Building a package that fits into an evolving ecosystem

OCRUG

Emil Hvitfeldt

2019-11-19

Overview

- Open Source Software Development

- My journey

- What I learned

This talk is based on anecdotes, I dearly hope that they generalize!

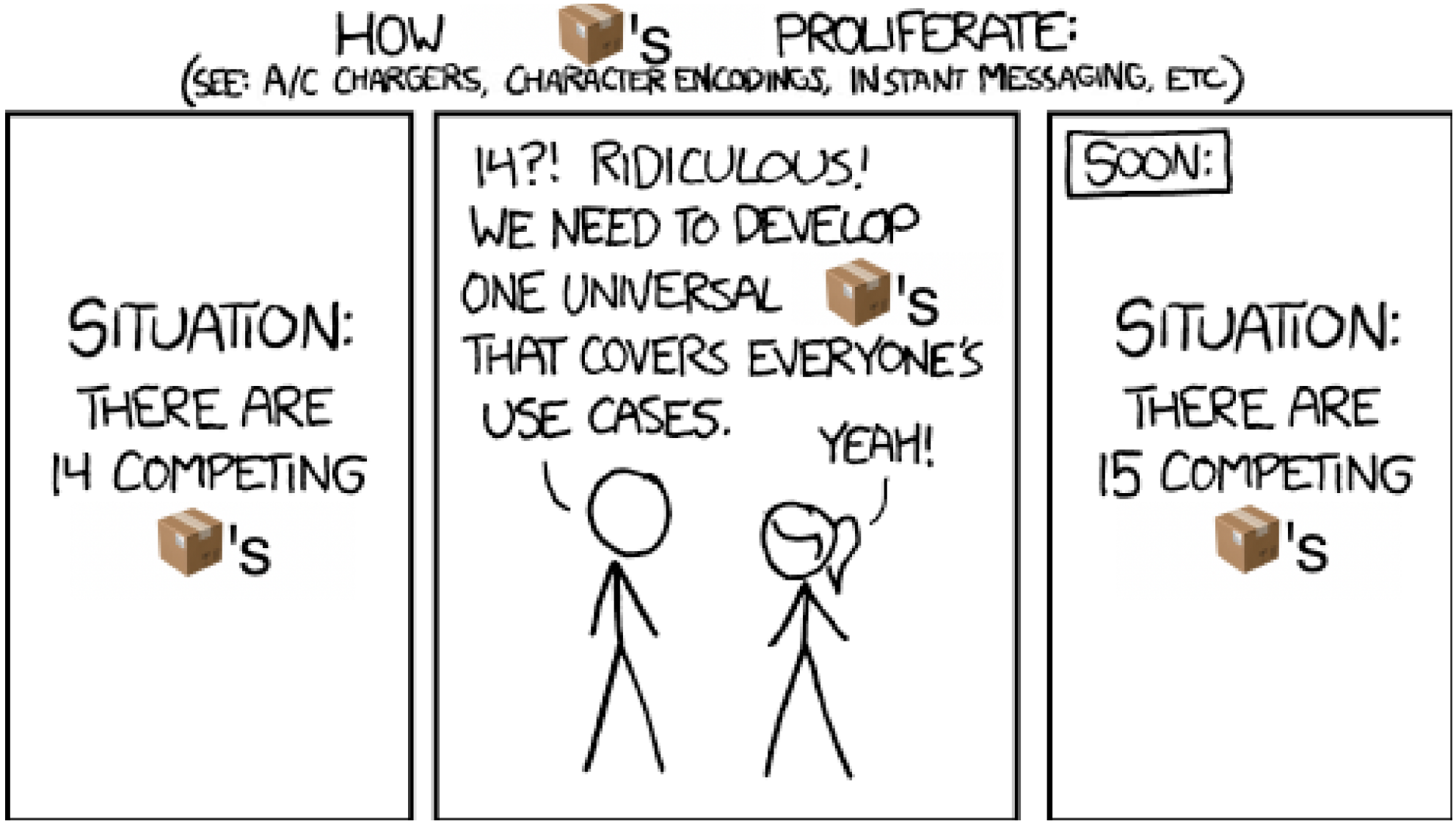
As a new developer it can be hard to find problems that are:

- Easy enough for you do
- Prominent enough that they are worth solving

My advice

- Keep list of ideas
- Google early and often about implementations

Working on a implentation can still be fruitful even if it doesn't end up on CRAN.


HOW 's PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION:
THERE ARE
14 COMPETING
's
14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL 's
THAT COVERS EVERYONE'S
USE CASES.
YEAH!
SOON:
SITUATION:
THERE ARE
15 COMPETING
's

Look for Github tags

Document font size units in theme_bw() etc documentation good first issue help wanted 2

#3458 opened on Jul 26 by hadley

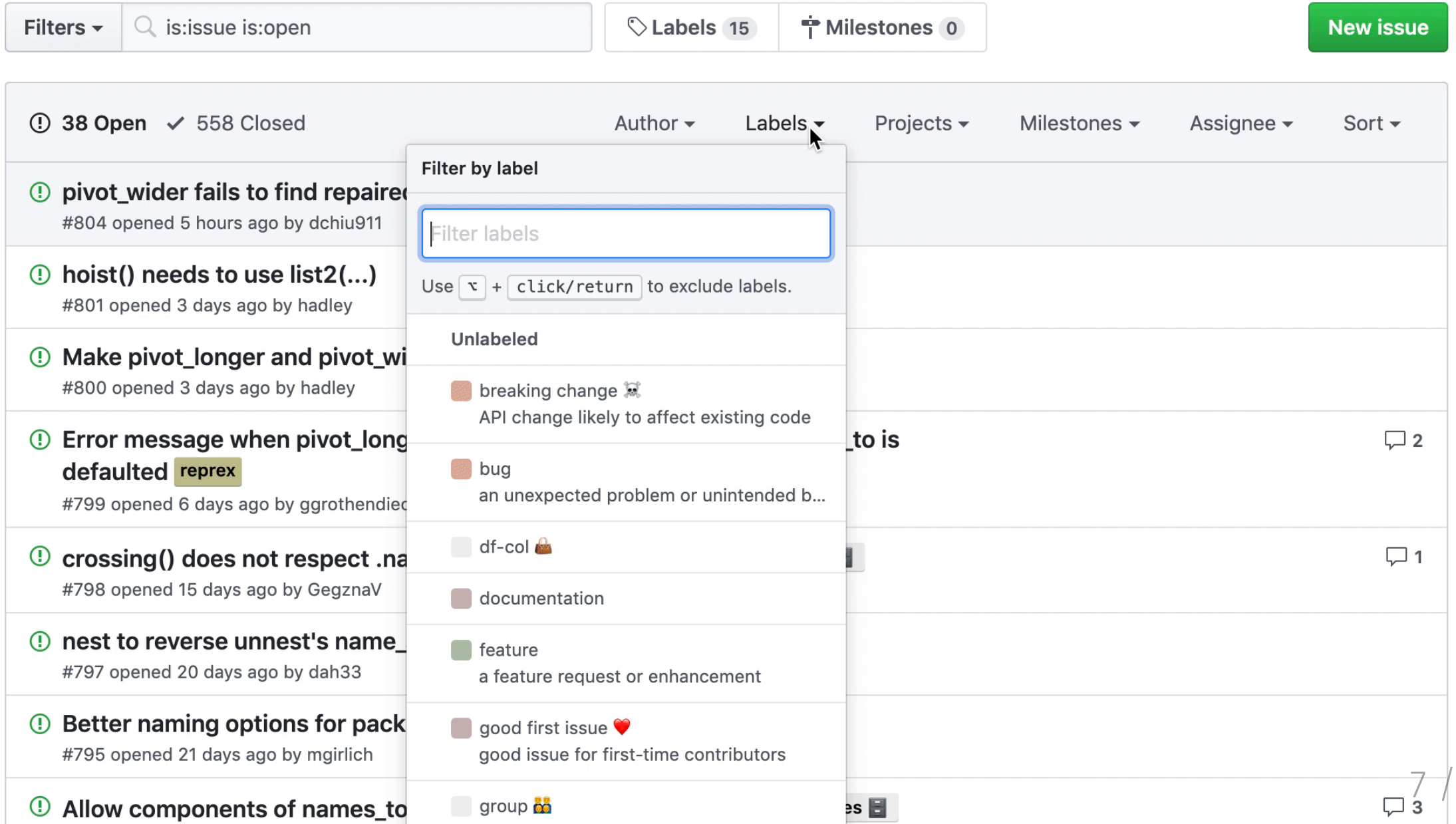
Updating a theme to remove panel.grid.major.x documentation good first issue help wanted 3

#3019 opened on Nov 29, 2018 by llrs

Possibility to create an `as_word()` function? Difficulty: Advanced Effort: High Help Wanted Priority: High Type: Enhancement 6

#121 opened on Jan 3 by rich-iannone

good first issue	good issue for first-time contributors
help wanted	we'd love your help!


Filters
is:issue is:open
Labels 15
Milestones 0
New issue
38 Open
558 Closed
Author
Labels
Projects
Milestones
Assignee
Sort
pivot_wider fails to find repaire
#804 opened 5 hours ago by dchiu911
hoist() needs to use list2(...)
#801 opened 3 days ago by hadley
Make pivot_longer and pivot_wi
#800 opened 3 days ago by hadley
Error message when pivot_long
defaulted reprex
#799 opened 6 days ago by ggrothendiec
_to is
2
crossing() does not respect .na
#798 opened 15 days ago by GegznaV
1
nest to reverse unnest's name_
#797 opened 20 days ago by dah33
Better naming options for pack
#795 opened 21 days ago by mgirlich
Allow components of names_to
3
Filter by label
Filter labels
Use ⌥ + click/return to exclude labels.
Unlabeled
breaking change ☠️
API change likely to affect existing code
bug
an unexpected problem or unintended b...
df-col 👜
documentation
feature
a feature request or enhancement
good first issue ❤️
good issue for first-time contributors
group 👪

Ask before starting the work

Makes sure that:

- you are not doing the same work as someone else
- the work is wanted

Artwork by @allison_horst

Github Projects

Text Processing Recipes Closed

Updated on Dec 18, 2018

currently in development with the `textrecipes` package.

- seriousness: medium
- complexity: low

Add steps for stop words, stemming, tf-idf, n-grams and so on.


ME! ME! ME!

text recipe steps #192

Open **topepo** opened this issue on Sep 5, 2018 · 14 comments

topepo commented on Sep 5, 2018 Member

For this project, the idea is to have steps that can be used to process text data (contained in a new package). I've made placeholders in that project for some obvious processing candidates.

@EmilHvitfeldt has volunteered to get started. Perhaps **@juliasilge**, **@skeydan**, and others might have some suggestions and opinions. I'd be happy to include tensorflow methods for text processing too.

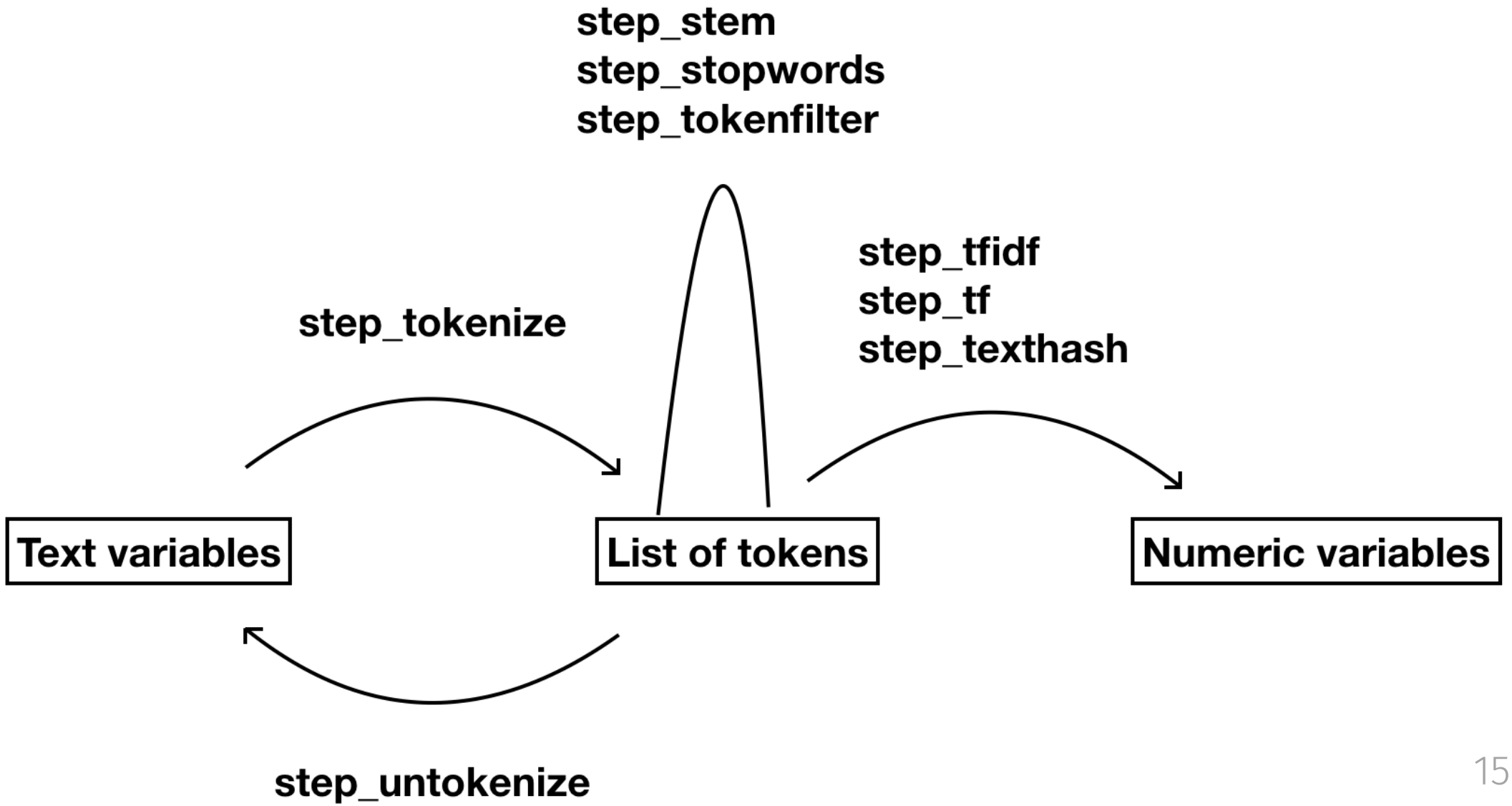
I can add anyone interested in helping to the project page. We can use this issue to discuss ideas and kick around implementation questions.

Challenges

- Inheret messy structure of text
- steps don't have specified order

Design choice

Flexibily > speed


step_stem
step_stopwords
step_tokenfilter
step_tokenize
step_tfidf
step_tf
step_texthash
Text variables
List of tokens
Numeric variables
step_untokenize

```r
step_stem <-
  function(recipe,
           ...,
           role = NA,
           trained = FALSE,
           columns = NULL,
           options = list(),
           custom_stemmer = NULL,
           skip = FALSE,
           id = rand_id("stem")
  ) {
    add_step(
      recipe,
      step_stem_new(
        terms = ellipse_check(...),
        role = role,
        trained = trained,
        options = options,
        custom_stemmer = custom_stemmer,
        columns = columns,
        skip = skip,
        id = id
      )
    )
  }

step_stem_new <-
  function(terms, role, trained, columns, options, custom_stemmer, skip, id) {
    step(
      subclass = "stem",
      terms = terms,
      role = role,
      trained = trained,
      columns = columns,
      options = options,
      custom_stemmer = custom_stemmer,
      skip = skip,
      id = id
    )
  }

prep.step_stem <- function(x, training, info = NULL, ...) {
  col_names <- terms_select(x$terms, info = info)

  check_list(training[, col_names])

  step_stem_new(
    terms = x$terms,
    role = x$role,
    trained = TRUE,
    columns = col_names,
    options = x$options,
    custom_stemmer = x$custom_stemmer,
    skip = x$skip,
    id = x$id
  )
}
```

```r
bake.step_stem <- function(object, new_data, ...) {
  col_names <- object$columns

  stem_fun <- object$custom_stemmer %||%
    SnowballC::wordStem

  for (i in seq_along(col_names)) {
    stemmed_text <- map(new_data[, col_names[i], drop = TRUE],
                        stem_fun)

    new_data[, col_names[i]] <- tibble(stemmed_text)
  }
  new_data <- factor_to_text(new_data, col_names)
  as_tibble(new_data)
}

print.step_stem <-
  function(x, width = max(20, options()$width - 30), ...) {
    cat("Stemming for ", sep = "")
    printer(x$columns, x$terms, x$trained, width = width)
    invisible(x)
}

tidy.step_stem <- function(x, ...) {
  if (is_trained(x)) {
    res <- tibble(terms = x$terms,
                  is_custom_stemmer = is.null(x$custom_stemmer))
  } else {
    term_names <- sel2char(x$terms)
    res <- tibble(terms = term_names,
                  value = na_chr)
  }
  res$id <- x$id
  res
}
```

```r
step_stem <-
  function(recipe,
           ...,
           role = NA,
           trained = FALSE,
           columns = NULL,
           options = list(),
           custom_stemmer = NULL,
           skip = FALSE,
           id = rand_id("stem")
  ) {
    add_step(
      recipe,
      step_stem_new(
        terms = ellipse_check(...),
        role = role,
        trained = trained,
        options = options,
        custom_stemmer = custom_stemmer,
        columns = columns,
        skip = skip,
        id = id
      )
    )
  }

step_stem_new <-
  function(terms, role, trained, columns, options, custom_stemmer, skip, id) {
    step(
      subclass = "stem",
      terms = terms,
      role = role,
      trained = trained,
      columns = columns,
      options = options,
      custom_stemmer = custom_stemmer,
      skip = skip,
      id = id
    )
  }

prep.step_stem <- function(x, training, info = NULL, ...) {
  col_names <- terms_select(x$terms, info = info)

  check_list(training[, col_names])

  step_stem_new(
    terms = x$terms,
    role = x$role,
    trained = TRUE,
    columns = col_names,
    options = x$options,
    custom_stemmer = x$custom_stemmer,
    skip = x$skip,
    id = x$id
  )
}
```

```r
bake.step_stem <- function(object, new_data, ...) {
  col_names <- object$columns

  stem_fun <- object$custom_stemmer %||%
    SnowballC::wordStem

  for (i in seq_along(col_names)) {
    stemmed_text <- map(new_data[, col_names[i], drop = TRUE],
                        stem_fun)

    new_data[, col_names[i]] <- tibble(stemmed_text)
  }
  new_data <- factor_to_text(new_data, col_names)
  as_tibble(new_data)
}

print.step_stem <-
  function(x, width = max(20, options()$width - 30), ...) {
    cat("Stemming for ", sep = "")
    printer(x$columns, x$terms, x$trained, width = width)
    invisible(x)
}

tidy.step_stem <- function(x, ...) {
  if (is_trained(x)) {
    res <- tibble(terms = x$terms,
                  is_custom_stemmer = is.null(x$custom_stemmer))
  } else {
    term_names <- sel2char(x$terms)
    res <- tibble(terms = term_names,
                  value = na_chr)
  }
  res$id <- x$id
  res
}
```

The bake step

```
bake.step_stem <- function(object, new_data, ...) {
  col_names <- object$columns

  stem_fun <- object$custom_stemmer %||%
    SnowballC::wordStem

  for (i in seq_along(col_names)) {
    stemmed_text <- map(new_data[, col_names[i], drop = TRUE],
                        stem_fun)

    new_data[, col_names[i]] <- tibble(stemmed_text)
  }
  new_data <- factor_to_text(new_data, col_names)
  as_tibble(new_data)
}
```

the traceback

```
> library(textrecipes)
> library(recipes)
> xxx <- recipe(~ essay0, data = okc_text) %>%
+   step_tokenize(essay0) %>%
+   step_stem(essay0) %>%
+   prep() %>%
+   juice()
Error in bake.step_stem(x$steps[[i]], new_data = training) : test
> traceback()
12: stop("test") at stem.R#151
11: bake.step_stem(x$steps[[i]], new_data = training)
10: bake(x$steps[[i]], new_data = training)
9: prep.recipe(.)
8: prep(.)
7: function_list[[i]](value)
6: freduce(value, `_function_list`)
5: `_fseq`(`_lhs`)
4: eval(quote(`_fseq`(`_lhs`)), env, env)
3: eval(quote(`_fseq`(`_lhs`)), env, env)
2: withVisible(eval(quote(`_fseq`(`_lhs`)), env, env))
1: recipe(~essay0, data = okc_text) %>% step_tokenize(essay0) %>%
       step_stem(essay0) %>% prep() %>% juice()
```

- I call `prep()`
- `prep()` calls `prep.recipe()`
- `prep.recipe()` calls `bake()` in a loop
- `bake()` calls `bake.step_stem()`

quite a few levels deep.

Browser to the rescue

Plenty of follow up

Don't reinvent the wheel

textrecipes stands on the shoulders of

- recipe (obviously)
- tokenizers
- SnowballC
- stopwords
- text2vec
- textfeatures

My biggest git mistake

revdep start

master 0.0.1

Browse files

EmilHvitfeldt committed on Nov 25, 2018 1 parent 5c4634d commit 5a37841c70c5484bcc50dbfd492bf9bd828b2857

Showing **13,543 changed files** with **2,620,784 additions** and **0 deletions**.

Unified Split

The diff you're trying to view is too large. We only load the first 3000 changed files.

1 .Rbuildignore

```
@@ -7,3 +7,4 @@
7  7    ^\.Rproj\.user$
8  8    ^docs$
9  9    ^_pkgdown\.yml$
   10 + ^revdep$
```

The reward - joined my first organization

Joined the **tidymodels** organization on Nov 18

tidymodels

The reward - CRAN releases

The reward - Post on tidyverse.org

Thank you!

EmilHvitfeldt

@Emil_Hvitfeldt

emilhvitfeldt

www.hvitfeldt.me

Slides created via the R package xaringan.