regular-expression

Max Lang

10/27/2022

‘find_url“

This function will find URLs in long text and splits them into the respective components such as protocol, domainname etc.

#ä Input * `input`: A `character(1)` containing the input string.

Output:

- A data.frame with columns protocol, domainname and path.

Code

```
find_url <- function(input) {
  assertString(input)
  rx <- "(?<![[:alpha:]])(https?)://([[:alnum:]][-[:alnum:]]*(?:\\.[[:alnum:]][-[:alnum:]]*)+)/([-a-z[:

  matches <- gregexpr(rx, input, perl = TRUE, ignore.case = TRUE)
  urls <- regmatches(input, matches)[[1]]

  m <- regexec(rx, urls, perl = TRUE, ignore.case = TRUE)
  urlparts <- regmatches(urls, m)
  res <- t(vapply(urlparts, identity, character(4), USE.NAMES = FALSE))
  res <- res[, -1]
  res <- as.data.frame(res, stringsAsFactors = FALSE)
  colnames(res) <- c("protocol", "domainname", "path")
  res
}
```

Worked example

```
string <- "https://www.google.com/ is probably the most popular url, sometimes used ashttp://www.google
find_url(string)
```

```
##   protocol     domainname                           path
## 1    https www.google.com
## 2    HTTPS     GITHUB.COM MAXMLANG/STRUCTURED-PROGRAMMING
```