Projects for Text Version Comparison and Analytics in R

Contents

1 Overview

Status

R code: 1616 *C++ code:* 112 *test code:* 1010

Version

0.1.12

Description

Provides data structures and methods for manual as wells as automated R based text comparison and text as well as change coding.

Funding

This software was created as part of the "Institutional Design in Western European Democracies" research project, funded by DFG (Deutsche Forschungsgemeinschaft), lead by Ulrich Sieberer and based at University Konstanz.

License

Citation

Meißner P (2016). *diffrprojects: Projects for Text Version Comparison and Analytics in R*. R package version 0.1.12, <URL: https://github.com/petermeissner/diffrprojects>.

Sieberer U, Meißner P, Keh J and Müller W (2016). "Mapping and Explaining Parliamentary Rule Changes in Europe: A Research Program." *Legislative Studies Quarterly*, *41*(1), pp. 61-88. ISSN 1939-9162, doi: 10.1111/lsq.12106 (URL: http://doi.org/10.1111/lsq.12106), <URL: http://dx.doi.org/10.1111/lsq.12106>.

BibTex for citing

```
toBibtex(citation("diffrprojects"))
```

Installation

stable CRAN version

```
install.packages("diffrprojects")
library(rtext)
```

(stable) development version

```
standard_repos <- options("repos")$repos
install.packages(
  "diffrprojects",
  repos = c(standard_repos, "https://petermeissner.github.io/drat/")
)
library(rtext)
```

Contribution

Note, that this package uses a Contributor Code of Conduct. By participating in this project you agree to abide by its terms: http://contributor-covenant.org/version/1/0/0/ (basically this should be a place were people get along with each other respectful and nice because it's simply more fun that way for everybody)

Contributions are very much welcome, e.g. in the form of:

- **typo fixing** (edit file directly on Github)
- **bug reporting** (file an issue - after having searched if the issue came up before - as - if possible - minimal reproducable example)
- **extending help files** (e.g. edit the respective files directly on Github or fork the package and later on make a pull request; note, that the package use roxygen2 for easing documentation)
- **writing example** (e.g. edit the respective files directly on Github or fork the package and later on make a pull request; note, that the package use roxygen2 for easing documentation)
- **vignette writing** (file an issue first so that we can discuss htings than fork the package and later on make a pull request)
- **test writing** (have a look at the test coverage than fork the package and later on make a pull request)
- **feature suggestions** (file an issue describing the idea, why this is important, possible alternative solutions and an example)
- **general discussion** of approach and or implementation (file an issue)
- implementation **improvements** (file an issue naming whats to be improved, why and how)

2 Usage

2.1 Fast Introduction for the Impatient

For those in a hurry here is a very brief

```r
# loading package
library(diffrprojects)


# the first chapter of Robinson Crusoe from three different sources
rcs <- rtext:::testfile(pattern="rc.*ch1.txt", full.names = TRUE)


# creating a new project
dp <- diffrproject$new()


# setting options
dp$options$verbose <- FALSE


# adding texts to the corpus
dp$text_add(text_file = rcs)
dp$text_data(1) %>% head(11)
```

```
##     i char         name
## 1   1    T rc_1_ch1.txt
## 2   2    h rc_1_ch1.txt
## 3   3    e rc_1_ch1.txt
## 4   4      rc_1_ch1.txt
## 5   5    P rc_1_ch1.txt
## 6   6    r rc_1_ch1.txt
## 7   7    o rc_1_ch1.txt
## 8   8    j rc_1_ch1.txt
## 9   9    e rc_1_ch1.txt
## 10 10    c rc_1_ch1.txt
## 11 11    t rc_1_ch1.txt
```

```r
# linking the files (which file should be compared to which)
dp$text_link()
dp$link %>% as.data.frame()
```

```
##           from           to                       link
## 1 rc_1_ch1.txt rc_2_ch1.txt rc_1_ch1.txt~rc_2_ch1.txt
## 2 rc_2_ch1.txt rc_3_ch1.txt rc_2_ch1.txt~rc_3_ch1.txt
```

```r
# calculating text alignments
dp$text_align(tokenizer=text_tokenize_words)
dp$alignment[[1]] %>% head(30)
```

```
##    alignment_i token_i_1 token_i_2 distance      type from_1 to_1 from_2  to_2
## 1            1         1      1932        0 no-change      1    3  10326 10328
## 2            2         2        NA        7  deletion      5   11     NA    NA
## 3            3         3        NA        9  deletion     13   21     NA    NA
## 4            4         4        NA        5  deletion     23   27     NA    NA
## 5            5         5      1932        0 no-change     30   32  10326 10328
## 6            6         6        NA        4  deletion     34   37     NA    NA
## 7            7         7        87        0 no-change     39   41    513   515
## 8            8         8        NA       10  deletion     43   52     NA    NA
## 9            9         9        57        0 no-change     54   55    355   356
```

```
## 10          10         10          3         0 no-change     57   64     15     22
## 11          11         11          4         0 no-change     66   71     24     29
## 12          12         12         85         0 no-change     74   75    497    498
## 13          13         13          1         0 no-change     77   82      1      6
## 14          14         14          2         0 no-change     84   88      8     12
## 15          15         15       1520         0 no-change     92   95   8187   8190
## 16          16         16         NA         5  deletion     97  101     NA     NA
## 17          17         17       3667         0 no-change    103  104  19215  19216
## 18          18         18        263         0 no-change    106  108   1495   1497
## 19          19         19         51         0 no-change    110  112    328    330
## 20          20         20         NA         3  deletion    114  116     NA     NA
## 21          21         21         57         0 no-change    118  119    355    356
## 22          22         22         NA         6  deletion    121  126     NA     NA
## 23          23         23         NA         8  deletion    128  135     NA     NA
## 24          24         24         50         0 no-change    137  138    325    326
## 25          25         25         51         0 no-change    140  142    328    330
## 26          26         26         NA         6  deletion    144  149     NA     NA
## 27          27         27         NA         6  deletion    151  156     NA     NA
## 28          28         28         87         0 no-change    158  160    513    515
## 29          29         29        513         0 no-change    162  165   2853   2856
## 30          30         30        306         0 no-change    167  171   1724   1728
```

2.2 Creating a Diffrprojects Instance

To create a diffrproject we use the diffrproject creator object - its simply an object with an function that knows how to create a project.

Creating a project looks like this:

```
library(diffrprojects)
dp <- diffrproject$new()
```

Et violà - we created a first, for now empty, project that we will use throughout the tutorial.

2.3 Some Help Please

To get a better idea about what this thing called *diffrproject* really is you can consult its help page which gives a broad overview over its capabilities:

```
?diffrproject
```

Another way is to call the ls() method. This will present us with a data frame listing all fields where data is stored and all the methods (aka object specific functions) of our diffrprojects instance. Those methods and fields located in *private* are not for the user to mess around with while non-private (*self* aka public) data fields can be read by the user and public methods can be triggered by the user to manipulate the data or retrieve data in a specific format.

```
dp$ls()
```

```
##                            name   where                    class
## 1                  execute_load private                 function
## 2                          hash private                 function
## 3                       hashed private                 function
## 5                  prepare_save private                 function
## 4                        hashes private                     list
```

```
## 9                   alignment_data    self alignment_data_list, list
## 6                        alignment    self      alignment_list, list
## 21                            link    self      alignment_list, list
## 7                    alignment_add    self                  function
## 8                   alignment_code    self                  function
## 10             alignment_data_full    self                  function
## 11              alignment_data_set    self                  function
## 12                alignment_delete    self                  function
## 13                           clone    self                  function
## 14                           debug    self                  function
## 15                      export_csv    self                  function
## 16                   export_sqlite    self                  function
## 17                             get    self                  function
## 18                      import_csv    self                  function
## 19                   import_sqlite    self                  function
## 20                      initialize    self                  function
## 22                            load    self                  function
## 23                              ls    self                  function
## 24                         message    self                  function
## 27                            save    self                  function
## 29                        text_add    self                  function
## 30                       text_align    self                  function
## 31                        text_code    self                  function
## 32       text_code_alignment_token    self                  function
## 33 text_code_alignment_token_regex    self                  function
## 34                  text_code_regex    self                  function
## 35                        text_data    self                  function
## 36                text_data_inherit    self                  function
## 37                      text_delete    self                  function
## 38                        text_link    self                  function
## 39                   text_meta_data    self                  function
## 40         tokenize_text_data_lines    self                  function
## 41         tokenize_text_data_regex    self                  function
## 42         tokenize_text_data_words    self                  function
## 43                           warning    self                  function
## 25                              meta    self                      list
## 26                           options    self                      list
## 28                              text    self                      list
```

The base R class() function furthermore reveals from which classes the diffrproject class inherits:

```
class(dp)
```

```
## [1] "diffrproject"      "dp_inherit"        "dp_align"          "dp_export"
## [5] "rtext_loadsave"    "dp_base"           "R6_rtext_extended" "R6"
```

2.4 Adding Texts to Projects

Our diffrproject (`dp`) has one method called `text_add()` that allows to add texts to the project. Basically the method can be used in three different flavors: adding character vectors, adding texts stored on disk, or by adding rtext objects (see rtext package: https://CRAN.R-project.org/package=rtext; rtext objects are the way individual texts are represented within diffrprojects). For each of these use cases there is one option: `text`, `text_file`, `rtext`; respectively.

Below are shown examples using each of these methods:

adding text files

```
test_file1 <- stringb:::test_file("rc_1_ch1.txt")
test_file2 <- stringb:::test_file("rc_2_ch1.txt")
dp$text_add(text_file = c(test_file1, test_file2) )
```

adding rtext objects

```
test_file <- stringb:::test_file("rc_1_ch1.txt")
rt <- rtext$new( text_file = test_file)
dp$text_add(rtext = rt)
```

adding character vectors

```
test_file1 <- stringb:::test_file("rc_1_ch1.txt")
test_file2 <- stringb:::test_file("rc_2_ch1.txt")
cv <- ""
cv[1] <- text_read(test_file1, NULL)
cv[2] <- text_read(test_file2, NULL)
dp$text_add(text = cv)
```

In the last case make sure to put each text in one separate line. Functions like readLines() or text_read() read in texts such that each line corresponds to one element in a character vector. With e.g. text_read()'s tokenize parameter to NULL the text will be read in as one long string.

2.5 Piping methods

Now is a good time to mention a feature of diffrprojects that comes in handy: All functions that do not explicitly extract data (those usually have some 'get' as part of their name) do return return the object itself so that one can pipe together a series of method calls.

Consider the following example where we initiate a new diffrprojects instance and add two texts in just one pipe:

```
dp <-
  diffrproject$
  new()$
  text_add(text_version_1, name = "version1")$
  text_add(text_version_2, name = "version2")

length(dp$text)
```

```
## [1] 2
```

2.6 Getting Infos About Texts

If we want to get some general overview about the texts gathered in our project we can use the text_meta_data() method to do so. The method has no parameters and return a data.frame with several variables informing us about its source, length, encoding used for storage, and its name.

```
dp$text_meta_data()
```

```
##   text_file character encoding sourcetype     name
## 1      <NA>       479    UTF-8       text version1
## 2      <NA>       539    UTF-8       text version2
```

2.7 Getting And Setting Infos About the Project

Similar to the text_meta_data() method we can access the projects meta data via data fields meta and options. But contrary to the text_meta_data() method that gathers data from all the texts within the project and does not allow for manipulation of the data, the data fields allow reading and writing.

First let us have a look and thereafter turn of the message notification service:

getting data fields

```
dp$options
```

```
## $verbose
## [1] TRUE
##
## $warning
## [1] TRUE
##
## $ask
## [1] TRUE
```

setting data fields

```
dp$options$verbose <- FALSE
```

(note, ask is deprecated and only remains for compatibility reasons but has no function anymore)

Now its time to have a look at the projects meta data. It tells us when the project was created, which path to use for SQLite exports, which path to use for saving data as in RData format and what is the projects id. The id is a hash of a time stamp as well as session information which should ensure uniqueness across space and time.

All these values can manipulated by the user to her liking.

```
dp$meta
```

```
## $ts_created
## [1] "2016-11-05 20:23:02 UTC"
##
## $db_path
## [1] "./diffrproject.db"
##
## $file_path
## [1] ""
##
## $project_id
## [1] "c0ead7afcda40f05bd4b7eb5fde43c19"
```

```
dp$meta$file_path = "./diffrproject.RData"
```

2.8 Deleting Texts

Of cause we can not only add texts but delete them from the project as well. For this purpose there is the text_delete() method.

Let's just add two texts and delete one by providing its index number and the second by providing its name to the text_delete() method.

```
dp$text_add(text = "nonesense", "n1")
dp$text_add(text = "nonesense", "n2")

dp$text_delete(3)
dp$text_delete("n2")

length(dp$text)
```

```
## [1] 2
```

```
names(dp$text)
```

```
## [1] "version1" "version2"
```

2.9 Defining Relationships Between Texts: Linking

The purpose of diffrprojects is to enable data collection on the difference of texts. Having filled a project with various texts there are endless possibilities to form pairs of text for comparison and change measurement - where endless actually is equal to: $n^2 - n$.

Linking can be done via the text_link method which accepts either index numbers or text names for its from and to arguments (a third argument delete will delete a specified link if set to TRUE).

```
dp$text_link(from = 1, to = 2)
dp$text_link(from = 1, to = 2, delete = TRUE)
```

If no arguments are specified text_link will linkthe first text to the second, the third to th fourth, the fourth to the fifths and so on.

```
dp$text_link()
```

To get an idea of what links are currently specified we can directly access the link data field or/and ask R to transforme the list found there into a data.frame.

```
dp$link
```

```
## $`version1~version2`
## $`version1~version2`$from
## [1] "version1"
##
## $`version1~version2`$to
## [1] "version2"
##
##
## attr(,"class")
## [1] "alignment_list" "list"
```

```
dp$link %>% as.data.frame()
```

```
##       from       to              link
## 1 version1 version2 version1~version2
```

2.10 Aligning Texts and Measuring Change

At the heart of each diffrproject lies the text_align method. This method compares two texts and tries to align parts of one text with parts of the other text. The first two arguments (`t1` and `t2`) are for specifying

which pair of texts to compare - if left as-is, all text pairs that are specified within the link data field will be aligned.

Text parts are arbitrary character spans defined by the `tokenizer` argument. This argument expects a function splitting text into a token data.frame. If the tokenizer argument is left as-is it will default to text_tokenize_lines function from the stringb package.

Text tokens can be pre-processed before alignment. The `clean` argument allows to hand over a function tranforming a charactr vector of text tokens into their clean counterparts.

The `ignore` arguments expects a function that is able to transform a character vector of tokens into a logical vector of same length indicating which tokens to ignore throughout the alignment process and which to consider.

The next argument - `distance` - specifies which distance metrics to use to calculate distances between strings.

Since the text_align method basically is a wrapper around diff_align you can get more information via `?diff_align` and since again diff_align is a wrapper around stringdist from the stringdist package `?stringdist::stringdist` and also `?stringdist::‘stringdist-metrics‘` will provide further insights about possible metrics and how to use the rest of the arguments to text_align (these are passed through to stringdist).

Let's have an example:

```
dp$text_align(distance = "lv", maxDist = 1)

dp$alignment
```

```
## $`version1~version2`
##     alignment_i token_i_1 token_i_2 distance      type from_1 to_1 from_2 to_2
## 1             1         1         6        0 no-change      1   16     97  112
## 2             2         2         7        0 no-change     18   40    114  136
## 3             3         3         8        0 no-change     42   61    138  157
## 4             4         4         9        0 no-change     63   84    159  180
## 5             5         5        10        0 no-change     86  107    182  203
## 6             6         6        11        0 no-change    109  132    205  228
## 7             7         7        12        0 no-change    134  156    230  252
## 8             8         8        NA       24  deletion    158  181     NA   NA
## 9             9         9        NA        8  deletion    183  190     NA   NA
## 11           10        10        23        0 no-change    193  215    475  497
## 12           11        11        NA       22  deletion    217  238     NA   NA
## 13           12        12        NA       25  deletion    240  264     NA   NA
## 14           13        13        NA       12  deletion    266  277     NA   NA
## 16           14        14        14        0 no-change    280  303    274  297
## 17           15        15        15        1    change    305  327    299  321
## 18           16        16        16        0 no-change    329  345    323  339
## 19           17        17        17        0 no-change    347  367    341  361
## 20           18        18        18        0 no-change    369  389    363  383
## 21           19        19        19        0 no-change    391  412    385  406
## 22           20        20        20        0 no-change    414  436    408  430
## 23           21        21        21        0 no-change    438  459    432  453
## 24           22        22        22        0 no-change    461  478    455  472
## 15           23        NA         1       20 insertion     NA   NA      1   20
## 25           24        NA         2       17 insertion     NA   NA     22   38
## 31           25        NA         3       23 insertion     NA   NA     40   62
## 41           26        NA         4       21 insertion     NA   NA     64   84
## 51           27        NA         5        9 insertion     NA   NA     86   94
## 141          28        NA        13       18 insertion     NA   NA    254  271
```

```
## 27           29        NA        24       23 insertion     NA   NA    499  521
## 28           30        NA        25       17 insertion     NA   NA    523  539
##
## attr(,"class")
## [1] "alignment_list" "list"
```

2.11 Coding Texts

Now let us put some data into our diffrproject.

The most basic method to do so is simply called text_code. Text_code takes up to five arguments (the first three are mandatory) where one specifies the text to be coded (`text`, either by index number or by name), how the variable to store the information is called (`x`), and the index number or a vector of those indicating which characters of the text should be coded. The last two parameters are optional and specify which value the variable should hold (`val`) and at which hierarchy level the coding is placed (`hl`, higher or equal hierarchy levels will overwrite existing codings of lower hierarchy level for the same text, character span, and variable).

```
dp$text_code(text = 1, x = "start", i=1:10, val = TRUE, hl = 0)
dp$text_code(text = "version2", x = "start", i=1:10, val = TRUE, hl = 0)
```

The text_code method is quite verbose and in most cases more suited to be accessed by a machine or algorithm than by a human. Therefore, there are three other methods to code text: text_code_regexp, text_code_alignment_token, text_code_alignment_token_regexp.

2.12 Getting Text Codings

2.13 Aggregating Text Codings

2.14 Text Coding Inheritence

3 Technicalities

3.1 Naming Conventions and General Structure of Methods and Data

The methods and data fields of diffrprojects can be categorized into five realms - *cursive*: methods; (parentheses): private; rest: data:

- **text**: everything related to individual texts starts with text
  - text, *text_add*, *text_delete*, *text_align*, *text_code*, *text_code_alignment_token*, *text_code_alignment_token_regex*, *text_code_regex*,
  - *text_data*, *text_data_inherit*, *tokenize_text_data_lines*, *tokenize_text_data_regex*, *tokenize_text_data_words*
  - text_meta_data
- **alignment**: everything that concerns the relation between two texts
  - alignment, *alignment_add*, *alignment_code*, *alignment_delete*, *alignemtn_data_full*, *alignment_data_set*
  - text_link, *link*
- **misc**:
  - meta, options, *load*, *save*, *export_sqlite*, *import_sqlite*, (*execute_load*), (*prepare_save*)
- **inherited from R6_rtext_extended**:
  - options, *message*, *warning*, (*hash*), (hashed), (hashes)
- **inherited from R6**:

– *clone*, *initialize*

3.2 Data formats

3.2.1 meta

Meta is a list with only a few items providing/storing general information for the whole project - i.e. time stamp the project was created, path to store data, path to export data, an project id.

3.2.2 text

Text is a list of rtext instances. Each rtext instances stores text's actual text as data gathered on the text.

The text_data method will return a data.frame containing all text data, while tokenize_text_data_xxx methods will aggregate text data to specific token levels: words, lines or user defined patterns.

3.2.3 link

Link is a list of links between texts. Link defines for which text combination alignments should be calculated. Each list item hold a from and to field which stores the names of texts to be aligned. The method to create links is text_link, it also allows to delete specific links.

Link data can be transformed to one big data.frame via: as.data.frame function.

3.2.4 alignment

Alignment is a list of data.frames. Each alignment list item stores which part of one text (character span) is connected to which part of another text (character span).

The list of alignments can be transformed to one big data.frame via: as.data.frame function.

3.2.5 alignment_data

The list of alignment_data can be transformed to one big data.frame via: as.data.frame function.

[[[???!!!]]]

3.3 The Diffrprojects Universe

Diffrprojects has two other packages it relies heavily on and one package that add further features.

3.3.1 Rtext

Rtext is a package providing a data structure and accompanying methods to handle texts / strings / characters as well as data bound to these texts / strings / characters. All string manipulations are based upon the stringb package. All diffrproject texts are actually rtext instances. Unfortunate you cannot yet manipulate rtext objects once they are part of a diffrproject and expect that data on the relation between texts (i.e. alignment and alignment_data) gets updated as well - hence manipulating texts might lead to inconsistencies in alignments and alignment_data.

A strategy to implement such a feature would be to extend rtext in such a way that text manipulation methods would pass change information to a list of call back functions. Furthermore, diffrprojects need to methods that allow for handling shifts in the character sequences of texts. Those update methods can than be passed to rtext instances once they become part of a diffrproject. Then whenever e.g. some characters are deleted alignments as well as alignment data touching these character spans get deleted as well and character span information for all other alignments get shifted by the appropriate amount.

For those preferring a version using stringi/stringr - go ahead - since rtext and diffrprojects provide tests for all respectively for all vital parts and stringb copied the function naming scheme from stringr anyways this should be a small matter.

3.3.2 Stringb

Stringb is a package providing convenience functions for string handling and manipulation using R's own regular expression engine. All string manipulations are based upon the stringb package.

In addition stringb provides very flexible text tokenization functions that are very much in line with the needs of diffrprojects.

3.3.3 Diffrprojectswidget

This package enhances diffrprojects by providing HTMLwidgets for visualizing diffrproject data: as interactive table or as interactive graph.

HTMLwidgets (see: http://www.htmlwidgets.org/) are a framework that allows for interactive, web technology based graphics that are furthermore easily integrate able into e.g. R-shiny (http://shiny.rstudio.com/) applications.

3.4 Two words or more about Objects / R6 / Classes / Instances

Diffrprojects is written in object oriented programming style because it seemed adequate to do so. Why? Because in OOP in comparison to functional programming one does more stuff like in-place-modification of data, data and its modifiers (methods) come in one big bundle, its easier to work on the current state of the object / to only allow consistent states of the object. Yeap everything here could have done with FP as well - please go ahead.

The downside of using OOP in R is that what happens becomes much mor intrasparent and harder to reason about - I am sorry for that.

3.4.1 Classes and Instances

Classes are object blueprints - a schema that describes how an object of this class should look like. Classes might be objects too but they are not the objects they describe. To get an object instance of an object - a manifestion of the idea of the object described in the class - one has to explicitly translate execute the instructions led out in the class, e.g. via: `diffrproject$new()`, or `rtext$new()`

3.4.2 R6

R6 is a package that provides a framework that makes it very very easy to build objects in R that are more like things known from traditional all purpose programming languages like say Java or C++.

3.4.3 Methods

Methods are exactly like functions only that they are not floating around loosely in your global environment or elsewhere but are bound to specific instances of an object. So there is not one text_add function that can be used with any diffrproject but there is one specific text_add method for each instance of an diffrproject. This quite strange right? Why the duplication? Well with that you can e.g. pass this method around, hand it over to a function that calls it or put it into another object maybe that than can decide to use it or not. A silly example:

```
dp1 <- diffrproject$new()
add_text_to_dp1 <- dp1$text_add

add_text_to_dp1("ahhh")
add_text_to_dp1("behhh")
add_text_to_dp1("cehhh")

names(dp1$text)
```

```
## [1] "noname_1" "noname_2" "noname_3"
```