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 exracting 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

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

Install solr

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

```
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.

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

Search

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

```
##                              id
## 1 10.1371/journal.pone.0025014
## 2 10.1371/journal.pone.0055525
```

Facet

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

```
## $facet_queries
##   term value
## 1 cell 79476
## 2 bird  7965
##
## $facet_fields
## $facet_fields$journal
##                                  X1     X2
## 1                         plos one 663280
## 2                     plos genetics  33284
## 3                    plos pathogens  29244
## 4        plos computational biology  24845
## 5                      plos biology  23926
## 6  plos neglected tropical diseases  18781
## 7                     plos medicine  17031
## 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, url = url, key = key)
```

```
## $`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 = "counter_total_all,alm_twitterCount",
    stats.facet = "journal,volume", url = url, key = key)
```

```
out$data
```

```
##                    min    max count missing      sum sumOfSquares     mean
## counter_total_all    0 291798 18090       0 58248156    9.639e+11 3219.909
## alm_twitterCount     0   1288 18090       0    56281    7.406e+06    3.111
##                     stddev
## counter_total_all 6551.12
## alm_twitterCount    19.99
```

```
out$facet
```

```
## $counter_total_all
## $counter_total_all$journal
##    min    max count missing      sum sumOfSquares  mean stddev
## 1    0  37364   404       0  2067577    1.767e+10  5118   4193
## 2    0  42118   529       0  3035262    2.790e+10  5738   4456
## 3    0 291798 13909       0 35301226    5.395e+11  2538   5688
## 4 4168   8060     2       0    12228    8.234e+07  6114   2752
## 5    0  82757   208       0  2158539    4.224e+10 10378   9789
## 6 1083 156837   746       0  8466420    2.151e+11 11349  12638
## 7    0  53230   365       0  1885392    1.917e+10  5165   5089
## 8    0 156975   676       0  2144469    3.551e+10  3172   6521
##                         facet_field
## 1                   plos pathogens
## 2                    plos genetics
## 3                        plos one
## 4             plos clinical trials
## 5                    plos medicine
## 6                     plos biology
## 7       plos computational biology
## 8 plos neglected tropical diseases
##
## $counter_total_all$volume
##     min    max count missing      sum sumOfSquares  mean stddev
## 1   816 107405   741       0  5068779    9.137e+10  6840   8754
## 2  1132  85278   482       0  3949081    7.702e+10  8193   9636
## 3  1372 108353    81       0  1065357    3.599e+10 13153  16573
## 4     0  59941    71       0   708999    1.306e+10  9986   9246
## 5     0 178757  4823       0 12104091    1.717e+11  2510   5414
## 6   505 156975  2946       0  9871464    1.220e+11  3351   5495
## 7   470  73727  1538       0  7245872    8.175e+10  4711   5566
## 8   493 291798  1010       0  6224943    1.807e+11  6163  11877
## 9     0 156837   354       0  1880616    4.070e+10  5312   9327
## 10    0 149871  5983       0  9502785    1.356e+11  1588   4489
## 11 1147  66540    61       0   626169    1.393e+10 10265  11180
##    facet_field
## 1            3
## 2            2
## 3            1
## 4           10
## 5            7
## 6            6
## 7            5
## 8            4
## 9            9
## 10           8
## 11          11
##
##
## $alm_twitterCount
## $alm_twitterCount$journal
##   min  max count missing   sum sumOfSquares  mean stddev
## 1   0   73   404       0  1172        30074 2.901  8.136
```

```
## 2   0   48   529       0  1146        19558 2.166  5.687
## 3   0  733 13909       0 38274      4148472 2.752 17.050
## 4   0    3     2       0     3            9 1.500  2.121
## 5   0  201   208       0  1568       138226 7.538 24.711
## 6   0 1288   746       0  4975      2034243 6.669 51.827
## 7   0  102   365       0  1081        35411 2.962  9.407
## 8   0  784   676       0  1711       625745 2.531 30.342
##                        facet_field
## 1                   plos pathogens
## 2                    plos genetics
## 3                         plos one
## 4               plos clinical trials
## 5                     plos medicine
## 6                      plos biology
## 7         plos computational biology
## 8 plos neglected tropical diseases
##
## $alm_twitterCount$volume
##    min  max count missing   sum sumOfSquares    mean  stddev facet_field
## 1    0   17   741       0   292         2136  0.3941   1.653           3
## 2    0   35   482       0   256         3778  0.5311   2.752           2
## 3    0   28    81       0    80         1582  0.9877   4.334           1
## 4    0  201    71       0  1735       140243 24.4366  37.387          10
## 5    0  733  4823       0 16890      1547170  3.5020  17.567           7
## 6    0  784  2946       0  2634       750518  0.8941  15.939           6
## 7    0  110  1538       0  1004        38182  0.6528   4.941           5
## 8    0  142  1010       0   472        25576  0.4673   5.013           4
## 9    0  150   354       0  2871       112269  8.1102  15.877           9
## 10   0  727  5983       0 26011      2785113  4.3475  21.135           8
## 11   1 1288    61       0  4036      1998982 66.1639 169.899          11
```

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, url = url,
    key = key)
out$docs
```

```
##                             id counter_total_all
## 1 10.1371/journal.pbio.0020440             15977
## 2 10.1371/journal.pone.0040117              1589
## 3 10.1371/journal.pone.0072525               635
## 4 10.1371/journal.ppat.1002320              4612
## 5 10.1371/journal.pone.0015143             11003
```

```
out$mlt
```

```
##                              id counter_total_all
## 1  10.1371/journal.pone.0035964              2247
## 2  10.1371/journal.pone.0003259              1693
## 3  10.1371/journal.pone.0068814              3953
```

```
## 4  10.1371/journal.pbio.0020148            11186
## 5  10.1371/journal.pbio.0030105             2761
## 6  10.1371/journal.pone.0069352              647
## 7  10.1371/journal.pone.0014065             3311
## 8  10.1371/journal.pone.0035502             1757
## 9  10.1371/journal.pone.0078369              455
## 10 10.1371/journal.pone.0048646             1357
## 11 10.1371/journal.pone.0060766              831
## 12 10.1371/journal.pcbi.1002928             6051
## 13 10.1371/journal.pcbi.0020144            11556
## 14 10.1371/journal.pcbi.1000350             7925
## 15 10.1371/journal.pone.0068714             1363
## 16 10.1371/journal.pbio.1001332            12315
## 17 10.1371/journal.ppat.1000222             9901
## 18 10.1371/journal.pone.0052612             1223
## 19 10.1371/journal.pntd.0001693             2402
## 20 10.1371/journal.pntd.0001283             3505
## 21 10.1371/journal.pbio.1001702             1576
## 22 10.1371/journal.pone.0008413             5687
## 23 10.1371/journal.pone.0014451             4823
## 24 10.1371/journal.ppat.1003500             2212
## 25 10.1371/journal.pone.0035348             5200
```

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, url = url,
    key = key, raw = TRUE))
```

```
## [1] "{\"response\":{\"numFound\":11203,\"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
##                                                                                                  ab
## 1   Background: <em>Alcohol</em> consumption causes an estimated 4% of the global disease burden, pr
## 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 = "latitude,longitude,scientific_name", url = url2)
```

```
##    longitude latitude        scientific_name
## 1     -75.12    40.23 Catostomus commersonii
## 2     -75.12    40.23  Ambloplites rupestris
## 3     -75.12    40.23      Anguilla rostrata
## 4     -75.12    40.23      Anguilla rostrata
## 5     -75.12    40.23 Catostomus commersonii
## 6     -75.12    40.23  Ambloplites rupestris
## 7     -75.12    40.23      Lepomis cyanellus
## 8     -75.12    40.23      Lepomis cyanellus
## 9     -75.12    40.23     Fundulus diaphanus
## 10    -75.12    40.23    Etheostoma olmstedi
```

The species names service

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

```
## [1] "{\"responseHeader\":{\"status\":0,\"QTime\":509},\"response\":{\"numFound\":111109690,\"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.