# Package 'rodham'

July 18, 2017

**Type** Package

**Title** Fetch Hillary Rodham Clinton's Emails

**Version** 0.1.1

**Maintainer** John Coene <jcoenep@gmail.com>

**Description** Fetch and process Hillary Rodham Clinton's ``personal'' emails.

**License** MIT + file LICENSE

**LazyData** TRUE

**Depends** R (>= 2.10)

**Imports** jsonlite, splitstackshape, plyr, stringr, tibble, methods, utils

**RoxygenNote** 6.0.1

**Suggests** testthat, covr, knitr, rmarkdown, igraph, lintr

**VignetteBuilder** knitr

**URL** https://github.com/JohnCoene/rodham

**BugReports** https://github.com/JohnCoene/rodham/issues

**NeedsCompilation** no

**Author** John Coene [aut, cre]

**Repository** CRAN

**Date/Publication** 2017-07-18 10:21:24 UTC

## R topics documented:

---

`clean_content` *Clean emails*

---

## Description

Clean emails by removing *useless* lines.

## Usage

```
clean_content(content)
```

## Arguments

| | |
|---|---|
| `content` | list of content as returned by `get_content`. |

## Details

Example of line removed `UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C05765911 Dat`
look at the source code for more details `clean_content`.

## Author(s)

John Coene <jcoenep@gmail.com>

## Examples

```
## Not run:
hrc_emails <- load_emails(emails_bengh) # load emails
cont <- get_content(hrc_emails)
cont <- clean_content(hrc_emails)

## End(Not run)
```

`download_emails` *Download emails*

## Description

Download emails manually

## Usage

```
download_emails(release, save.dir = getwd())
```

## Arguments

| | |
|---|---|
| `release` | Name of the batch of release of emails; see details. |
| `save.dir` | Directory where to save the downloaded emails, defaults to `getwd()` |

## Details

Below are the valid values for `release`; follows the WSJ naming convention.

- Benghazi
- June
- July
- August
- September
- October
- November
- January 7
- January 29
- February 19
- february 29
- December
- Non-disclosure

## Value

Returns full path to downloaded zip or tar.

## Author(s)

John Coene <jcoenep@gmail.com>

## See Also

`get_xpdf`, `extract_emails`

## edges_emails *Network the treacherous*

### Description

Builds edge list from emails using `from` and `to`, edge source and target respectively.

### Usage

```
edges_emails(emails = emails, ...)
```

### Arguments

| | |
|---|---|
| `emails` | Data frame of emails as returned by `search_emails`, defaults to `emails` (see `emails`) |
| `...` | any additional column to keep as meta-data |

### Author(s)

John Coene <jcoenep@gmail.com>

### See Also

`search_emails`

### Examples

```
## Not run:
emails <- search_emails()

edges <- edges_emails(emails)

## End(Not run)
```

## emails *Hillary Rodham Clinton emails*

### Description

A dataset containing 29444 emails from/to Hillary Rodham Clinton sent/received between 2009-08-14 and 2014-08-13.

### Usage

```
emails
```

### Format

A data frame with 29444 rows and 9 variables:

**docID** Primary key

**docDate** Date when document was sent or received

**to** Who the emails was sent to

**from** Who the email is received from

**originalTo** From whom the email originally comes from

**originalFrom** To whom the email was originally sent to

**subject** Subject of the email

**interesting** Rating, relevancy of email

**not_interesting** Rating, irrelevancy of email

### Source

http://graphics.wsj.com/hillary-clinton-email-documents/api/search.php?subject=&to=&from=&start=&end=&sort=docDate&order=desc&docid=&limit=27159&offset=0

---

## extract_emails *Extract emails contents*

---

### Description

Extract content of manually downloaded emails.

### Usage

```
extract_emails(release, save.dir = getwd(), extractor, ...)
```

### Arguments

| | |
|---|---|
| release | Name of the batch of release of emails; see details. |
| save.dir | Directory where to save the extracted text defaults to getwd() |
| extractor | Full path to pdf extractor pdftotext, see details. |
| ... | additional parameters to pass to pdftotext. |

### Author(s)

John Coene <jcoenep@gmail.com>

### See Also

get_xpdf, download_emails

## Examples

```
## Not run:
# download emails
download_emails("August") # August release

dir.create("emails_pdf") # dir to extract zip

unzip("August.zip", exdir = "./emails_pdf")

# create directory to store extracted contents
dir.create("emails_txt")

ext <- get_xpdf()

extarct_contents(emails = "HRC_Email_296", dest = "./emails_txt", extractor = ext)

## End(Not run)
```

---

`get_com` *get_com: get emails sender's and receiver's name*

---

## Description

Get emails communication

## Usage

```
get_com(emails)

## S3 method for class 'rodham'
get_com(emails)
```

## Arguments

| | |
|---|---|
| `emails` | list of email contents, as returned by `load_emails` |

## Details

Get the sender's and receiver's name.

## Value

data.frame of names and document id.

## Examples

```
## Not run:
emails <- load_emails("emails")
com <- get_com(emails)

## End(Not run)
```

---

`get_content` *get_contents: get original emails senders and receivers' name*

---

## Description

Get emails original communication

## Usage

```
get_content(emails)

## S3 method for class 'rodham'
get_content(emails)
```

## Arguments

| | |
|---|---|
| `emails` | list of email contents, as returned by `load_emails` |

## Details

Get the senders and receivers' name of original email.

## Value

named list (document id) of email contents.

## Examples

```
## Not run:
emails <- load_emails("emails")
contents <- get_content(emails)

## End(Not run)
```

get_date *get_date: get emails date*

## Description

Get emails date

## Usage

```
get_date(emails)

## S3 method for class 'rodham'
get_date(emails)
```

## Arguments

| | |
|---|---|
| emails | list of email contents, as returned by load_emails |

## Details

Get the date on which the emails is received.

## Value

data.frame of dates and document id.

## Examples

```
## Not run:
emails <- load_emails("emails")
dates <- get_date(emails)

## End(Not run)
```

get_emails *Get emails and its contents*

## Description

Get the content of Hillary Rodham Clinton's emails by release.

## Usage

```
get_emails(release, save.dir = getwd(), extractor, ...)
```

## Arguments

| | |
|---|---|
| release | Name of the batch of release of emails; see details. |
| save.dir | Directory where to save the extracted text defaults to getwd() |
| extractor | Full path to pdf extractor pdftotext, see details. |
| ... | additional parameters to pass to pdftotext. |

## Details

Below are the valid values for release; follows the WSJ naming convention.

- Benghazi
- June
- July
- August
- September
- October
- November
- January 7
- January 29
- February 19
- february 29
- December
- Non-disclosure

The extractor argument is the full path to your pdftotext.exe extractor; visit xpdf to download or try get_xpdf which attempts to download and unzip the text to pdf extractor. See examples.

## Value

Fetches email zip file from the WSJ and extract text files in save.dir, returns full path to directory that contains parsed txt files.

## Author(s)

John Coene <jcoenep@gmail.com>

## See Also

get_xpdf, download_emails, extract_emails

## Examples

```
## Not run:
# get xpdf extractor
ext <- get_xpdf()

# create
dir.create("emails")

# get emails released in august
emails_aug <- get_emails(release = "August", save.dir = "./emails",
                     extractor = ext)

# use manually downloaded extractor
# ext <- "C:/xpdfbin-win-3.04/bin64/pdftotext.exe"

# get emails related to Benghazi released in December
emails_bengh <- get_emails(release = "Benghazi", extractor = ext,
                           save.dir = "./emails")

## End(Not run)
```

---

`get_id` *get_id: get emails subjects*

---

## Description

Get emails ID

## Usage

```
get_id(emails)

## S3 method for class 'rodham'
get_id(emails)
```

## Arguments

| | |
|---|---|
| `emails` | list of email contents, as returned by `load_emails` |

## Details

Get emails' document id

## Value

vector of emails' document ids.

## Examples

```
## Not run:
emails <- load_emails("emails")
docids <- get_id(emails)

## End(Not run)
```

---

`get_interest` *get_interest: get emails subjects*

---

## Description

Get emails interest

## Usage

```
get_interest(emails)

## S3 method for class 'rodham'
get_interest(emails)
```

## Arguments

| | |
|---|---|
| `emails` | list of email contents, as returned by `load_emails` |

## Details

Get emails' subjects

## Value

data.frame of emails' interest and document id.

## Examples

```
## Not run:
emails <- load_emails("emails")
subjects <- get_interest(emails)

## End(Not run)
```

get_or *get_or: get emails senders and receivers' name*

## Description

Get emails original communication

## Usage

```
get_or(emails)

## S3 method for class 'rodham'
get_or(emails)
```

## Arguments

| | |
|---|---|
| emails | list of email contents, as returned by load_emails |

## Details

Get the senders and receivers' name of original email.

## Value

data.frame of names and document id.

## Examples

```
## Not run:
emails <- load_emails("emails")
original <- get_or(emails)

## End(Not run)
```

get_subject *get_subject: get emails subjects*

## Description

Get emails subjects

## Usage

```
get_subject(emails)

## S3 method for class 'rodham'
get_subject(emails)
```

### Arguments

| | |
|---|---|
| emails | list of email contents, as returned by load_emails |

### Details

Get emails' subjects

### Value

data.frame of emails' subjects and document id.

### Examples

```
## Not run:
emails <- load_emails("emails")
subjects <- get_subject(emails)

## End(Not run)
```

---

## get_xpdf *Get pdf to text extractor (xpdf)*

---

### Description

Downloads and extracts pdf to text extractor, see details.

### Usage

```
get_xpdf(dest = getwd())
```

### Arguments

| | |
|---|---|
| dest | Destination folder defaults to getwd() |

### Details

If the function fails you can download the extractor manually from http://www.foolabs.com/xpdf/ then set it manually as shown in examples get_emails

Tested on:

- Windows
- Linux

### Value

Returns full path to pdftotext executable

## Author(s)

John Coene <jcoenep@gmail.com>

## See Also

get_emails

---

hrc_names *Hillary Rodham Clinton emails*

---

## Description

List that pairs sender and recipient names provided by the State Department website with that person's commonly-used name.

## Usage

```
hrc_names
```

## Format

A data frame with 912 rows and 2 variables:

**originalName** Original name

**commonName** Commonly-used name in emails

## Details

For example, HRC becomes Hillary Clinton.

## Source

https://github.com/wsjdata/clinton-email-cruncher/blob/master/HRCEMAIL_names.csv

load_emails *load emails from text files*

## Description

Load all emails.

## Usage

```
load_emails(dir)
```

## Arguments

| | |
|---|---|
| dir | directory where txt emails can be found. |

## Value

named list of emails; names are file names without extension.

## Author(s)

John Coene <jcoenep@gmail.com>

## Examples

```
## Not run:
# get xpdf extractor
ext <- get_xpdf()

# create
dir.create("emails")

# get emails released in august
emails_aug <- get_emails(release = "August", save.dir = "./emails",
                     extractor = ext)

# use manually downloaded extractor
# ext <- "C:/xpdfbin-win-3.04/bin64/pdftotext.exe"

# get emails related to Benghazi released in December
emails_bengh <- get_emails(release = "Benghazi", extractor = ext,
                           save.dir = "./emails")

contents <- load_emails(emails_bengh)

## End(Not run)
```

search_emails *Search Rodham's emails*

## Description

Search Hillary Rodham Clinton's *personal* emails.

## Usage

```
search_emails(subject = NULL, to = NULL, from = NULL, start = NULL,
  end = NULL, internal = TRUE)
```

## Arguments

| | |
|---|---|
| subject | Filter by subject, defaults to NULL(no filter). If `internal = TRUE` then matches pattern, if `internal = FALSE` then looks for exact match. |
| to | Filter by Receiver, defaults to NULL(no filter). |
| from | Filter by Sender, defaults to NULL(no filter). |
| start | Filter by date range, defaults to NULL(no filter). |
| end | Filter by date range, defaults to NULL(no filter). |
| internal | if TRUE (default) searches the internal data set (see `data(emails)`), if FALSE fetches the data through the Wall Street journal API. `data(emails)` is equivalent to internal TRUE |

## Details

There are a total of 29444 emails ranging from 2009-08-14 to 2014-08-13, please consider leaving internal to TRUE to not hammer the Wall Street Journal's API. `internal = TRUE` is equivalent to emails.

## Author(s)

John Coene <jcoenep@gmail.com>

## Examples

```
## Not run:
emails <- search_emails()

# only emails on cuba
emails <- search_emails(subject = "Cuba")

# only emails from Jake Sullivan since 2014
j_s <- search_emails(from = "Jake Sullivan", start = as.Date("2014-01-01"))

## End(Not run)
```

tidy_emails *Tidy contents*

## Description

Tidy email contents

## Usage

```
tidy_emails(content)
```

## Arguments

| | |
|---|---|
| content | email content as returned by get_content. |

## Value

A two-column tibble with emails document id in one column (emails) and the email content in another.

## Author(s)

John Coene <jcoenep@gmail.com>

## Examples

```
## Not run:
content <- get_content(content)
content <- clean_content(content)
tidy <- tidy_emails(content)

## End(Not run)
```

# Index