Alex Farach

Let’s start at the beginning - bits to character encoding in R

2022-06-28

Hello! My name is Alex Farach and I’m a data scientist and analytics manager at Accenture Federal Services.

I think a lot about:

- R & RStudio
- Natural language processing (NLP)
- Data visualization
- Statistical learning

Currently working on:

- github/farach/huggingfaceR

What is bits to character encoding?

Computer → Human

Computer ← Human

← *Protocol* →

How would you describe the letter “A” to a computer?

A	1	0
B	2	1
C	3	00
D	4	01
E	5	10
F	6	11
…	…	…

How many bits are needed to represent 256 unique values?

2 values	1 bit
4 values	2 bits
8 …	3 …
16 …	4 …
32 …	5 …
64 …	6 …
128 …	7 …
256 values	***8 bits = 1 byte***

ASCII, Latin1, and Unicode

ASCII (American Standard Code for Information Interchange):

7 bits = 128 values

ASCII, Latin1, and Unicode

ASCII (American Standard Code for Information Interchange):
7 bits = 128 values

Latin-1 (ISO-8859-1):
8 bits = 256 values

ASCII, Latin1, and UTF-8

ASCII (American Standard Code for Information Interchange):
7 bits = 128 values

Latin-1 (ISO-8859-1):
8 bits = 256 values

UTF-8 (Unicode Transformation 8-bit):
1:4 bytes = 1,112,064 values (or code points)!

Character String Encoding in R

R < 4.2.0

```
print(c("coffee", "café", "caf\u00E9", "caf\xe9"))

Encoding(c("coffee", "café", "caf\u00E9", "caf\xe9"))
## [1] "coffee" "café"  "caf<e9>"   "caf<e9>"
## [1] "unknown" "latin1"  "UTF-8"   "latin1"
```

R >= 4.2.0

```
print(c("coffee", "café", "caf\u00E9", "caf\xe9"))

Encoding(c("coffee", "café", "caf\u00E9", "caf\xe9"))
## [1] "coffee"  "café"    "café"    "caf\xe9"
## [1] "unknown" "UTF-8"   "UTF-8"   "unknown"
```

R < 4.2.0

```
Sys.getlocale()
## [1] "LC_COLLATE=English_United
States.1252;LC_CTYPE=English_United
States.1252;LC_MONETARY=English_United
States.1252;LC_NUMERIC=C;LC_TIME=English_United States.1252"
```

R >= 4.2.0

```
Sys.getlocale()
## [1] "LC_COLLATE=English_United
States.utf8;LC_CTYPE=English_United
States.utf8;LC_MONETARY=English_United
States.utf8;LC_NUMERIC=C;LC_TIME=English_United States.utf8"
```

R < 4.2.0

```
l10n_info()
## $MBCS
## [1] FALSE
##
## $`UTF-8`
## [1] FALSE
##
## $`Latin-1`
## [1] TRUE
##
## $codepage
## [1] 1252
## $system.codepage
## [1] 1252
```

R >= 4.2.0

```
l10n_info()
## $MBCS
## [1] TRUE
##
## $`UTF-8`
## [1] TRUE
##
## $`Latin-1`
## [1] FALSE
##
## $codepage
## [1] 65001
## $system.codepage
## [1] 65001
```

R < 4.2.0

```r
x <- "café"

Encoding(x)

x <- iconv(x, from =
Encoding(x), to =
"UTF-8")

Encoding(x)
## [1] "latin1"
## [1] "UTF-8"
```

R >= 4.2.0

```r
x <- "café"

Encoding(x)

x <- iconv(x, from =
Encoding(x), to =
"latin1")

Encoding(x)
## [1] "UTF-8"
## [1] "latin1"
```

Character encoding, Tidyverse style

```
library(tidyverse)

str_conv(string = "café", encoding = "latin1")
str_conv(string = "café", encoding = "UTF-8")
str_conv(string = "café", encoding =
sample(stringi::stri_enc_list(), 1))
```

```
## [1] "cafÃ©"
## [1] "café"
## [1] "cafΓ©"
```

Thank you!

Where to find me:

LinkedIn: https://www.linkedin.com/in/alex-farach/

GitHub: https://github.com/farach