Package 'hypoparsr'

September 6, 2016

Version 0.1.0

Title Multi-Hypothesis CSV Parser

Author Till Doehmen [aut, cre], Hannes Muehleisen [ctb]

Maintainer Hannes Muehleisen <hannes@cwi.nl>

Description A Multi-Hypothesis CSV Parser. Stresses your computer not you.

License MPL (== 2.0)

URL https://bitbucket.org/doehmen/open-information-integration-masters-thesis

Imports data.tree, RecordLinkage, readr, tibble

Suggests testthat

Collate misc.R encoding.R dialect.R table_area.R row_function.R
col_function.R data_type.R quality_assessment.R parser_full.R

NeedsCompilation no

Repository CRAN

Date/Publication 2016-09-06 10:04:43

R topics documented:

`parse_file` *Multi-hypothesis parsing of CSV files.*

Description

hypoparsr takes a different approach to CSV parsing by creating different parsing hypotheses for a given file and ranking them based on data quality features. `parse_file` creates and returns the ranked parsing results.

Usage

```
parse_file(file, pruning_level = 0.1, quality_weights =
                 c(warnings = -1, edits = -1, moves = -1, confidence = 1,
                 total_cells = 1, typed_cells = 1, empty_header = -1,
                 empty_cells = -1, non_latin_chars = -1, row_col_ratio =
                 1))
```

Arguments

`file`	Path to a CSV file.
`pruning_level`	Numeric value between 0-1 which defined the lower threshold for confidence values of parsing hypotheses. The higher the value, the less hypotheses are created and the correct hypothesis might be omitted.
`quality_weights`	A named list of numeric quality feature weights which influence the hypothesis ranking. Positive weights improve the ranking of results with the respective characteristic and negative weights penalize the same.

Value

A `hypoparsr_result`, which contains all created hypotheses and their ranking. Call `as.data.frame()` on this object to retrieve the highest ranked parsing result.

Examples

```
# generate a CSV
csv <- tempfile()
write.csv(iris, csv, row.names=FALSE)

# call hypoparsr
res <- hypoparsr::parse_file(csv)

# get result data frames
best_guess <- as.data.frame(res)
second_best_guess <- as.data.frame(res, rank=2)
```

Index