Scraping, Reading, and Creating JSON Data with R

Samuel L. Ventura

Department of Statistics, Carnegie Mellon University

March 18, 2015

This Talk Will

1. Overview simple techniques for scraping JSON data in R
2. Overview simple techniques for creating JSON data in R
3. Use live demonstrations!

Why Do We Need to Scrape?

1. Data is good
2. More data is better
3. Large datasets are best
4. Many large datasets live on the web

What is JSON?

JavaScript = "an object-oriented programming language commonly used to create **interactive effects within web browsers**"

JSON = JavaScript Object Notation

"Lightweight data-interchange format"

"It is easy for humans to read and write"

"It is easy for machines to parse and generate"

What does JSON Data Look Like?


```
"Transactions":["SortOrder":"1","Date":"2005-07-30T00:00:0
by the Pittsburgh Penguins in the 1st round (1st overall)
in 2005."],"RosterMoves":"SortOrder":"1","Date":"2015-03-15
1 game (illness).","SortOrder":"2","Date":"2015-03-14T00:0
day-to-day.","SortOrder":"3","Date":"2015-01-28T00:00:00-05
1 game (lower body injury).","SortOrder":"4","Date":"2015-0
body injury, day-to-day.","SortOrder":"5","Date":"2014-12-1
3 games (mumps).","SortOrder":"6","Date":"2014-12-12T00:00
day-to-day.","SortOrder":"7","Date":"2014-04-09T00:00:00-04
1 game (upper body injury).","SortOrder":"8","Date":"2014-0
body injury, day-to-day.","SortOrder":"9","Date":"2013-05-0
the last 12 regular season and first playoff games (oral
surgery).","SortOrder":"10","Date":"2013-04-02T00:00:00-04
surgery, sidelined indefinitely"
```

What does JSON Data Look Like?


```
"SortOrder":"2",
"Date":"2015-03-14T00:00:00-04:00",
"Desc":"Undisclosed, day-to-day.",

"SortOrder":"3",
"Date":"2015-01-28T00:00:00-05:00",
"Desc":"Missed 1 game (lower body injury).",

"SortOrder":"4",
"Date":"2015-01-27T00:00:00-05:00",
"Desc":"Lower body injury, day-to-day.",

"SortOrder":"5",
"Date":"2014-12-18T00:00:00-05:00",
"Desc":"Missed 3 games (mumps)."
```

How to Scrape – Before JSON

1. Download the HTML code

   `try(readLines(URLencode(link)))`

2. Parse/format the HTML code to extract the data

   `strsplit()`, `grep()`, `gsub()`, `paste()`

3. Organize the parsed data into a database/table

   `rbind()`, `cbind()`, `c()`

4. Or, use functions in the `XML` package

   `readHTMLTable()`

But I Can't Find JSON Data!

"When I scraped HTML data, it was easy to find." – everyone

"JSON data is hard to find!" – You, until now

Demo time

Relevant JSON Packages and Code

Package:
`RJSONIO` (there are others, but this will do)

Get JSON Data:
```
json.to.R <- fromJSON(getURL(url.of.website))
```

Create JSON Data:
```
R.to.json <- toJSON(my.data.frame)
```

Other stuff in package:
`readJSONStream()`, `isValidJSON()`, more

Sample Code

```
# Sam Ventura
# 18 March 2015
# JSON / R


# install and load necessary packages

#install.packages("RJSONIO")
#install.packages("RCurl")
library(RJSONIO)
library(RCurl)
```

Sample Code

```
# Get the JSON data from a URL

my.url <- "http://stats.tsn.ca/HGET/urn:tsn:nhl:team:
pittsburgh-penguins/roster?type=json"
json <- fromJSON(getURL(my.url))


# Explore the data

class(json)
length(json)
names(json)
length(json$Players[[1]])
json$Players[[1]]
```

Sample Code

```
# Convert the data you want into a data.frame

my.dtf <- data.frame(do.call(rbind, json$Players))
#my.dtf <- data.frame(t(sapply(json$Players, c)))
head(my.dtf)


# Create a JSON object from a data.frame, save it to a
text file

r.to.json <- toJSON(my.dtf)
write.table(r.to.json, file = "Desktop/r.to.json.txt")
```

Thanks!

For more information:

- sventura@stat.cmu.edu
- www.stat.cmu.edu/∼sventura