HowTo - Text manipulation

A. Blejec

July 30, 2010

Contents

1 Reading the text file

```
> text <- readLines("../dat/text.txt")
```

Parse text into words

```
> x <- unlist(strsplit(text, " "))
> x <- unlist(strsplit(x, "\\("))
> x <- unlist(strsplit(x, "\\)"))
> x <- unlist(strsplit(x, "\\t"))
> x <- unlist(strsplit(x, "\\\\n"))
> x <- unlist(strsplit(x, ","))
> x <- unlist(strsplit(x, "\\."))
> x <- unlist(strsplit(x, "\\\\"))
> x <- unlist(strsplit(x, "="))
> words <- unlist(x)
> str(words)
 chr [1:501] "Read" "Data" "Values" "Description" ...
```

Calculate word lengths

```
> nCharacters <- sapply(words, nchar)
> head(nCharacters)
       Read        Data      Values Description        Read
          4           4           6          11           4
       data
          4


> minChar <- 4
```

Delete the words with less than 4 characters

```
> select <- which(nCharacters >= minChar)
> head(select, 20)
```

```
       Read        Data      Values Description        Read
          1           2           3           4           5
       data        into      vector        list        from
          6           7           9          11          12
    console        file       Usage        scan        file
         14          16          17          18          19
       what      double        nmax       quote   identical
         22          24          26          35          38
```

```
> words <- words[select]
```

Count the words

```
> tbl <- rev(sort(table(words)))
> head(tbl, 20)
```

```
words
      file       read       data       what       will       list
        17          8          7          6          5          5
      from      FALSE  character     values      input      which
         5          5          5          4          4          3
      this       scan     number    maximum       line   encoding
         3          3          3          3          3          3
connection      value
         3          2
```

Which word is most common?

```
> tbl[1]
```

```
file
  17
```

```
> topWord <- names(tbl[1])
> topWord <- "file"
```

Show neighborhood of the topmost word: 'file'

```
> select <- which(words == topWord)
> select
 [1]  12  15  49  51  57  81 102 108 113 114 116 136 141 146 150 176
[17] 221
> n <- 1
> nbgh <- lapply(select, FUN = function(x, n) seq(max(x -
+     n, 1), x + n), n = n)
> lapply(nbgh, FUN = function(x, words) words[x], words = words)
[[1]]
[1] "console" "file"    "Usage"

[[2]]
[1] "scan" "file" "what"

[[3]]
[1] "Arguments" "file"      "name"

[[4]]
[1] "name" "file" "read"

[[5]]
[1] "specified" "file"      "then"

[[6]]
[1] "Otherwise" "file"      "name"

[[7]]
[1] "script"    "file"      "\"stdin\""

[[8]]
[1] "stdin"  "file"   "stream"

[[9]]
[1] "compressed" "file"       "file"

[[10]]
[1] "file"          "file"          "Alternatively"

[[11]]
[1] "Alternatively" "file"          "connection"

[[12]]
[1] "match" "file"  "also"

[[13]]
[1] "data"    "file"    "current"

[[14]]
[1] "Latin-1" "file"    "UTF-8"

[[15]]
[1] "conversely" "file"       "connection"

[[16]]
[1] "data"    "file"    "records"

[[17]]
[1] "read"     "file"     "integer:"
```

SessionInfo

Windows XP (build 2600) Service Pack 3

- R version 2.10.0 (2009-10-26), `i386-pc-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250`, `LC_CTYPE=Slovenian_Slovenia.1250`, `LC_MONETARY=Slovenian_Slovenia.1250`, `LC_NUMERIC=C`, `LC_TIME=Slovenian_Slovenia.1250`
- Base packages: base, datasets, graphics, grDevices, stats, utils
- Other packages: patchDVI 1.5

Project path: C:/_Y/R/I2R

View as vignette

Project files can be viewed by pasting this code to R console:

```
> projectName <-"I2R";  mainFile <-"HowTo-text"

> commandArgs()
> library(tkWidgets)
> openPDF(file.path(dirname(getwd()), "doc", paste(mainFile,
+     "PDF", sep = ".")))
> viewVignette("viewVignette", projectName, file.path("../doc",
+     paste(mainFile, "RNW", sep = ".")))
```