Package: ojoregex (via r-universe)

May 26, 2026

Title Regex Categorization Tools for OJO Analysts

Version 0.10.1

Description This package houses all the regex strings we use in our work. The main functionality cleaning charge descriptions from raw OSCN data.

License GPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Imports cli, dplyr, lubridate, purrr, readr, rlang, stringi, stringr

Suggests googlesheets4, gt, here, knitr, ojodb, rmarkdown, testthat (>= 3.0.0), usethis

Depends R (>= 4.1.0)

Remotes openjusticeok/ojodb

LazyData true

URL https://openjusticeok.github.io/ojoregex/

VignetteBuilder knitr

Config/testthat/edition 3

Config/pak/sysreqs libicu-dev libx11-dev

Repository https://openjusticeok.r-universe.dev

Date/Publication 2025-10-28 16:21:52 UTC

RemoteUrl https://github.com/openjusticeok/ojoregex

RemoteRef HEAD

RemoteSha eab8a8a23209ef38ec8e58185f2380238f91f0b8

Contents

`ojo_add_controlling_charges` *Add controlling charges to the dataset*

Description

This function processes a dataset of charges and adds columns to classify the maximum sentence types and calculate control ranks based on the severity of the sentences.

Usage

```
ojo_add_controlling_charges(ojo_regex_cats)
```

Arguments

`ojo_regex_cats`	A data frame containing charge information, including columns `max_sentence_first_offense` and `max_sentence_any`.

Value

A data frame with additional columns for sentence classifications and control ranks.

`ojo_apply_regex` *Apply OJO Regex*

Description

This function applies regular expressions patterns to clean and categorize charge descriptions in a given dataset.

Usage

```
ojo_apply_regex(
  data,
  col_to_clean = "count_as_filed",
  .keep_flags = FALSE,
  .include_cats = TRUE,
  .quiet = FALSE
)
```

Arguments

data	A data frame containing the dataset to be processed.
col_to_clean	The name of the column in the dataset containing the charge descriptions to be cleaned and categorized.
.keep_flags	Logical value indicating whether to keep the concept flags generated during processing. Defaults to FALSE, which returns only the cleaned dataset without the flags.
.include_cats	Logical value indiciating whether the categories / subcategories should be included in the returned data
.quiet	Should the progress bar be shown?

Value

A cleaned and categorized dataset with charge descriptions in the specified column, along with any additional columns present in the original dataset.

Examples

```
## Not run:
# Load example dataset
data(example_data)

# Apply OJO Regex to clean and categorize charge descriptions
cleaned_data <- apply_ojo_regex(data = example_data, col_to_clean = "charge_description")

## End(Not run)
```

---

ojo_get_flag_regex *Return OJO Regex for a given flag*

Description

This function returns the regex string for a given flag

Usage

```
ojo_get_flag_regex(flag = NA)
```

Arguments

flag	The flag you want to get the regex pattern for

Value

A string of regex

---

ojo_get_statute_regex *Return OJO Regex for a given statute*

Description

This function returns the regex string for a given statute

Usage

```
ojo_get_statute_regex(statute = NA)
```

Arguments

statute	The statute you want to get the regex pattern for

Value

A string of regex

---

ojo_regex_cats *OJO Regex Categories dataset*

Description

OJO Regex Categories dataset

Usage

```
ojo_regex_cats
```

Format

A data frame with X rows and 16 columns:

in_ojoregex Description of in_ojoregex

clean_charge_description Description of clean_charge_description

category Description of category

subcategory Description of subcategory

title Description of title

statutes Description of statutes

chapter Description of chapter

gist Description of gist

description Description of description

statute_link Description of statute_link

cf_cm Description of cf_cm

cf_cm_notes Description of cf_cm_notes

control_rank Description of control_rank

max_sentence_any Description of max_sentence_any

max_sentence_first_offense Description of max_sentence_first_offense

outdated Description of outdated

notes Description of notes

sq780_status Description of sq780_status

violent_crimes_list Description of violent_crimes_list

---

ojo_regex_flags *OJO Regex Flags Dataset*

Description

OJO Regex Flags Dataset

Usage

```
ojo_regex_flags
```

Format

A data frame with 139 rows and 8 columns:

flag Description of column1

regex Description of column2

group Description of column3

criminal_or_civil Description of column4

genre Description of column5

word_boundary Description of column6

examples Description of column7

notes Description of column8 ...

---

ojo_regex_unhoused *Unhoused / Homeless address regex*

Description

Detects whether an address is likely to indicate "homeless" / "unhoused" / "NA", etc.

Usage

```
ojo_regex_unhoused
```

Format

An object of class `character` of length 1.

---

regex_pre_clean *Pre clean charge descriptions to be matched*

Description

This function pre-cleans charge descriptions to be matched by removing specific patterns that are not relevant for matching. It removes phrases like "in concert with" from the end of the charge descriptions.

Usage

```
regex_pre_clean(count_as_filed)
```

Arguments

`count_as_filed` A character vector containing the charge descriptions to be pre-cleaned.

Value

A character vector with pre-cleaned charge descriptions.

Examples

```
## Not run:
# Example usage
clean_text <- regex_pre_clean("TAXS, FAIL TO DISPLAY TAX STAMP ON CDS IN CONCERT W/J POOLE")
clean_text

## End(Not run)
```

Index