Reading in and cleaning the data

Solutions

Tom Paskhalis

29 November, 2018

```
# Load libraries
library("pdftools")
library("stringr")
```

```
# Read-in the file
draft <- pdftools::pdf_text("../data/draft_withdrawal_agreement_0.pdf")
```

Challenge 1

Easy mode Now extract all the Directives mentioned in the Withdrawal agreement. You can either use `str_extract_all` function from `stringr` package. Otherwise, use `gregexpr` function from base R and pass its output to `regmatches`. The solution should be able to detect such directives as **Directive 92/84/EEC**, **Directive 2011/64/EU** and **Directive 2008/118/EC**.

Medium: Compute a frequency table of directives mentioned and return the first 10 and their corresponding frequencies.

Advanced: Produce a bar chart showing the frequencies of the top 10 directives.

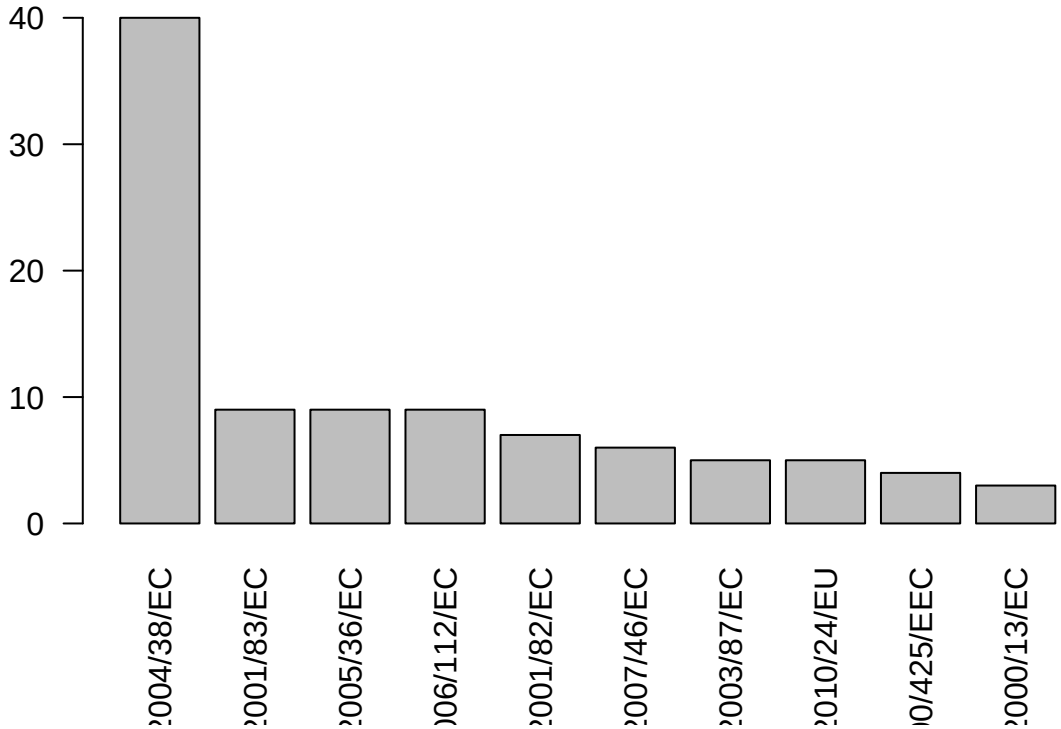
Subject expert: Explore the topics to which the most mentioned directives refer.

```
pattern <- "Directive [0-9]{2,4}/[0-9]{2,3}/[A-Z]{2,3}"

# `base` R solution
directives1 <- regmatches(draft, gregexpr(pattern, draft))

# `stringr` solution
directives2 <- stringr::str_extract_all(draft, pattern)

directives <- unlist(directives1)
directives <- sort(table(directives), decreasing = TRUE)
sum(directives)
```

```
## [1] 328
```

```
directives10 <- directives[1:10]
```

```
barplot(directives10, las = 2)
```



The most mentioned directive, *2004/38/EC* is also called *Citizens' Rights Directive* or *Free Movement Directive*. As its name suggests, it covers the rights of the EU citizens with respect to the freedom of movement. Interestingly, the second most cited directive is *2001/83/EC* is related to medicinal products for human use. While the third, *Directive 2005/36/EC* is about the recognition of professional qualifications.