

Package ‘ssrch’

May 25, 2024

Title a simple search engine

Description Demonstrate tokenization and a search gadget for collections of CSV files.

Version 1.21.0

Author Vince Carey

Suggests knitr, testthat, rmarkdown, BiocStyle

Depends R (>= 3.6), methods

Imports shiny, DT, utils

Maintainer VJ Carey <stvjc@channing.harvard.edu>

License Artistic-2.0

LazyLoad yes

LazyData yes

biocViews Infrastructure

VignetteBuilder knitr

RoxygenNote 6.1.1

Encoding UTF-8

git_url <https://git.bioconductor.org/packages/ssrch>

git_branch devel

git_last_commit 47dc6e1

git_last_commit_date 2024-04-30

Repository Bioconductor 3.20

Date/Publication 2024-05-24

Contents

| | |
|----------------------------|---|
| ctxsearch | 2 |
| DocSet | 3 |
| DocSet-class | 4 |
| docset_cancer68 | 5 |
| docset_searchapp | 5 |

| | |
|----------------------------|-----------|
| ds_can1009b | 6 |
| parseDoc | 7 |
| study_publ_dates | 8 |
| titles68 | 8 |
| urls68 | 9 |
| Index | 10 |

| | |
|-----------|--|
| ctxsearch | <i>ssrch demo with metadata documents from 68 cancer transcriptomics studies</i> |
|-----------|--|

Description

ssrch demo with metadata documents from 68 cancer transcriptomics studies

Usage

```
ctxsearch()
```

Value

Simply starts an app.

Note

The metadata were derived by extracting sample.attributes fields from a search with github.com/seandavi/SRAdbV2. The sample.attributes content varies between studies and sometimes between experiments within studies. The field sets were unified with the sampleAtts function of github.com/vjcitn/HumanTranscriptomeCompendium. After unification records were stacked and CSVs were written.

Examples

```
if (interactive()) {
  oask = options()$example.ask
  options(example.ask=FALSE)
  try(ctxsearch2())
  options(example.ask=oask)
}
```

| | |
|--------|-------------------------------|
| DocSet | <i>constructor for DocSet</i> |
|--------|-------------------------------|

Description

constructor for DocSet

Usage

```
DocSet(kw2docs = new.env(hash = TRUE), docs2recs = new.env(hash =
  TRUE), docs2kw = new.env(hash = TRUE), titles = character(),
  urls = character(), doc_retriever = function(...) NULL)
```

Arguments

| | |
|---------------|--|
| kw2docs | an environment mapping keywords to documents |
| docs2recs | an environment mapping document identifiers to records |
| docs2kw | an environment mapping documents to keywords |
| titles | a named character vector with titles; names are document identifiers |
| urls | a named character vector with document-associated URLs; names are document identifiers |
| doc_retriever | a function that, given a document identifier, will produce the document |

Value

instance of DocSet

Note

Titles must be bound in post-hoc. `parseDoc` produces data including parsed titles but does not bind the title into the resulting object.

Examples

```
getClass("DocSet")
```

| | |
|--------------|--|
| DocSet-class | <i>Container for simple documents with arbitrary numbers/shapes of records</i> |
|--------------|--|

Description

Container for simple documents with arbitrary numbers/shapes of records
utilities for srch

Usage

```
kw2docs(sdata)
docs2kw(sdata)
docs2recs(sdata)
searchDocs(string, obj, ...)
retrieve_doc(x, obj, ...)
```

Arguments

| | |
|--------|----------------------------------|
| sdata | instance of srchData class |
| string | character(1) query string |
| obj | instance of DocSet class |
| ... | passed to base::grep |
| x | character(1) document identifier |

Value

an environment
an environment
an environment
a data.frame with tokens queried (hits) and associated document ids (docs)
result of calling obj@doc_retriever on arguments x, ...

Examples

```
getClass("DocSet")
```

| | |
|-----------------|---|
| docset_cancer68 | <i>DocSet instance with metadata from 68 cancer studies</i> |
|-----------------|---|

Description

DocSet instance with metadata from 68 cancer studies

Usage

```
docset_cancer68
```

Format

S4 class DocSet defined in ssrch

| | |
|------------------|---|
| docset_searchapp | <i>interactive app for ssrch DocSet instances</i> |
|------------------|---|

Description

interactive app for ssrch DocSet instances

Usage

```
docset_searchapp(docset, se = NULL, sefilter = function(se, ...) se)
```

Arguments

| | |
|----------|---|
| docset | an instance of DocSet |
| se | (defaults to NULL) an instance of SummarizedExperiment; samples will be filtered by selection method prescribed in sefilter |
| sefilter | a function accepting (se, ...) and returning a SummarizedExperiment |

Value

Returns list of data.frames of metadata on studies requested. Can provide a SummarizedExperiment download when 'se' is non-null, but this is not yet returned to the session.

Note

The handling of SummarizedExperiments by this app is specialized. The 'sefilter' for the cancer example would be 'function(se, y) se[,which(se\$study_accession will be called with 'y' bound to the study accession numbers selected in the app.

Examples

```
if (interactive()) {  
  oask = options()$example.ask  
  options(example.ask=FALSE)  
  n1 = try(docset_searchapp(ssrch::docset_cancer68))  
  str(n1)  
  options(example.ask=oask)  
}
```

ds_can1009b

DocSet instance with metadata from 1009 cancer studies

Description

DocSet instance with metadata from 1009 cancer studies

Usage

```
ds_can1009b()
```

Format

S4 class DocSet defined in ssrch

Value

DocSet instance

Note

This is part of a sequence of datasets assessing how far we can go with environments of keywords. Annotation for 43000 samples is indexed here.

Examples

```
ds_can1009b()
```

| | |
|----------|---|
| parseDoc | <i>parse a document and place content in a DocSet</i> |
|----------|---|

Description

parse a document and place content in a DocSet

Usage

```
parseDoc(csv, DocSetInstance = new("DocSet"), doctitle = NA_character_,
  docabst = NA_character_, rec_id_field = "experiment.accession",
  exclude_fields = c("study.accession"),
  substrings_to_omit = c("http://purl.obolibrary.org/obo/"),
  patterns_to_kill = "...-..-..|.*...,..",
  token_fixups = list(c("t'", "t'"), c(":$", "")), max_tok_nchar = 25,
  min_tok_nchar = 4, cleanFields = list(".*id$", ".name$", "_name$",
  "checksum", "isolate", "filename", "^ID$", "barcode", "Sample.Name"))
```

Arguments

| | |
|--------------------|--|
| csv | a character(1) CSV file path |
| DocSetInstance | if missing, DocSet is initialized in this function, otherwise the instance is updated with new content |
| doctitle | character(1) document title |
| docabst | character(1) a string: the document abstract |
| rec_id_field | character(1) field in CSV identifying records |
| exclude_fields | character vector of fields to ignore while parsing |
| substrings_to_omit | character vector of strings to remove from candidate keywords via gsub |
| patterns_to_kill | character(1) regexp that identifies tokens to be omitted from keyword set |
| token_fixups | a list of character(2) vectors that will be |
| max_tok_nchar | numeric(1) defaults to 25, tokens with more characters will be truncated to this length and suffixed with ellipsis |
| min_tok_nchar | numeric(1) defaults to 4, tokens shorter than this are not in index used with gsub() to repair irregularities. For example 'c("t'", "t'")' will transform 'Burkitt's' to 'Burkitt's' |
| cleanFields | list of regular expressions identifying fields to ignore |

Value

instance of DocSet

Note

The expected use case has ‘DocSetInstance’ being updated in a loop. Sharing of environments across multiple DocSetInstances can occur and unexpected behaviors may ensue. Note also that many of the parameter defaults to parseDoc are for the use case of processing SRA metadata.

Examples

```
myob = ssrch::docset_cancer68
td = tempdir()
alld = ls(docs2kw(myob))
r1 = retrieve_doc(alld[1], myob)
expo = write.csv(r1, paste0(td, "/expo.csv"))
pd = parseDoc(paste0(td, "/expo.csv"), doctitle=ssrch::titles68[alld[1]],
  docabst="qwerty")
pd
searchDocs("quer", pd) # query will fail
searchDocs("qwer", pd) # should succeed
```

| | |
|------------------|---|
| study_publ_dates | <i>publication dates for 6000 SRA transcriptome studies</i> |
|------------------|---|

Description

publication dates for 6000 SRA transcriptome studies

Usage

```
study_publ_dates
```

Format

```
data.frame
```

| | |
|----------|-------------------------------------|
| titles68 | <i>titles for 68 cancer studies</i> |
|----------|-------------------------------------|

Description

titles for 68 cancer studies

Usage

```
titles68
```

Format

```
character vector
```

| | |
|--------|--|
| urls68 | <i>pubmed URLs for subset of 68 cancer studies</i> |
|--------|--|

Description

pubmed URLs for subset of 68 cancer studies

Usage

urls68

Format

character vector

Index

* datasets

- docset_cancer68, [5](#)
- study_publ_dates, [8](#)
- titles68, [8](#)
- urls68, [9](#)

ctxsearch, [2](#)

docs2kw (DocSet-class), [4](#)
docs2recs (DocSet-class), [4](#)
DocSet, [3](#)
DocSet-class, [4](#)
docset_cancer68, [5](#)
docset_searchapp, [5](#)
ds_can1009b, [6](#)

kw2docs (DocSet-class), [4](#)

parseDoc, [7](#)

retrieve_doc (DocSet-class), [4](#)

searchDocs (DocSet-class), [4](#)
study_publ_dates, [8](#)

titles68, [8](#)

urls68, [9](#)