

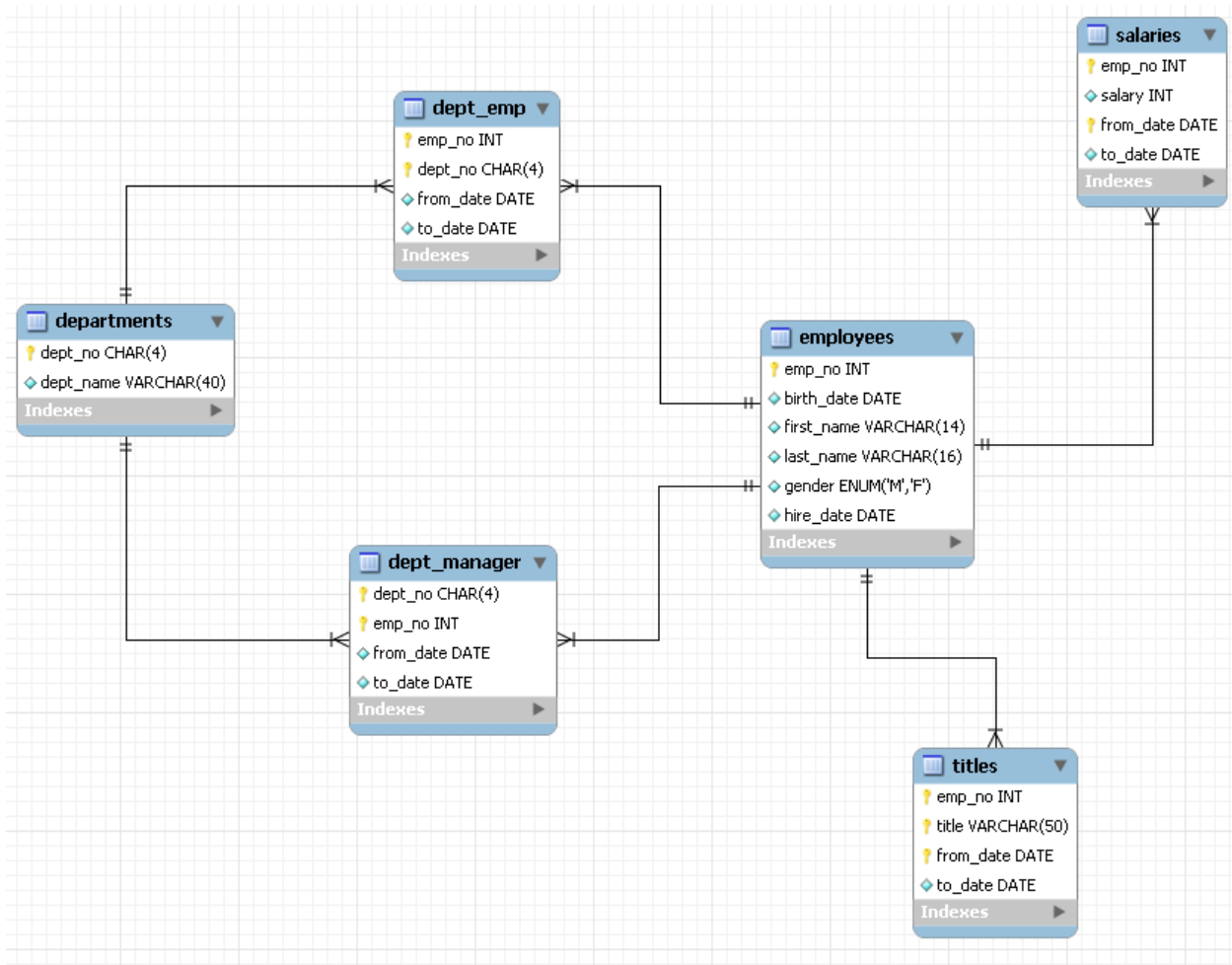
semana_2

Luis Ambrocio

/tableofcontents

Leyendo MySQL

- Software de base de datos de código abierto gratuito y ampliamente utilizado
- Ampliamente utilizado en aplicaciones basadas en Internet.
- Los datos están estructurados en
 - Bases de datos
 - Tablas dentro de bases de datos
 - Campos dentro de tablas
- Cada fila se llama registro



primero se instala MySQL

<http://dev.mysql.com/doc/refman/5.7/en/installing.html>

luego instalar RMySQL

- Official instructions - <http://biostat.mc.vanderbilt.edu/wiki/Main/RMySQL> (may be useful for Mac/UNIX users as well)
- Potentially useful guide - <http://www.ahschulz.de/2013/07/23/installing-rmysql-under-windows/>

conectar y enumerar bases de datos

```
ucscDb <- dbConnect(MySQL(),user="genome",  
                    host="genome-mysql.cse.ucsc.edu")  
result <- dbGetQuery(ucscDb,"show databases;"); dbDisconnect(ucscDb);
```

```
[1] TRUE
```

```
result
```

```
      Database  
1  information_schema  
2      ailMel1  
3      allMis1  
4      anoCar1  
5      anoCar2  
6      anoGam1  
7      apiMel1  
8      apiMel2
```

Conexión a hg19 y listas de tablas

```
hg19 <- dbConnect(MySQL(),user="genome", db="hg19",
                  host="genome-mysql.cse.ucsc.edu")
allTables <- dbListTables(hg19)
length(allTables)
```

```
[1] 10949
```

```
allTables[1:5]
```

```
[1] "HInv"          "HInvGeneMrna" "acembly"      "acemblyClass" "acemblyPep"
```

obteniendo dimensiones de una tabla especifica

```
dbListFields(hg19,"affyU133Plus2")
```

```
[1] "bin"          "matches"      "misMatches"   "repMatches"   "nCount"       "qNumInsert"
[7] "qBaseInsert"  "tNumInsert"   "tBaseInsert"  "strand"       "qName"        "qSize"
[13] "qStart"       "qEnd"         "tName"        "tSize"        "tStart"       "tEnd"
[19] "blockCount"   "blockSizes"   "qStarts"      "tStarts"
```

```
dbGetQuery(hg19, "select count(*) from affyU133Plus2")
```

```
count(*)
1      58463
```

leyendo de la tabla

```
affyData <- dbReadTable(hg19, "affyU133Plus2")
head(affyData)
```

	bin	matches	misMatches	repMatches	nCount	qNumInsert	qBaseInsert	tNumInsert	tBaseInsert	strand
1	585	530	4	0	23	3	41	3	898	-
2	585	3355	17	0	109	9	67	9	11621	-
3	585	4156	14	0	83	16	18	2	93	-
4	585	4667	9	0	68	21	42	3	5743	-
5	585	5180	14	0	167	10	38	1	29	-
6	585	468	5	0	14	0	0	0	0	-

	qName	qSize	qStart	qEnd	tName	tSize	tStart	tEnd	blockCount
1	225995_x_at	637	5	603	chr1	249250621	14361	15816	5
2	225035_x_at	3635	0	3548	chr1	249250621	14381	29483	17
3	226340_x_at	4318	3	4274	chr1	249250621	14399	18745	18
4	1557034_s_at	4834	48	4834	chr1	249250621	14406	24893	23
5	231811_at	5399	0	5399	chr1	249250621	19688	25078	11
6	236841_at	487	0	487	chr1	249250621	27542	28029	1

	blockSizes
1	93,144,229,70,21,

seleccionando un subconjunto especifico

```
query <- dbSendQuery(hg19, "select * from affyU133Plus2 where misMatches between 1 and 3")
affyMis <- fetch(query); quantile(affyMis$misMatches)
```

0%	25%	50%	75%	100%
1	1	2	2	3

```
affyMisSmall <- fetch(query,n=10); dbClearResult(query);
```

```
[1] TRUE
```

```
dim(affyMisSmall)
```

cerrar la conexion

```
dbDisconnect(hg19)
```

```
[1] TRUE
```

Recursos adicionales

- Viñeta RMySQL <http://cran.r-project.org/web/packages/RMySQL/RMySQL.pdf>
- Lista de comandos <http://www.pantz.org/software/mysql/mysqlcommands.html>
 - **No, no elimine, agregue o combine elementos de ensambl. Solo seleccione .**
 - En general, tenga cuidado con los comandos mysql
- Una buena publicación de blog que resume algunos otros comandos <http://www.r-bloggers.com/mysql-and-r/>
- <http://en.wikipedia.org/wiki/MySQL>
- <http://www.mysql.com/>

Leyendo HD5F

HDF5

- Se utiliza para almacenar grandes conjuntos de datos.
- Admite el almacenamiento de una variedad de tipos de datos
- Formato de datos jerárquico
- *grupos* que contienen cero o más conjuntos de datos y metadatos
 - Tener un *encabezado de grupo* con el nombre del grupo y una lista de atributos
 - Tener una *tabla de símbolos de grupo* con una lista de objetos en el grupo
- *datasets* matriz multidimensional de elementos de datos con metadatos
 - Tener un *encabezado* con nombre, tipo de datos, espacio de datos y diseño de almacenamiento
 - Tener un *data array* con los datos

<http://www.hdfgroup.org/>

```
invisible(if(file.exists("example.h5")){file.remove("example.h5")})
```

```
library(rhdf5)
file <- h5createFile("example.h5")
file
```

```
## [1] TRUE
```

- Esto instalará paquetes de Bioconductor <http://bioconductor.org/>, que se usa principalmente para genómica pero también tiene buenos paquetes de “big data”
- Se puede utilizar para interactuar con conjuntos de datos hdf5.
- Esta conferencia se basa muy de cerca en el tutorial de rhdf5 que se puede encontrar aquí <http://www.bioconductor.org/packages/release/bioc/vignettes/rhdf5/inst/doc/rhdf5.pdf>

crear grupos

```
created = h5createGroup("example.h5", "foo")
created = h5createGroup("example.h5", "baa")
```

```
created = h5createGroup("example.h5","foo/foobaa")
h5ls("example.h5")
```

```
##      group   name      otype dclass dim
## 0      /      baa  H5I_GROUP
## 1      /      foo  H5I_GROUP
## 2 /foo foobaa H5I_GROUP
```

Escribir a grupos

```
A = matrix(1:10,nr=5,nc=2)
h5write(A, "example.h5","foo/A")
B = array(seq(0.1,2.0,by=0.1),dim=c(5,2,2))
attr(B, "scale") <- "liter"
h5write(B, "example.h5","foo/foobaa/B")
h5ls("example.h5")
```

```
##      group   name      otype dclass      dim
## 0      /      baa  H5I_GROUP
## 1      /      foo  H5I_GROUP
## 2      /foo      A H5I_DATASET INTEGER    5 x 2
## 3      /foo foobaa H5I_GROUP
## 4 /foo/foobaa      B H5I_DATASET  FLOAT 5 x 2 x 2
```

escribir un conjunto de datos

```
df = data.frame(1L:5L,seq(0,1,length.out=5),
  c("ab","cde","fghi","a","s"), stringsAsFactors=FALSE)
h5write(df, "example.h5","df")
h5ls("example.h5")
```

```
##      group   name      otype dclass      dim
## 0      /      baa  H5I_GROUP
## 1      /      df  H5I_DATASET COMPOUND      5
## 2      /      foo  H5I_GROUP
## 3      /foo      A H5I_DATASET INTEGER    5 x 2
## 4      /foo foobaa H5I_GROUP
## 5 /foo/foobaa      B H5I_DATASET  FLOAT 5 x 2 x 2
```

leyendo datos

```
readA = h5read("example.h5","foo/A")
readB = h5read("example.h5","foo/foobaa/B")
readdf= h5read("example.h5","df")
readA
```

```
##      [,1] [,2]
## [1,]    1    6
## [2,]    2    7
## [3,]    3    8
## [4,]    4    9
## [5,]    5   10
```

Fragmentos de escritura y lectura

```
h5write(c(12,13,14),"example.h5","foo/A",index=list(1:3,1))
h5read("example.h5","foo/A")
```

```
##      [,1] [,2]
```

```
## [1,] 12 6
## [2,] 13 7
## [3,] 14 8
## [4,] 4 9
## [5,] 5 10
```

Notas y otros recursos

- hdf5 se puede utilizar para optimizar la lectura / escritura desde el disco en R
- El tutorial de rhdf5:
 - <http://www.bioconductor.org/packages/release/bioc/vignettes/rhdf5/inst/doc/rhdf5.pdf>
- El grupo HDF tiene información sobre HDF5 en general <http://www.hdfgroup.org/HDF5/>

Leyendo datos de la web

Webscraping

Webscraping: Extracción de datos de forma programada del código HTML de sitios web.

- Puede ser una excelente manera de obtener datos [Cómo Netflix hizo ingeniería inversa a Hollywood] (<http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/>)
- Muchos sitios web tienen información que puede querer leer programáticamente
- En algunos casos, esto va en contra de los términos de servicio del sitio web.
- Intentar leer demasiadas páginas demasiado rápido puede bloquear su dirección IP

ejemplo

Jeff Leek Edit
 Assistant Professor of Biostatistics, Johns Hopkins Bloomberg School of Public Health Edit
[Statistics](#) - [Computing](#) - [Genomics](#) - [Personalized Medicine](#) - [Scientific Communication](#) Edit
 Verified email at [jhsph.edu](mailto:jtleek@jhsph.edu) Edit
 My profile is public Edit [Link](#) [Homepage](#) Edit

Citation indices

	All	Since 2008
Citations	1285	1146
h-index	10	10
i10-index	11	11

Citations to my articles

Select: **All**, None Actions Show: 20 1-20 Next >

Title / Author	Cited by	Year
Significance analysis of time course microarray experiments <input type="checkbox"/> JD Storey, W Xiao, JT Leek, RG Tompkins, RW Davis Proceedings of the National Academy of Sciences of the United States of ...	338	2005
Capturing heterogeneity in gene expression studies by surrogate variable analysis <input type="checkbox"/> JT Leek, JD Storey PLoS Genetics 3 (9), e161	171	2007
EDGE: extraction and analysis of differential gene expression <input type="checkbox"/> JT Leek, E Monsen, AR Dabney, JD Storey Bioinformatics 22 (4), 507-508	140	2006
Tackling the widespread and critical impact of batch effects in high-throughput data <input type="checkbox"/> JT Leek, RB Scharpf, HC Bravo, D Simcha, B Langmead, WE Johnson, D Geman, K ... Nature Reviews Genetics 11 (10), 733-739	133	2010
The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments <input type="checkbox"/> JD Storey, JY Dai, JT Leek UW Biostatistics Working Paper Series, 260	107	2005
Systems-level dynamic analyses of fate change in murine embryonic stem		

Follow this author
 5 Followers
[Follow new articles](#)
[Follow new citations](#)

Add co-authors

John D. Storey	Add - X
Rafael A. Irizarry	Add - X
Ben Langmead	Add - X
Hector Corrada Br...	Add - X
wenzhong xiao	Add - X
W. Evan Johnson	Add - X
Alexander Lachm...	Add - X
Olga Troyanskaya	Add - X
Avi Ma'ayan	Add - X
Edoardo M Airolidi	Add - X

[View all co-authors](#)

Co-authors
 No co-authors
 Name
 Email
☐ Inviting co-author
[Send invitation](#)

`http://scholar.google.com/citations?user=HI-I6C0AAAAJ&hl=en`

```
con = url("http://scholar.google.com/citations?user=HI-I6C0AAAAJ&hl=en")
htmlCode = readLines(con)
close(con)
```

obteniendo informacion con paquete httr

```
library(httr)
library(XML)
url <- "http://scholar.google.com/citations?user=HI-I6C0AAAAJ&hl=en"
html2 = GET(url)
content2 = content(html2,as="text")
parsedHtml = htmlParse(content2,asText=TRUE)
xpathSApply(parsedHtml, "//title", xmlValue)
```

```
## [1] "Jeff Leek - Google Scholar"
```

accesando a sitios web con contraseña

```
pg1 = GET("http://httpbin.org/basic-auth/user/passwd")
pg1
```

```
## Response [http://httpbin.org/basic-auth/user/passwd]
##   Date: 2021-08-26 03:53
##   Status: 401
##   Content-Type: <unknown>
## <EMPTY BODY>
```

```
pg2 = GET("http://httpbin.org/basic-auth/user/passwd",
  authenticate("user","passwd"))
pg2
```

```
## Response [http://httpbin.org/basic-auth/user/passwd]
##   Date: 2021-08-26 03:53
##   Status: 200
##   Content-Type: application/json
##   Size: 47 B
## {
##   "authenticated": true,
##   "user": "user"
## }
```

```
names(pg2)
```

```
## [1] "url"          "status_code" "headers"      "all_headers" "cookies"
## [6] "content"      "date"         "times"        "request"      "handle"
```

usando handle

```
google = handle("http://google.com")
pg1 = GET(handle=google,path="/")
pg2 = GET(handle=google,path="search")

pg1
```

```
## Response [http://www.google.com/]
##   Date: 2021-08-26 03:53
##   Status: 200
```



```
## Content-Type: text/html; charset=ISO-8859-1
## Size: 14.2 kB
## <!doctype html><html itemscope="" itemtype="http://schema.org/WebPage" lang="...
## var f=this||self;var h,k=[];function l(a){for(var b;a&&(!a.getAttribute)||!(b=...
## function n(a,b,c,d,g){var e="";c||-1!==b.search("&ei=")|| (e="&ei="+l(d),-1===...
## google.y={};google.sy=[];google.x=function(a,b){if(a)var c=a.id;else{do c=Mat...
## document.documentElement.addEventListener("submit",function(b){var a;if(a=b.t...
## </style><style>body,td,a,p,.h{font-family:arial,sans-serif}body{margin:0;over...
## if (!iesg){document.f&&document.f.q.focus();document.gbqf&&document.gbqf.q.fo...
## }
## })();</script><div id="mngb"><div id=gbar><noobr><b class=gb1>Búsqueda</b> <a ...
## else top.location='/doodles/';};})();</script><input value="AINFCbYAAAAAYSceS...
## ...
```

pg2

```
## Response [http://www.google.com/webhp]
## Date: 2021-08-26 03:53
## Status: 200
## Content-Type: text/html; charset=ISO-8859-1
## Size: 14.2 kB
## <!doctype html><html itemscope="" itemtype="http://schema.org/WebPage" lang="...
## var f=this||self;var h,k=[];function l(a){for(var b;a&&(!a.getAttribute)||!(b=...
## function n(a,b,c,d,g){var e="";c||-1!==b.search("&ei=")|| (e="&ei="+l(d),-1===...
## google.y={};google.sy=[];google.x=function(a,b){if(a)var c=a.id;else{do c=Mat...
## document.documentElement.addEventListener("submit",function(b){var a;if(a=b.t...
## </style><style>body,td,a,p,.h{font-family:arial,sans-serif}body{margin:0;over...
## if (!iesg){document.f&&document.f.q.focus();document.gbqf&&document.gbqf.q.fo...
## }
## })();</script><div id="mngb"><div id=gbar><noobr><b class=gb1>Búsqueda</b> <a ...
## else top.location='/doodles/';};})();</script><input value="AINFCbYAAAAAYSceS...
## ...
```

Notas y otros recursos

- R Bloggers tiene varios ejemplos de raspado web <http://www.r-bloggers.com/?s=Web+Scraping>
- El archivo de ayuda httr tiene ejemplos útiles <http://cran.r-project.org/web/packages/httr/httr.pdf>
- Ver conferencias posteriores sobre API
- <http://cran.r-project.org/web/packages/httr/httr.pdf>
- http://en.wikipedia.org/wiki/Web_scraping

Leyendo de APIs

Application programming interfaces(Interfaces de programación de aplicaciones)

cursor to be -1 if it isn't supplied.
Example Values: 12893764510938

Example Request

GET `https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include_entities=true`

```

1. {
2.   "previous_cursor": 0,
3.   "previous_cursor_str": "0",
4.   "next_cursor": 0,
5.   "users": [
6.     {
7.       "profile_sidebar_border_color": "CODEED",
8.       "name": "Javier Heady \r",
9.       "profile_sidebar_fill_color": "DDEEF6",
10.      "profile_background_tile": false,
11.      "location": null,
12.      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
13.      "profile_image_url":
14.      "http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
15.      "is_translator": false,
16.      "id_str": "509466276",
17.      "profile_link_color": "0084B4",
18.      "follow_request_sent": false,
19.      "contributors_enabled": false,
20.      "default_profile": true,
21.      "url": null,
22.      "favourites_count": 0,
23.      "utc_offset": null,
24.      "id": 509466276,
25.      "profile_image_url_https":
26.      "https://s10.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
27.      "listed_count": 0,
28.      "profile_use_background_image": true,
29.      "profile_text_color": "333333",
30.      "lang": "en",
31.      "protected": false,
32.      "followers_count": 0,
33.      "geo_enabled": false,
34.      "description": null,

```

crear una cuenta para ingresar a la API

<https://dev.twitter.com/apps>

luego crear una “app” y buscar los pasos para acceder a la api, el ejemplo siguiente es de una cuenta ya creada para fines academicos

```

myapp = oauth_app("twitter",
                  key="yourConsumerKeyHere", secret="yourConsumerSecretHere")
sig = sign_oauth1.0(myapp,
                    token = "yourTokenHere",
                    token_secret = "yourTokenSecretHere")
homeTL = GET("https://api.twitter.com/1.1/statuses/home_timeline.json", sig)

```

convertir a JSON el objeto

```
json1 = content(homeTL)
json2 = jsonlite::fromJSON(toJSON(json1))
json2[1,1:4]
```

```
      created_at      id      id_str
1 Mon Jan 13 05:18:04 +0000 2014 4.225984e+17 422598398940684288

1 Now that P. Norvig's regex golf IPython notebook hit Slashdot, let's see if our traffic spike tops the prev
```

En general, mira la documentación.

- http permite solicitudes GET,POST, PUT,DELETE si está autorizado
- Puede autenticarse con un nombre de usuario o una contraseña
- La mayoría de las API modernas usan algo como oauth
- http funciona bien con Facebook, Google, Twitter, Github, etc.

Reading from other sources

Hay un paquete para eso

- Roger tiene un buen video sobre cómo hay paquetes R para la mayoría de las cosas a las que querrá acceder.
- Aquí voy a revisar brevemente algunos paquetes útiles
- En general, la mejor forma de averiguar si el paquete R existe es el “paquete R del mecanismo de almacenamiento de datos” de Google
 - Por ejemplo: “Paquete MySQL R”

Interactuar más directamente con archivos

- file: abre una conexión a un archivo de texto
- url: abre una conexión a una URL
- gzfile: abre una conexión a un archivo .gz
- bzfile: abre una conexión a un archivo .bz2
- *?connections* para más información
- Recuerde cerrar conexiones

paquete foreign

- Carga datos de Minitab, S, SAS, SPSS, Stata, Systat
- Funciones básicas *read.foo*
 - read.arff (Weka)
 - read.dta (Stata)
 - read.mtp (Minitab)
 - read.octave (octava)
 - read.spss (SPSS)
 - read.xport (SAS)
- Consulte la página de ayuda para obtener más detalles.

<http://cran.r-project.org/web/packages/foreign/foreign.pdf>

Ejemplos de otros paquetes de bases de datos

- RPostgreSQL proporciona una conexión de base de datos compatible con DBI desde R. Tutorial- <https://code.google.com/p/rpostgresql/>, archivo de ayuda- <http://cran.r-project.org/web/packages/RPostgreSQL/RPostgreSQL.pdf>
- RODB proporciona interfaces para múltiples bases de datos, incluidas PostgreSQL, MySQL, Microsoft Access y SQLite. Tutorial - <http://cran.r-project.org/web/packages/RODBC/vignettes/RODBC.pdf>, archivo de ayuda - <http://cran.r-project.org/web/packages/RODBC/RODBC.pdf>
- RMongo <http://cran.r-project.org/web/packages/RMongo/RMongo.pdf> (ejemplo de Rmongo <http://www.r-bloggers.com/r-and-mongodb/>) y rmongodb proporcionan interfaces a MongoDB.

Leer imágenes

- jpeg - <http://cran.r-project.org/web/packages/jpeg/index.html>
- readbitmap - <http://cran.r-project.org/web/packages/readbitmap/index.html>
- png - <http://cran.r-project.org/web/packages/png/index.html>
- EBImage (Bioconductor) - <http://www.bioconductor.org/packages/2.13/bioc/html/EBImage.html>

Lectura de datos GIS

- rgdal - <http://cran.r-project.org/web/packages/rgdal/index.html>
- rgeos - <http://cran.r-project.org/web/packages/rgeos/index.html>
- raster - <http://cran.r-project.org/web/packages/raster/index.html>

Leer datos musicales

- tuneR - <http://cran.r-project.org/web/packages/tuneR/>
- seewave - <http://rug.mnhn.fr/seewave/>

leer tablas de web

En la pagina <https://www.ncdc.noaa.gov/snow-and-ice/rsi/> se puede encontrar la siguiente imagen

Home > Climate Monitoring > Snow and Ice > RSI

August US Release: Thu, 9 Sep 2021, 11:00 AM EDT

Regional Snowfall Index (RSI)

Overview | RSI and Societal Impacts | Historic Storms | FAQ | NESIS | References

Overview

NOAA's National Centers for Environmental Information is now producing the Regional Snowfall Index (RSI) for significant snowstorms that impact the eastern two thirds of the U.S. The RSI ranks snowstorm impacts on a scale from 1 to 5, similar to the Fujita scale for tornadoes or the Saffir-Simpson scale for hurricanes.

Data Visualization

RSI Map Viewer
NESIS

CATEGORY	RSI VALUE	DESCRIPTION
1	1-3	Notable
2	3-6	Significant
3	6-10	Major
4	10-18	Crippling
5	18.0+	Extreme

The RSI differs from these other indices because it includes population. RSI is based on the spatial extent of the storm, the amount of snowfall, and the juxtaposition of these elements with population. Including population information ties the index to societal impacts. Currently, the index

y la tabla se obtiene de la siguiente forma

```
rio::import("https://www.ncdc.noaa.gov/snow-and-ice/rsi/",format="html")
```

```
##   Category RSI Value Description
## 1         1    1-3    Notable
## 2         2    3-6 Significant
## 3         3   6-10      Major
## 4         4  10-18 Crippling
## 5         5  18.0+   Extreme
```