Package 'webtrackR'

March 13, 2023

Title Analysing Web Tracking Data and Online News Behaviour

Version 0.0.1

Description Implements data structures and methods to work with web tracking data. This includes data preprocessing steps, methods to construct audience networks as described in Mangold & Scharkow (2020) <doi:10.1080/19312458.2020.1724274>, and metrics of news audience polarization described in Mangold & Scharkow (2022) <doi:10.1080/19312458.2022.2085249>.

URL https://github.com/schochastics/webtrackR

BugReports https://github.com/schochastics/webtrackR/issues

Depends R (>= 3.2.0)

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

Imports data.table, tibble, igraph, urltools, utils

LazyData true

Suggests backbone, stats, testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author David Schoch [aut, cre] (<https://orcid.org/0000-0003-2952-4812>), Frank Mangold [aut] (<https://orcid.org/0000-0002-9776-3113>), Sebastian Stier [aut] (<https://orcid.org/0000-0002-1217-5778>)

Maintainer David Schoch <david@schochastics.net>

Repository CRAN

Date/Publication 2023-03-13 13:30:02 UTC

R topics documented:

`add_duration` *Add time spent in seconds on webpage*

Description

Derive the time spend on a website from the timestamps

Usage

```
add_duration(wt, reset = 3600)
```

Arguments

`wt`	webtrack data object
`reset`	numeric. If duration is greater than this value, it is reset to zero, assuming a new browsing session has started

Value

webtrack data.table with the same columns as wt and a new column called duration

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
wt <- add_duration(wt)
```

add_panelist_data *Add panelist features to webtrack data Add characteristics of panelists (e.g. from a survey) to the webtrack data*

Description

Add panelist features to webtrack data Add characteristics of panelists (e.g. from a survey) to the webtrack data

Usage

```
add_panelist_data(wt, data, cols = NULL)
```

Arguments

wt	webtrack data object
data	a data.table (or object that can be converted to data.table) which contains variables of panelists
cols	character vector of columns to add. If NULL, all columns are added

Value

webtrack object with the same columns and joined with panelist survey data

Examples

```
data("test_data")
data("test_survey")
wt <- as.wt_dt(test_data)
add_panelist_data(wt,test_survey)
```

aggregate_duration *Aggregate duration of consecutive visits to a website*

Description

Aggregate duration of consecutive visits to a website

Usage

```
aggregate_duration(wt, keep = FALSE)
```

Arguments

wt	webtrack data object
keep	logical. if intermediary columns should be kept or not. defaults to FALSE

Value

webtrack data.table with the same columns as wt with updated duration

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
wt <- add_duration(wt)
wt <- extract_domain(wt)
# the following step can take longer
wt <- wt[1:100,]
aggregate_duration(wt)
```

---

audience_incidence *Create incidence matrix for audience-outlet network*

Description

Create incidence matrix for audience-outlet network

Usage

```
audience_incidence(wt, cutoff = 3)
```

Arguments

wt	webtrack data object
cutoff	visits below this cutoff will not be considered as a visit

Details

The incidence matrix is a matrix A with entries `A[i,j]=1` if panelist i visited outlet j at least once.

Value

incidence audience-outlet network

See Also

to create audience networks see audience_network

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
wt <- add_duration(wt)
wt <- extract_domain(wt)
audience_incidence(wt)
```

audience_network *Create audience networks*

Description

audience network

Usage

```
audience_network(wt, cutoff = 3, type = "pmi", alpha = 0.05)
```

Arguments

wt	webtrack data object
cutoff	visits below this cutoff will not be considered as a visit
type	one of "pmi", "phi", "disparity", "sdsm, "or "fdsm".
alpha	significance level

Value

audience network as igraph object

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
wt <- add_duration(wt)
wt <- extract_domain(wt)
audience_network(wt, type = "pmi", cutoff = 120)
```

bakshy *Bakshy Top500 Ideological alignment of 500 domains based on facebook data*

Description

Bakshy Top500 Ideological alignment of 500 domains based on facebook data

Usage

```
bakshy
```

Format

An object of class `data.table` (inherits from `data.frame`) with 500 rows and 7 columns.

References

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. "Exposure to ideologically diverse news and opinion on Facebook." Science 348.6239 (2015): 1130-1132.

`classify_domains` *Classify domains according to prespecified classes*

Description

Classify domains according to prespecified classes

Usage

```
classify_domains(
  wt,
  domain_classes = NULL,
  prev_type = TRUE,
  preprocess_newsportals = FALSE,
  return.only = NULL
)
```

Arguments

`wt`	webtrack data object
`domain_classes`	a data.table containing a column "domain" and "type". If NULL, an internal list is used
`prev_type`	logical. If TRUE (default) the type of the domain visited before the current visit is added
`preprocess_newsportals`	logical. add suffix "NEWS" to domains which are classified as portals. If TRUE there needs to be a domain type "newsportals"
`return.only`	if not null, only return the specified domain type

Value

webtrack data.table with the same columns as wt and a new column called type. If prev_type is TRUE, a column prev_type is added with the type of the visit before the current one. If newsportals are processed, found newsportals have an added "/NEWS" in the domain column. If return.only is used, only rows that contain a specific domain type are returned

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
wt <- extract_domain(wt)
wt <- add_duration(wt)
wt <- classify_domains(wt)
```

create_urldummy *Create an urldummy variable from a data.table object*

Description

Create an urldummy variable from a data.table object

Usage

```
create_urldummy(wt, dummy, name)
```

Arguments

wt	webtrack data object
dummy	a vector of urls that should be dummy coded
name	name of dummy variable to create.

Value

webtrack object with the same columns and a new column called "name" including the dummy variable

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
wt <- extract_domain(wt)
code_urls <- c("Ccj4QELzbJe6.com/FrKrkvugBVJWwfSobV")
create_urldummy(wt,dummy = code_urls, name = "test_dummy")
```

domain_list *Domain list classification of domains into news,portals, search, and social media*

Description

Domain list classification of domains into news,portals, search, and social media

Usage

```
domain_list
```

Format

An object of class `data.table` (inherits from `data.frame`) with 663 rows and 2 columns.

References

Stier, S., Mangold, F., Scharkow, M., & Breuer, J. (2022). Post Post-Broadcast Democracy? News Exposure in the Age of Online Intermediaries. American Political Science Review, 116(2), 768-774.

`extract_domain` *Extract domain from url*

Description

Extracts the domain and subdomain from the urls

Usage

```
extract_domain(wt)
```

Arguments

`wt`	webtrack data object

Value

webtrack data.table with the same columns as wt and a new column called domain

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
wt <- extract_domain(wt)
```

`isolation_index` *Isolation Index In terms of news exposure, the isolation index captures the extent to which conservatives disproportionately visit outlets whose other visitors are conservative*

Description

Isolation Index In terms of news exposure, the isolation index captures the extent to which conservatives disproportionately visit outlets whose other visitors are conservative

Usage

```
isolation_index(left, right)
```

Arguments

left	vector (usually corresponds to a column in a webtrack data.table) indicating the number of left leaning individuals using an outlet
right	vector (usually corresponds to a column in a webtrack data.table) indicating the number of right leaning individuals using an outlet

Details

a value of 1 indicates that left leaning and right leaning users do not have any outlet overlap. A value of 0 means both use exactly the same outlets

Value

numeric value between 0 and 1. 0 indicates no isolation and 1 perfect isolation

References

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. "The rise and decline of the American ghetto." Journal of political economy 107.3 (1999): 455-506. Gentzkow, Matthew, and Jesse M. Shapiro. "Ideological segregation online and offline." The Quarterly Journal of Economics 126.4 (2011): 1799-1839.

Examples

```
# perfect isolation
left <-  c(5,5,0,0)
right <- c(0,0,5,5)
isolation_index(left,right)

#perfect overlap
left <-  c(5,5,5,5)
right <- c(5,5,5,5)
isolation_index(left,right)
```

---

news_types *News Types*

Description

Classification of domains into different news types

Usage

```
news_types
```

Format

An object of class `data.table` (inherits from `data.frame`) with 690 rows and 2 columns.

References

Stier, S., Mangold, F., Scharkow, M., & Breuer, J. (2022). Post Post-Broadcast Democracy? News Exposure in the Age of Online Intermediaries. American Political Science Review, 116(2), 768-774.

`print.wt_dt` *Print web tracking data*

Description

Print web tracking data

Usage

```
## S3 method for class 'wt_dt'
print(x, ...)
```

Arguments

`x`	object of class wt_dt
`...`	additional parameters for print

Value

No return value, called for side effects

`summary.wt_dt` *Summary function for webtrack data*

Description

Summary function for webtrack data

Usage

```
## S3 method for class 'wt_dt'
summary(object, ...)
```

Arguments

`object`	object of class wt_dt
`...`	additional parameters for summary

Value

No return value, called for side effects

test_data *Test data*

Description

Randomly generated webtrack data only used for illustrative purposes

Usage

```
test_data
```

Format

An object of class `data.table` (inherits from `data.frame`) with 45290 rows and 3 columns.

test_survey *Test survey*

Description

Randomly generated survey data only used for illustrative purposes

Usage

```
test_survey
```

Format

An object of class `data.table` (inherits from `data.frame`) with 9 rows and 2 columns.

vars_exist *Check if columns are present*

Description

checks if the required columns are present in the webtrack data

Usage

```
vars_exist(wt, vars = c("panelist_id", "url", "timestamp"))
```

Arguments

`wt`	webtrack data as data.table object
`vars`	character vector of variables

Value

A data.table object

`wt_dt` *An S3 class, based on data.table, to store webtrack data*

Description

An S3 class, based on data.table, to store webtrack data

Usage

```
as.wt_dt(x)

is.wt_dt(x)
```

Arguments

`x`	data.table containing the correct set of variables (panelist_id,url and timestamp)

Details

A `wt_dt` table is a data.table. Therefore, it can be used by any function that would work on a data.frame or a data.table. Most of the operation such as variable creation, subsetting and joins are inherited from the data.table `[]` operator, following the convention `DT[i,j,by]` (see data table package for detail). These operations are applied on the data.

Value

a webtrack data object

logical. TRUE if x is a webtrack data object and FALSE otherwise

See Also

- data.table – on which `wt_dt` is based

Examples

```
data("test_data")
wt <- as.wt_dt(test_data)
is.wt_dt(wt)
```

Index