Local setup of Solr and querying using solr R package, on Mac OSX

A general purpose R interface to Solr

This package only deals with extracting data from a Solr endpoint, not writing data (pull request or holla if you're interested in writing solr data).

Solr info

- Solr home page
- Highlighting help
- Faceting help
- Installing Solr on Mac using homebrew
- Install and Setup SOLR in OSX, including running Solr

Quick start

Install

Install dependencies

```r
install.packages(c("rjson", "plyr", "httr", "XML", "assertthat"))
```

Install solr

```r
install.packages("devtools")
library(devtools)
install_github("ropensci/solr")
```

```r
library(solr)
```

Define stuff Your base url and a key (if needed). This example should work. You do need to pass a key to the Public Library of Science search API, but it apparently doesn't need to be a real one.

```r
url <- "http://api.plos.org/search"
key <- "key"
```

Search

```r
solr_search(q = "*:*", rows = 2, fl = "id", base = url, key = key)
```

```
## http://api.plos.org/search?q=*:*&start=0&rows=2&wt=json&fl=id
```

```
##                               id
## 1 10.1371/journal.pone.0060627
## 2 10.1371/journal.pbio.0000080
```

Facet

```
solr_facet(q = "*:*", facet.field = "journal", facet.query = c("cell",
    "bird"), base = url, key = key)
```

```
## http://api.plos.org/search?q=*:*&facet.query=cell&facet.query=bird&facet.field=journal&key=key&wt=js

## $facet_queries
##   term value
## 1 cell 85941
## 2 bird  8588
##
## $facet_fields
## $facet_fields$journal
##                                  X1     X2
## 1                          plos one 742824
## 2                     plos genetics  35463
## 3                    plos pathogens  31152
## 4        plos computational biology  26016
## 5                      plos biology  24699
## 6  plos neglected tropical diseases  20115
## 7                     plos medicine  17444
## 8               plos clinical trials    521
## 9                      plos medicin      9
## 10                 plos collections      5
##
##
## $facet_dates
## NULL
##
## $facet_ranges
## NULL
```

Highlight

```
solr_highlight(q = "alcohol", hl.fl = "abstract", rows = 2, base = url, key = key)
```

```
## http://api.plos.org/search?wt=json&q=alcohol&start=0&rows=2&hl=true&fl=DOES_NOT_EXIST&hl.fl=abstract

## $`10.1371/journal.pmed.0040151`
## $`10.1371/journal.pmed.0040151`$abstract
## [1] "Background: <em>Alcohol</em> consumption causes an estimated 4% of the global disease burden, p
##
##
## $`10.1371/journal.pone.0027752`
## $`10.1371/journal.pone.0027752`$abstract
## [1] "Background: The negative influences of <em>alcohol</em> on TB management with regard to delays
```

Stats

```
out <- solr_stats(q = "ecology", stats.field = c("counter_total_all", "alm_twitterCount"),
    stats.facet = c("journal", "volume"), base = url, key = key)
```

```
## http://api.plos.org/search?q=ecology&stats.field=counter_total_all&stats.field=alm_twitterCount&stats
```

```
out$data
```

```
##                    min    max count missing      sum sumOfSquares     mean
## counter_total_all   0 297294 19679       0 64851389    1.097e+12 3295.462
## alm_twitterCount    0   1446 19679       0    71992    1.011e+07    3.658
##                     stddev
## counter_total_all 6699.81
## alm_twitterCount    22.37
```

```
out$facet
```

```
## $counter_total_all
## $counter_total_all$journal
##    min    max count missing      sum sumOfSquares  mean stddev
## 1    0  39085   427       0  2285267    2.027e+10  5352   4343
## 2    0  43592   557       0  3336132    3.196e+10  5989   4642
## 3    0 297294 15379       0 40023738    6.223e+11  2602   5804
## 4 4638   8607     2       0    13245    9.559e+07  6622   2807
## 5  513  85165   213       0  2361321    5.359e+10 11086  11371
## 6  768  57904   378       0  2071231    2.359e+10  5479   5698
## 7  574 168945   758       0  8871519    2.341e+11 11704  13116
## 8    0 164090   714       0  2394341    3.951e+10  3353   6645
##                        facet_field
## 1                   plos pathogens
## 2                    plos genetics
## 3                         plos one
## 4             plos clinical trials
## 5                    plos medicine
## 6       plos computational biology
## 7                     plos biology
## 8 plos neglected tropical diseases
##
## $counter_total_all$volume
##     min    max count missing      sum sumOfSquares  mean stddev
## 1   859 108653   741       0  5231098    9.622e+10  7060   8951
## 2  1159  86761   482       0  4062160    8.123e+10  8428   9885
## 3     0  82673   136       0   991749    2.279e+10  7292  10736
## 4  1391 111334    81       0  1088239    3.765e+10 13435  16965
## 5     0 179433  4825       0 13328457    1.883e+11  2762   5604
## 6     0 164090  2948       0 10560418    1.396e+11  3582   5876
## 7     0  74838  1539       0  7624055    8.949e+10  4954   5799
## 8   513 297294  1010       0  6467119    1.909e+11  6403  12172
## 9     0 168945  1709       0  3117421    6.074e+10  1824   5677
## 10    0 188324  6131       0 11597343    1.716e+11  1892   4941
## 11  610  74895    66       0   714981    1.722e+10 10833  12076
## 12  574  33078    11       0    68349    1.241e+09  6214   9036
##    facet_field
## 1            3
## 2            2
## 3           10
## 4            1
## 5            7
## 6            6
## 7            5
## 8            4
## 9            9
## 10           8
## 11          11
## 12          12
##
##
## $alm_twitterCount
## $alm_twitterCount$journal
##   min  max count missing   sum sumOfSquares   mean stddev
## 1   0   74   427       0  1387        35947  3.248  8.591
## 2   0  141   557       0  1648        49984  2.959  9.007
## 3   0  781 15379       0 50416      5548300  3.278 18.710
## 4   0    3     2       0     3            9  1.500  2.121
## 5   0  524   213       0  2370       439366 11.127 44.137
## 6   0  104   378       0  1224        39048  3.238  9.647
## 7   0 1446   758       0  6591      2966605  8.695 61.993
## 8   0  800   714       0  1937       654019  2.713 30.165
##                        facet_field
## 1                   plos pathogens
## 2                    plos genetics
## 3                         plos one
## 4             plos clinical trials
## 5                    plos medicine
## 6       plos computational biology
## 7                     plos biology
## 8 plos neglected tropical diseases
##
## $alm_twitterCount$volume
##    min  max count missing   sum sumOfSquares    mean  stddev facet_field
## 1    0   29   741       0   342         3146  0.4615   2.009           3
## 2    0   36   482       0   282         4512  0.5851   3.006           2
## 3    0  524   136       0  2981       456107 21.9191  53.801          10
## 4    0   28    81       0    87         1655  1.0741   4.418           1
## 5    0  781  4825       0 17405      1696211  3.6073  18.401           7
## 6    0  800  2948       0  2904       820122  0.9851  16.653           6
## 7    0  111  1539       0  1142        43334  0.7420   5.256           5
## 8    0  151  1010       0   533        28965  0.5277   5.332           4
## 9    0  307  1709       0 11031       696865  6.4547  19.139           9
## 10   0  767  6131       0 29602      3428324  4.8282  23.151           8
## 11   1 1446    66       0  4602      2504276 69.7273 183.277          11
## 12   7  630    11       0  1081       430679 98.2727 180.124          12
```

More like this

solr_mlt is a function to return similar documents to the one

```
out <- solr_mlt(q = "title:\"ecology\" AND body:\"cell\"", mlt.fl = "title",
    mlt.mindf = 1, mlt.mintf = 1, fl = "counter_total_all", rows = 5, base = url,
    key = key)
```

```
## http://api.plos.org/search?q=title:"ecology" AND body:"cell"&mlt=true&fl=id,counter_total_all&mlt.fl=
out$docs
```

```
##                              id counter_total_all
## 1 10.1371/journal.pbio.1001805               574
## 2 10.1371/journal.pbio.0020440             16114
## 3 10.1371/journal.pone.0087217              1095
## 4 10.1371/journal.pone.0040117              1754
## 5 10.1371/journal.pone.0072525               714
```

```
out$mlt
```

```
## $`10.1371/journal.pbio.1001805`
##                              id counter_total_all
## 1 10.1371/journal.pone.0082578               573
## 2 10.1371/journal.pone.0087380               291
## 3 10.1371/journal.pcbi.1003408              2521
## 4 10.1371/journal.pcbi.1002915              4132
## 5 10.1371/journal.pcbi.1002652              2110
##
## $`10.1371/journal.pbio.0020440`
##                              id counter_total_all
## 1 10.1371/journal.pone.0035964              2660
## 2 10.1371/journal.pone.0003259              1728
## 3 10.1371/journal.pone.0068814              4539
## 4 10.1371/journal.pbio.0020215              4274
## 5 10.1371/journal.pbio.0020148             11359
##
## $`10.1371/journal.pone.0087217`
##                              id counter_total_all
## 1 10.1371/journal.pcbi.0020092             13333
## 2 10.1371/journal.pone.0063375               988
## 3 10.1371/journal.pcbi.1000986              2650
## 4 10.1371/journal.pntd.0000694              1806
## 5 10.1371/journal.pone.0015143             11368
##
## $`10.1371/journal.pone.0040117`
##                              id counter_total_all
## 1 10.1371/journal.pone.0069352               946
## 2 10.1371/journal.pone.0014065              3501
## 3 10.1371/journal.pone.0035502              2009
## 4 10.1371/journal.pone.0078369               980
## 5 10.1371/journal.pone.0084920               653
##
## $`10.1371/journal.pone.0072525`
##                              id counter_total_all
## 1 10.1371/journal.pone.0060766               914
## 2 10.1371/journal.pcbi.1002928              6369
## 3 10.1371/journal.pcbi.0020144             11857
## 4 10.1371/journal.pcbi.1000350              8200
## 5 10.1371/journal.pone.0068714              2164
```

Parsing

`solr_parse` is a general purpose parser function with extension methods `solr_parse.sr_search`, `solr_parse.sr_facet`, and `solr_parse.sr_high`, for parsing `solr_search`, `solr_facet`, and `solr_highlight` function output, respectively. `solr_parse` is used internally within those three functions (`solr_search`, `solr_facet`, `solr_highlight`) to do parsing. You can optionally get back raw `json` or `xml` from `solr_search`, `solr_facet`, and `solr_highlight` setting parameter `raw=TRUE`, and then parsing after the fact with `solr_parse`. All you need to know is `solr_parse` can parse

For example:

```
(out <- solr_highlight(q = "alcohol", hl.fl = "abstract", rows = 2, base = url,
    key = key, raw = TRUE))
```

```
## http://api.plos.org/search?wt=json&q=alcohol&start=0&rows=2&hl=true&fl=DOES_NOT_EXIST&hl.fl=abstract
```

```
## [1] "{\"response\":{\"numFound\":12306,\"start\":0,\"docs\":[{},{}]},\"highlighting\":{\"10.1371/jou
## attr(,"class")
## [1] "sr_high"
## attr(,"wt")
## [1] "json"
```

Then parse

```
solr_parse(out, "df")
```

```
##                            names
## 1 10.1371/journal.pmed.0040151
## 2 10.1371/journal.pone.0027752
##                                                                                                     al
## 1   Background: <em>Alcohol</em> consumption causes an estimated 4% of the global disease burden, pro
## 2 Background: The negative influences of <em>alcohol</em> on TB management with regard to delays in s
```

Using specific data sources

USGS BISON service

The occurrences service

```
url2 <- "http://bisonapi.usgs.ornl.gov/solr/occurrences/select"
solr_search(q = "*:*", fl = c("latitude", "longitude", "scientific_name"), base = url2)
```

```
## http://bisonapi.usgs.ornl.gov/solr/occurrences/select?q=*:*&start=0&wt=json&fl=latitude&fl=longitude&
```

```
## data frame with 0 columns and 0 rows
```

The species names service

```
solr_search(q = "*:*", base = url2, raw = TRUE)
```

```
## http://bisonapi.usgs.ornl.gov/solr/occurrences/select?q=*:*&start=0&wt=json
## [1] "{\"responseHeader\":{\"status\":0,\"QTime\":1033},\"response\":{\"numFound\":126357352,\"start\
## attr(,"class")
## [1] "sr_search"
## attr(,"wt")
## [1] "json"
```

PLOS Search API

Most of the examples above use the PLOS search API… :)

Please report any issues or bugs.