Package 'rdomains'

May 13, 2026

Title Get the Category of Content Hosted by a Domain

Version 0.4.0

Description Get the category of content hosted by a domain. Use Shallalist (service discontinued), 'VirusTotal' (which provides access to lots of services) <https://www.virustotal.com/>, 'DMOZ' <https://archive.org/details/dmoz-rdf-20150327>, University Domain list <https://github.com/Hipo/university-domains-list>, 'OpenAI' 'GPT' models, 'Anthropic' 'Claude' models, or validated machine learning classifiers based on 'Shallalist' data to learn about the kind of content hosted by a domain.

Depends R (>= 4.1.0)

Imports Matrix, urltools, glmnet, stats, methods, XML, httr, xml2, curl, virustotal, jsonlite, R.utils, dplyr (>= 1.1.0), purrr (>= 1.0.0), tibble (>= 3.2.0), stringr (>= 1.5.0), rlang (>= 1.1.0), cli (>= 3.6.0), checkmate (>= 2.3.0), glue (>= 1.6.0), readr (>= 2.1.0)

Suggests testthat, rmarkdown, knitr (>= 1.11)

VignetteBuilder knitr

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.3

NeedsCompilation no

Author Gaurav Sood [aut, cre]

Maintainer Gaurav Sood <gsood07@gmail.com>

Repository CRAN

Date/Publication 2026-05-13 04:50:02 UTC

Contents

`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, aws, OpenAI GPT models, Anthropic Claude models, and validated ML model based off shallalist data to predict content of a domain.

Details

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)
```

---

claude_cat *Get Category from Anthropic Claude*

Description

Fetches category of content hosted by a domain using Anthropic's Claude API. The function uses Claude models to classify domains into specified categories.

Usage

```
claude_cat(
  domains = NULL,
  api_key = NULL,
  categories = NULL,
  model = "claude-3-haiku-20240307",
  rate_limit = 0.5
)
```

Arguments

domains	vector of domain names
api_key	Anthropic API key. If not provided, looks for ANTHROPIC_API_KEY or CLAUDE_API_KEY environment variable
categories	vector of categories to classify into. If NULL, uses default web categories
model	Claude model to use (default: "claude-3-haiku-20240307" for cost efficiency)
rate_limit	delay in seconds between API calls (default: 0.5)

Value

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

Examples

```
## Not run:
claude_cat("google.com")
claude_cat(c("google.com", "facebook.com"))
claude_cat("google.com", categories = c("search", "social", "ecommerce", "news", "other"))

## 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_dmoz_data *Get DMOZ Data*

Description

Downloads archived DMOZ (Open Directory Project) data. DMOZ was discontinued in March 2017. This function downloads our preserved copy of the final DMOZ dataset. 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://archive.org/details/dmoz-rdf-20150327

Examples

```
## Not run:
get_dmoz_data()

## End(Not run)
```

get_shalla_data *Get Shalla Data*

Description

Shallalist service was discontinued in January 2022. This function downloads the last archived copy (from 1/14/22) that we have preserved on GitHub. The original service at shallalist.de is no longer available. Downloads, unzips and saves the final 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

https://web.archive.org/web/20210502020725/http://www.shallalist.de/

Examples

```
## Not run:
get_shalla_data()

## End(Not run)
```

`get_stevenblack_data` *Get Steven Black's Host List Data*

Description

Downloads the latest version of Steven Black's unified hosts file. The hosts file contains domains known for serving ads, malware, and tracking.

Usage

```
get_stevenblack_data(outdir = "./", variant = "base", overwrite = FALSE)
```

Arguments

`outdir`	Optional; folder to which you want to save the file; Default is current directory
`variant`	Optional; which variant to download. Options: "base", "porn", "social", "gambling", "all"
`overwrite`	Optional; default is FALSE. If TRUE, the file is overwritten.

References

https://github.com/StevenBlack/hosts

Examples

```
## Not run:
get_stevenblack_data()
get_stevenblack_data(variant = "all")

## 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)
```

---

openai_cat *Get Category from OpenAI*

Description

Fetches category of content hosted by a domain using OpenAI's chat completion API. The function uses GPT models to classify domains into specified categories.

Usage

```
openai_cat(
  domains = NULL,
  api_key = NULL,
  categories = NULL,
  model = "gpt-4o-mini",
  rate_limit = 0.5
)
```

Arguments

domains	vector of domain names
api_key	OpenAI API key. If not provided, looks for OPENAI_API_KEY environment variable
categories	vector of categories to classify into. If NULL, uses default web categories
model	OpenAI model to use (default: "gpt-4o-mini" for cost efficiency)
rate_limit	delay in seconds between API calls (default: 0.5)

Value

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

Examples

```
## Not run:
openai_cat("google.com")
openai_cat(c("google.com", "facebook.com"))
openai_cat("google.com", categories = c("search", "social", "ecommerce", "news", "other"))

## 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)
```

`stevenblack_cat` *Get Category from Steven Black's Host List*

Description

Classifies domains based on Steven Black's unified host list which blocks ads, malware, and tracking domains. The function checks if a domain appears in the blocklist and categorizes it accordingly.

Usage

```
stevenblack_cat(domain = NULL, use_file = NULL)
```

Arguments

`domain`	domain names as character vector
`use_file`	path to a local Steven Black hosts file. If NULL, downloads from GitHub

Details

Steven Black's host list is a consolidated list from multiple sources including adaway.org, mvps.org, malwaredomainlist.com, and someonewhocares.org.

Value

data.frame with original domain name and category

References

https://github.com/StevenBlack/hosts

Examples

```
## Not run:
stevenblack_cat("doubleclick.net")
stevenblack_cat(c("google.com", "googleadservices.com", "malware-example.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 multiple security vendors using the VirusTotal API v3. The function retrieves domain analysis results including categories from various security services. Not all services will have categories for all domains.

Usage

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

Arguments

domains	domain names as character vector
apikey	virustotal API key

Details

Get the API Access Key from https://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 domain and VirusTotal analysis results

References

https://docs.virustotal.com/reference/domains

Examples

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

## End(Not run)
```

Index