multi-dimensional arrays we can use a single array and the following dimensions:

* Dimension 0: The date of the quote.
* Dimension 1: The stock
* Dimension 2: The quote characteristic (“open”, “high”, etc.)

Let´s encode all quotes from Jul. 2014 for the following stocks: Google, Microsoft, Yahoo and Apple. We define an MDArray with the following specification: [22, 4, 6]. The first dimension of size 22 represents the 22 business days of Jul. 2014. The second dimension of size 4 is for each of the four stocks, and dimension 3 of size 6 has the quote attributes “open”, “high”, “low”, “close”, “volume” and “adjusted volume”.

Getting the data from Yahoo finance, we have that the opening value of Google stock on 1/Jul/2014 was 578.32. So, we assign data[0, 0, 0] = 578,32. The opening value of Google stock on 2/Jul/2014 was 583.35. So, again we have data[1, 0, 0] = 583.35.

Now, Microsoft “high” stock value on Jul/03/2014 was 44.09, so data[2, 1, 1] = 44.09.

Let´s say that we want the have statistics about the opening price of Google stocks. We can slice the data array to create a view with only the values of interest:



The ‘section’ method gets a section of the original array. It takes two or three arguments. The first two arguments are arrays and the third in ‘true’ (when used). The first array is an array of indexes and the second is an array of sizes. So, looking at the first dimension, we start at index 0 and get 22 elements (all elements in that dimension), in this example, all dates on Jul. 2014. The second dimension gets stock 0 and size 1, i.e., only 1 stock is selected. In this example Google is indexed by 0. Finally, the third dimension is from index 0 (“open”) and of size 1, i.e., only the open attribute is selected. Printing sec gives:



Now, let´s convert this to R and call the summary function, by:



The result is:



**SciCom main properties are**

* Allows access to R scripts from inside Ruby scripts;
* Allows for R scripts written in R by accessing method ‘R.eval’;
* Allows R scripts to be embedded inside here docs in Ruby;
* Creates a new ‘language’ that allows regular Ruby scripts to call R methods in such a way that programmers can be unaware of the fact that they are using R (although, of course, knowing R is of great benefit).;
* Integrates with MDArray allowing multi-dimensional arrays to be slice and cut and passed to an R script.

**SciCom installation and download**

* Install Jruby
* jruby –S gem install scicom

**SciCom Homepages**

* <http://rubygems.org/gems/scicom>
* <https://github.com/rbotafogo/scicom/wiki>

**Contributors**

Contributors are welcome.

**How SciCom Works**

Almost everything is dynamically dispatched to Renjin.  The magic is basically done by Ruby method 'method\_missing'.  I´ll describe a bit how the whole thing works.

Let´s begin with a simple example:

> vec = R.c(1, 2.45, 3)

Ruby will try to find method 'c' on R class, since there is no such method, Ruby calls 'method\_missing' passing the argument list to it.  'method\_missing' calls 'parse' to parse the arguments.  In this case, every argument is an Numeric and 'parse' will basically convert the argument list to a string: "(1, 2.45, 3)".  'method\_missing' then makes the following call:

# R.eval(" <method\_name> + <arguments> "), which in this case is:

> R.eval("c(1, 2.45, 3)")  giving us the desired result.

Now let´s do:

> vec2 = R.c(vec, 4, 5)

Again, 'method\_missing' is called, which calls 'parse'.  The first argument to parse is now a Ruby::Vector and parse does the following:

# create a temporary variable, lets call it sc\_1234

# let vec.sexp be the actual java sexp for this vector, then:

> R.eval("sc\_1234 = vec.sexp")  # this stores the java vector in the sc\_1234 variable

# finally return all parameters as a string: "(sc\_1234, 4, 5)"

# Now let Renjin do the work:

> R.eval("c(sc\_1234, 4, 5)")

# return of the above is the vector [1, 2.5, 3, 4, 5] as expected

# finally, we remove the temporary variable sc\_1234, from Renjin name space in the hope that # this will allow gc to work.

Every statement in SciCom is converted to an evaluable string that can be passed to Renjin 'eval'.  Now, this adds some overhead, but unless we put a SciCom statment inside a large loop this should be reasonable.  One way that integration could be improved is if there was some way of calling 'eval' on a SEXP and not on a string.  Maybe you have this already, but I couldn´t find a way of doing it.

Now, if we have

> vec.mean

Then again we call 'method\_missing' of the Vector class.  In this case, first we check if 'mean' is a named element of the vector.  If it is, then we call:

> R.eval("vec[mean]")

If mean is not a named element of the vector, then we call:

R.eval("mean(vec)")

Which calculates the mean of the vector.  Again, this is just translating Ruby statements onto R evaluable strings.

Some new methods had to be defined.  For instance we have method '\_' in order to simulate methods such as %in%.  So here we do

> vec.\_ :in, vec2

This is a call to the '\_' method with 2 arguments ':in' and 'vec2'.  This translates to the correct R call:

> R.eval("vec %in% vec2")

In place of ':in' we can put any other string.

Methods and variable that have a '.' on their names are used in SciCom with '\_\_' notation.  So, read.csv in R is written as read\_\_csv.  'method\_missing' converts '\_\_' to '.' before calling eval.  So, if there is any variable in R written with '\_\_' there will be no way to access it, other than calling R.eval directly.

'method\_missing' is really the great magic!

**SciCom History:**

* 19/Nov/2014: Another small bug fix
* 16/Nov/2014: Version 0.2.1 – Small bug fix
* 16/Nov/2014: Version 0.2.0 – Initial release