Package 'medExtractR'

July 31, 2020

Title Extraction of Medication Information from Clinical Text

Version 0.2

Description Function and support for medication and dosing information extraction from free-text clinical notes. Medication entities that can be extracted include drug name, strength, dose amount, dose, frequency, intake time, and time of last dose.

License GPL (>= 2)

Depends R (>= 2.10)

Encoding UTF-8

LazyData true

Imports stringr

RoxygenNote 7.1.1

NeedsCompilation no

Author Hannah Weeks [aut, cre],
Cole Beck [aut] (<https://orcid.org/0000-0002-6849-6255>),
Leena Choi [aut]

Maintainer Hannah Weeks <hannah.l.weeks@vanderbilt.edu>

Repository CRAN

Date/Publication 2020-07-31 18:40:03 UTC

R topics documented:

medExtractR-package *Medication Extraction With R*

Description

Provides a function `medExtractR` for extracting dose attributes for medications within a given electronic health record (EHR) note.

Author(s)

Hannah Weeks <hannah.l.weeks@vanderbilt.edu>,
Cole Beck <cole.beck@vumc.org>,
Leena Choi <leena.choi@vumc.org>

Maintainer: Hannah Weeks <hannah.l.weeks@vanderbilt.edu>

Examples

```
note1 <- "Progrf Oral Capsule 1 mg 3 capsules by mouth twice a day - last
dose at 10pm"
note2 <- "Currently on lamotrigine 150-200, but will increase to lamotrigine 200mg bid"
medExtractR(note1, c("prograf", "tacrolimus"), 60, "mg", 2, lastdose=TRUE)
medExtractR(note2, c("lamotrigine", "ltg"), 130, "mg", 1, strength_sep = "-")
```

dosechange_vals *Keywords Specifying Dose Change*

Description

Vector of keywords indicating a dose change, meaning that the associated drug regimen may not be current. This includes phrases such as increase, reduce, or switch. In the following example of clinical text, the word 'increase' represents a dose change keyword: "Increase prograf to 5mg bid."

Usage

```
dosechange_vals
```

Format

A vector with 18 character strings.

Examples

```
data(dosechange_vals)
```

`extract_entities` *Extract Medication Entities From Phrase*

Description

This function searches a phrase for medication dosing entities of interest. It is called within `medExtractR` and generally not intended for use outside that function.

Usage

```
extract_entities(
  phrase,
  p_start,
  p_stop,
  unit,
  freq_fun = NULL,
  intaketime_fun = NULL,
  strength_sep = NULL,
  ...
)
```

Arguments

`phrase`	Text to search.
`p_start`	Start position of phrase within original text.
`p_stop`	End position of phrase within original text.
`unit`	Unit of measurement for medication strength, e.g. 'mg'.
`freq_fun`	Function used to extract frequency.
`intaketime_fun`	Function used to extract intaketime.
`strength_sep`	Delimiter for contiguous medication strengths.
`...`	Parameter settings used in extracting frequency and intake time, including additional arguments to `freq_fun` and `intaketime_fun`. Use `freq_dict` to identify custom frequency dictionaries and `intaketime_dict` to identify custom intake time dictionaries.

Details

Various medication dosing entities are extracted within this function including the following:

strength: The strength of an individual unit (i.e. tablet, capsule) of the drug.
dose amount: The number of tablets, capsules, etc taken with each dose.
dose: The total strength given intake. This quantity would be equivalent to strength x dose amount, and appears similar to strength when dose amount is absent.
frequency: The number of times per day a dose is taken, e.g. "once daily" or '2x/day'.
intaketime: The time period of the day during which a dose is taken, e.g. 'morning', 'lunch', 'in the

pm'.

Strength, dose amount, and dose are primarily numeric quantities, and are identified using a combination of regular expressions and rule-based approaches. Frequency and intake time, on the other hand, use dictionaries for identification.

By default and when `freq_fun` and/or `intaketime_fun` are `NULL`, the `extract_generic` function will be used for these entities.

The `stength_sep` argument is `NULL` by default, but can be used to identify shorthand for morning and evening doses. For example, consider the phrase "Lamotrigine 300-200" (meaning 300 mg in the morning and 200 mg in the evening). The argument `strength_sep = '-'` identifies the full expression *300-200* as *dose* in this phrase.

Value

data.frame with entities information. At least one row per entity is returned, using `NA` when no expression was found for a given entity.
Sample output for the phrase "Lamotrigine 200mg bid" would look like:

entity	expr
IntakeTime	<NA>
Strength	<NA>
DoseAmt	<NA>
Frequency	bid;19:22
Dose	200mg;13:18

Examples

```
note <- "Lamotrigine 25 mg tablet - 3 tablets oral twice daily"
extract_entities(note, 1, nchar(note), "mg")
# A user-defined dictionary can be used instead of the default
my_dictionary <- data.frame(c("daily", "twice daily"))
extract_entities(note, 1, 53, "mg", freq_dict = my_dictionary)
```

---

extract_generic *Extract Generic Entities From Phrase*

Description

This function searches a phrase for the position and length of expressions specified in a dictionary.

Usage

```
extract_generic(phrase, dict)
```

Arguments

phrase	Text to search.
dict	data.frame, the first column should contain expressions to find. These can be regular expressions or exact phrases.

Details

extract_generic is used to extract entities that are identified with an associated dictionary of phrases or regular expressions, such as frequency or intake time in medExtractR. This function is called within extract_entities.

Value

A numeric matrix with position and expression length.

Examples

```
data(freq_vals)
extract_generic("take two every day", freq_vals)
extract_generic("take two every morning",
                 data.frame(c("morning", "every morning")))
```

---

extract_lastdose *Extract Last Dose Time From Phrase*

Description

This function searches a phrase for the expression and postion of the time at which the last dose of a drug was taken. It is called within medExtractR and generally not intended for use outside that function.

Usage

```
extract_lastdose(phrase, p_start, d_start, d_stop, time_exp = "default")
```

Arguments

phrase	Text to search.
p_start	Start position of phrase in larger text.
d_start	Start position of drug name in larger text.
d_stop	Start position of drug name in larger text.
time_exp	Vector of regular expressions to identify time expressions.

Details

This function identifies the time at which the last dose of a drug of interest was taken. The arguments `p_start`, `d_start`, and `d_stop` represent global start or stop positions for the phrase or drug. These arguments are used to determine the position of any found last dose time expressions relative to the overall clinical note, not just within `phrase`.

The `time_exp` argument contains regular expressions for numeric or text representations of last dose time. See `time_regex` for more information about the default regular expressions used in `medExtractR`.

Value

data.frame with last dose time entity information. This output format is consistent with the output of `extract_entities`
Sample output for the phrase "Last prograf at 5pm" would look like:

entity	expr
LastDose	5pm;17:20

Examples

```
# Suppose this phrase begins at character 120 in the overall clinical note
extract_lastdose("took aspirin last night at 8pm", p_start = 120,
                 d_start = 125, d_stop = 131)
```

`freq_vals` *Keywords Specifying Frequency*

Description

A dictionary mapping frequency expressions to numeric values representing the corresponding number of doses per day. The form of each frequency is given as a regular expression.

Usage

```
freq_vals
```

Format

A data frame with 77 observations on the following variables.

expr A character vector, expressions to consider as frequency

value A numeric vector, numeric value of frequency

Examples

```
data(freq_vals)
```

intaketime_vals *Keywords Specifying Intake Time*

Description

A dictionary mapping intake time expressions to numeric values representing the corresponding number of doses. The form of each intake time is given as a regular expression.

Usage

```
intaketime_vals
```

Format

A data frame with 23 observations on the following variables.

expr A character vector, expressions to consider as intake time

value A numeric vector, numeric value of intake time

Examples

```
data(intaketime_vals)
```

medExtractR *Extract Medication Entities From Clinical Note*

Description

This function identifies medication entities of interest and returns found expressions with start and stop positions.

Usage

```
medExtractR(
  note,
  drug_names,
  window_length,
  unit,
  max_dist = 0,
  drug_list = "rxnorm",
  lastdose = FALSE,
  lastdose_window_ext = 1.5,
  strength_sep = NULL,
  flag_window = 30,
  dosechange_dict = "default",
  ...
)
```

Arguments

note	Text to search.
drug_names	Vector of drug names to locate.
window_length	Length (in number of characters) of window after drug in which to look.
unit	Strength unit to look for (e.g., 'mg').
max_dist	Numeric - edit distance to use when searching for drug_names.
drug_list	Vector of known drugs that may end search window. By default calls rxnorm_druglist.
lastdose	Logical - whether or not last dose time entity should be extracted.
lastdose_window_ext	Numeric - multiplicative factor by which window_length should be extended when identifying last dose time.
strength_sep	Delimiter for contiguous medication strengths (e.g., '-' for "LTG 200-300").
flag_window	How far around drug (in number of characters) to look for dose change keyword - default fixed to 30.
dosechange_dict	List of keywords used to determine if a dose change entity is present.
...	Parameter settings used in extracting frequency and intake time. Potentially useful parameters include freq_dict and intaketime_dict (see ... argument in extract_entities) to specify frequency or intake time dictionaries, as well as 'freq_fun' and 'intaketime_fun' for user-specified extraction functions. See extract_entities documentation for details.

Details

This function uses a combination of regular expressions, rule-based approaches, and dictionaries to identify various drug entities of interest. Specific medications to be found are specified with drug_names, which is not case-sensitive or space-sensitive (e.g., 'lamotrigine XR' is treated the same as 'lamotrigineXR'). Entities to be extracted include drug name, strength, dose amount, dose, frequency, intake time, and time of last dose. See extract_entities and extract_lastdose for more details.

When searching for medication names of interest, fuzzy matching may be used. The max_dist argument determines the maximum edit distance allowed for such matches. If using fuzzy matching, any drug name with less than 5 characters will only allow an edit distance of 1, regardless of the value of max_dist.

Most medication entities are searched for in a window after the drug. The dose change entity, or presence of a keyword to indicate a non-current drug regimen, may occur before the drug name. The flag_window argument adjusts the width of the pre-drug window.

The stength_sep argument is NULL by default, but can be used to identify shorthand for morning and evening doses. For example, consider the phrase 'Lamotrigine 300-200' (meaning 300 mg in the morning and 200 mg in the evening). The argument strength_sep = '-' identifies the full expression *300-200* as *dose* in this phrase.

By default, the drug_list argument is "rxnorm" which calls data(rxnorm_druglist). A custom drug list in the form of a character string can be supplied instead, or can be appended to

`rxnorm_druglist` by specifying `drug_list = c("rxnorm",custom_drug_list)`. This uses publicly available data courtesy of the U.S. National Library of Medicine (NLM), National Institutes of Health, Department of Health and Human Services; NLM is not responsible for the product and does not endorse or recommend this or any other product. See `rxnorm_druglist` documentation for details.

Value

data.frame with entity information
Sample output:

entity	expr	pos
DoseChange	decrease	66:74
DrugName	Prograf	78:85
Strength	2 mg	86:90
DoseAmt	1	91:92
Frequency	bid	101:104
LastDose	2100	121:125

References

Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011 Jul-Aug;18(4)441-8. doi: 10.1136/amiajnl-2011-000116. Epub 2011 Apr 21. PubMed PMID: 21515544; PubMed Central PMCID: PMC3128404.

Examples

```
note1 <- "Progrf Oral Capsule 1 mg 3 capsules by mouth twice a day - last
dose at 10pm"
note2 <- "Currently on lamotrigine 150-200, but will increase to lamotrigine 200mg bid"
medExtractR(note1, c("prograf", "tacrolimus"), 60, "mg", 2, lastdose=TRUE)
medExtractR(note2, c("lamotrigine", "ltg"), 130, "mg", 1, strength_sep = "-")
```

`rxnorm_druglist` *List of Medications*

Description

A dictionary that contains a vector of medication names, primarily derived from RxNorm.

Usage

```
rxnorm_druglist
```

Format

A vector with 59,333 character strings.

Details

RxNorm is provided by the U.S. National Library of Medicine. This dictionary uses the January 7, 2019 RxNorm files directly downloaded from https://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html.

This list contains ingredient and brand names, cleaned to remove expressions likely to be ambiguous (e.g., 'today' or 'date'). It has also been supplemented with abbreviations for various medications in a manually curated list from Vanderbilt University's Synthetic Derivative.

This product uses publicly available data courtesy of the U.S. National Library of Medicine (NLM), National Institutes of Health, Department of Health and Human Services; NLM is not responsible for the product and does not endorse or recommend this or any other product.

References

Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011 Jul-Aug;18(4)441-8. doi: 10.1136/amiajnl-2011-000116. Epub 2011 Apr 21. PubMed PMID: 21515544; PubMed Central PMCID: PMC3128404.

Examples

```
data(rxnorm_druglist)
```

---

time_regex *Keywords Specifying Time Expressions*

Description

A vector of regular expressions to identify different forms of time expressions for last dose time. This are the default values used in `link{extract_lastdose}`.

Usage

```
time_regex
```

Format

A vector with 5 regular expressions for the following categories.

am/pm Time is indicated by the presence of 'am' or 'pm' following a numeric expression.

military Time is given in military time, for unambiguous times of 13:00-23:59.

qualifier_after Am/pm indication is implicit through a qualifying term like 'last night' or 'this morning'. The qualifier occurs after the time, e.g. '10 last night.'

qualifier_before Am/pm indication is implicit through a qualifying term like 'last night' or 'this morning'. The qualifier occurs before the time, e.g. 'last night at 10.'

duration Time (in hours) between the last dose and most recent lab value

Details

Certain expressions which might be considered ambiguous are excluded from the regular expressions presented here. For instance, expressions such as '600' could refer to either 6am or 6pm.

Examples

```
data(time_regex)
```

Index