Package 'rdomains'

January 15, 2022

Title Get the Category of Content Hosted by a Domain

Version 0.2.1

Description Get the category of content hosted by a domain. Use Shallalist <http://shalla.de/>, Virustotal (which provides access to lots of services) <https://www.virustotal.com/>, Alexa <https://aws.amazon.com/awis/>, DMOZ <https://curlie.org/>, University Domain list <https://github.com/Hipo/university-domains-list> or validated machine learning classifiers based on Shallalist data to learn about the kind of content hosted by a domain.

Depends R (>= 4.0.0)

Imports Matrix, urltools, glmnet, stats, methods, XML, httr, xml2, curl, virustotal, aws.alexa, jsonlite, devtools, R.utils

Suggests testthat, rmarkdown, knitr (>= 1.11), lintr

VignetteBuilder knitr

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.2

NeedsCompilation no

Author Gaurav Sood [aut, cre]

Maintainer Gaurav Sood <gsood07@gmail.com>

Repository CRAN

Date/Publication 2022-01-15 12:02:41 UTC

R topics documented:

rdomains-package 2
adult_ml1_cat 2
alexa_cat 3
brightcloud_cat 4
dmoz_cat 5
get_alexa_data 5
get_dmoz_data 6

get_shalla_data 7
glm_shalla 7
not_news 8
shalla_cat 9
uni_cat 9
virustotal_cat 10

Index **11**

`rdomains-package` *rdomains: Classify Domains by their Content*

Description

Want to know what kind of content is carried on a domain? Get the results quickly using rdomains. The package provides access to virustotal API, shalla, brightcloud, aws, and validated ML model based off shallalist data to predict content of a domain.

To learn how to use rdomains, see this vignette: `../doc/rdomains.html`.

Author(s)

Gaurav Sood

`adult_ml1_cat` *Probability that Domain Hosts Adult Content Based on features of Domain Name and Suffix alone.*

Description

Uses a validated ML model that uses keywords in the domain name and suffix to predict probability that the domain hosts adult content. For more information see `https://github.com/themains/keyword_porn`

Usage

```
adult_ml1_cat(domains = NULL)
```

Arguments

`domains`	required; string; vector of domain names

Value

data.frame with original list and content category of the domains

Examples

```
## Not run:
adult_ml1_cat("http://www.google.com")

## End(Not run)
```

---

alexa_cat *Get Category from Alexa*

Description

To learn how to get the Access Key ID and Secret Access Key, see https://docs.aws.amazon.com/general/latest/gr/aws-sec-cred-types.html, clicking on the username followed by security credentials. Either pass the access key and secret or set two environmental variables AWS_ACCESS_KEY_ID and AWS_SECRET_ACCESS_KEY. These environment variables persist within a R session.

Usage

```
alexa_cat(domain = NULL, key = NULL, secret = NULL)
```

Arguments

domain	domain name
key	Alexa Access Key ID
secret	Alexa Secret Access Key

Value

data.frame with 2 columns Title and AbsolutePath

References

https://docs.aws.amazon.com/AlexaWebInfoService/latest/

Examples

```
## Not run:
alexa_cat(domain = "http://www.google.com")

## End(Not run)
```

`brightcloud_cat` *Get Category from Brightcloud*

Description

Returns category of content from Brighcloud

Usage

```
brightcloud_cat(domain = NULL, key = NULL, secret = NULL)
```

Arguments

`domain`	domain name
`key`	brightcloud API consumer key
`secret`	brightcloud API consumer secret

Details

Get the API Consumer Key and Secret from http://www.brightcloud.com/.

Value

named list

References

http://www.brightcloud.com/

Examples

```
## Not run:
brightcloud_cat("http://www.google.com", key = "XXXX", secret = "XXXX")

## End(Not run)
```

dmoz_cat *Get Category from DMOZ*

Description

Fetches category (or categories) of content hosted by a domain according to DMOZ. The function checks if path to the DMOZ file is provided by the user. If not, it looks for `dmoz_domain_cateory.csv` in the working directory. It also returns results for prominent subdomains.

Usage

```
dmoz_cat(domains = NULL, use_file = NULL)
```

Arguments

`domains`	vector of domain names
`use_file`	path to the dmoz file, which can be downloaded using `get_dmoz_data`

Value

data.frame with original list and content category of the domain

Examples

```
## Not run:
dmoz_cat(domains = "http://www.google.com")
dmoz_cat(domains = c("http://www.google.com", "http://plus.google.com"))

## End(Not run)
```

get_alexa_data *Get Alexa Traffic Data*

Description

Get Top 1M most visited domains list from Alexa. These data can be used to weight the classification error.

Usage

```
get_alexa_data(outdir = ".", overwrite = FALSE)
```

Arguments

`outdir`	Optional; folder to which you want to save the file; Default is same folder
`overwrite`	Optional; default is FALSE. If TRUE, the file is overwritten.

References

https://aws.amazon.com/marketplace/pp/B07QK2XWNV

Examples

```
## Not run:
get_alexa_data()

## End(Not run)
```

`get_dmoz_data` *Get DMOZ Data*

Description

Downloads, unzips and saves archived version of the DMOZ data. For more details, check: https://github.com/themains/rdomains/tree/master/data-raw/dmoz/

Usage

```
get_dmoz_data(outdir = ".", overwrite = FALSE)
```

Arguments

`outdir`	Optional; folder to which you want to save the file; Default is same folder
`overwrite`	Optional; default is FALSE. If TRUE, the file is overwritten.

References

https://dmoztools.net

Examples

```
## Not run:
get_dmoz_data()

## End(Not run)
```

get_shalla_data *Get Shalla Data*

Description

Shalla has discontinued. We downloaded the last copy (1/14/22). For more information see data-raw folder on github Downloads, unzips and saves the latest version of shallalist data. By default, saves shalla data as `shalla_domain_category.csv`.

Usage

```
get_shalla_data(outdir = "./", overwrite = FALSE)
```

Arguments

`outdir`	Optional; folder to which you want to save the file; Default is same folder
`overwrite`	Optional; default is FALSE. If TRUE, the file is overwritten.

References

http://www.shallalist.de/

Examples

```
## Not run:
get_shalla_data()

## End(Not run)
```

glm_shalla *ML Model*

Description

ML Model

Usage

```
glm_shalla
```

Format

A list

Author(s)

Gaurav Sood

Source

ML model based on shallalist using keywords and domain suffixes,

---

`not_news` *Classify News and Non-News Based on keywords in the URL*

Description

Based on a slightly amended version of the regular expression used to classify news, and non-news in: "Exposure to ideologically diverse news and opinion on Facebook" by Bakshy, Messing, and Adamic. Science. 2015.

Usage

```
not_news(url_list = NULL)
```

Arguments

`url_list`	vector of URLs

Details

Amendment: sport rather than sports

URL containing any of the following words is classified as soft news: "sport|entertainment|arts|fashion|style|lifestyle|leisure|ce

URL containing any of following words is classified as hard news: "politi|usnews|world|national|state|elect|vote|govern|campa

Note that it is based on patterns existing in a small set of domains. See paper for details.

Value

data.frame with 3 columns: url, not_news, news

References

https://www.science.org/doi/10.1126/science.aaa1160

Examples

```
## Not run:
not_news("http://www.bbc.com/sport")
not_news(c("http://www.bbc.com/sport", "http://www.washingtontimes.com/news/politics/"))

## End(Not run)
```

shalla_cat *Get Category from Shallalist*

Description

Fetches category of content hosted by a domain according to Shalla. The function checks if path to the shalla file is provided by the user. If not, it looks for `shalla_domain_category.csv` in the working directory.

Usage

```
shalla_cat(domains = NULL, use_file = NULL)
```

Arguments

`domains`	vector of domain names
`use_file`	path to the latest shallalist file downloaded using `get_shalla_data`

Value

data.frame with original list and content category of the domain

Examples

```
## Not run:
shalla_cat(domains = "http://www.google.com")

## End(Not run)
```

uni_cat *Get Category from University Domain List*

Description

Fetches university domain json from: https://raw.githubusercontent.com/Hipo/university-domains-list/master/world_universities_and_domains.json

Usage

```
uni_cat(domains = NULL)
```

Arguments

`domains`	vector of domain names

Value

data.frame with original list and all the other columns from the university json

Examples

```
## Not run:
uni_cat(domains = "http://www.google.com")

## End(Not run)
```

`virustotal_cat` *Get Category from Virustotal*

Description

Returns category of content from 6 major services including: BitDefender, Dr. Web, Alexa (DMOZ), Google, Websense, and Trendmicro. Not all services will have categories for all the domains. When the categories are not returned for a particular domain, we return a NA.

Usage

```
virustotal_cat(domain = NULL, apikey = NULL)
```

Arguments

`domain`	domain name
`apikey`	virustotal API key

Details

Get the API Access Key from http://www.virustotal.com/. Either pass the API Key to the function or set the environmental variable: `VirustotalToken`. Environment variables persist within a R session.

Value

data.frame with 7 columns: domain, bitdefender, dr_web, alexa, google, websense, trendmicro

References

https://developers.virustotal.com/v2.0/reference

Examples

```
## Not run:
virustotal_cat("http://www.google.com")

## End(Not run)
```

Index

∗ **keywords**
    glm_shalla, 7
∗ **model**
    glm_shalla, 7

adult_ml1_cat, 2
alexa_cat, 3

brightcloud_cat, 4

dmoz_cat, 5

get_alexa_data, 5
get_dmoz_data, 5, 6
get_shalla_data, 7, *9*
glm_shalla, 7

not_news, 8

rdomains *(*rdomains-package*)*, 2
rdomains-package, 2

shalla_cat, 9

uni_cat, 9

virustotal_cat, 10