Shuffled Text Generator

Chris Bentz

4/2/2020

Load libraries

If the libraries are not installed yet, you need to install them using, for example, the command: install.packages("ggplot2").

```
library(stringi)
library(gsubfn)
```

```
## Loading required package: proto
```

List files

Create list with all the files in the directory "writing".

```
file.list <- list.files(path = "/home/chris/Github/StringBase/corpus/generated/test/original",
                        recursive = T, full.names = T)
#print(file.list)
length(file.list)
```

```
## [1] 8
```

Shuffled text generator

Create shuffled texts.

```
# set the maximal number of units (n) to be used for analysis
n = 10000
for (file in file.list)
  {
  # loading textfile
  textfile <- scan(file, what = "char", quote = "", comment.char = "",
                   encoding = "UTF-8", sep = "\n" , skip = 7)
  # remove tabs and parentheses
  textfile <- gsubfn(".", list("\t" = "", "(" = "", ")" = "", "]" = "",
                               "[" = "", "}" = "",  "{" = ""), textfile)
  # remove annotations marked by '<>'
  textfile <- gsub("<.*>","",textfile)
  # split the textfile into individual utf-8 characters. The output of strsplit()
  # is a list, so it needs to be "unlisted"" to get a vector. Note that white spaces
  # are counted as utf-8 characters here.
  chars <- unlist(strsplit(textfile, ""))
```

```
  chars <- chars[1:n] # use only maximally n units
  chars <- chars[!is.na(chars)] # remove NAs for vectors which are
  # already shorter than n
  chars <- chars[chars != " "] # remove white spaces from character vector
  # collapse the character vector into a single string
  chars.collapsed <- paste(chars, collapse = "")
  chars.shuffled <- stri_rand_shuffle(chars.collapsed)
  # create meta information for file header
  meta.info <- c("#type:\tshuffled",
                 paste("#specification:\tshuffled text generated from text", basename(file)),
                "#scriptcode:\tLatn", "#source:\tRcode in shuffledTextGenerator.Rmd",
                "#encoding:\tutf-8", "#copyright:\tNA",
                "#comments:\t", "")
  # add meta information
  full.vec <- append(meta.info, chars.shuffled)
  # get filename
  filename <- basename(file)
  # get the three letter identification code + the running number of the original
  code <- substring(substring(filename, regexpr("_", filename) + 1), 1, 8)
  # write to file
  new.filename <- paste("shuffled_", code, sep = "")
  write(full.vec, file = paste("~/Github/StringBase/corpus/generated/test/shuffled/", new.filename,
                               sep = ""), append = FALSE, sep = "\n")
}
```