Regex practice

Cedric Arisdakessian

2022-09-07

Before we begin

Syntax review

- `"."`: any character.
- `x|y`: x OR y
- `[xyz]`: x OR y OR z
- `[x-y]`: any character in range (numbers or letters)
- `[^xyz]`: anything but x, y or z
- `"^"`: matches the beginning of the text.
- `"$"`: matches the end of the text
- `"*"`: 0+ repetitions
- `"+"`: 1+ repetitions
- `{n,m}`: between n and m repetitions

Some R functions that use REGEX

- Find pattern in text:
  `grep/grepl(pattern, text)`
- Replace pattern in text
  `gsub(pattern, replacement, text)`

Some bash commands that use REGEX

- Find pattern in file/stream
  `grep [PATTERN] FILE`
- Replace pattern in file/stream
  `sed 's/[PATTERN]/[REPLACEMENT]/g' FILE`
- And a few more

Remarks

- In complex cases, it's better to make multiple regex queries than try to account for everything in a single expression.
- Think about all the cases you could encounter for YOUR data. Don't try to be too broad
- But make sure you're specific enough: you don't want to match things that shouldn't (see example)

Exercises

Word combination (question)

Find any (realistic) combination of two words in a table.

Ex:

- “Little Pond”
- “little_pOnd”

Word combination (solution)

```
ponds <- c("Little Pond", "little_pOnd ", "little.pond",
           "little-pond", "pond")

# General expression
regex <- " *[Ll][Ii][Tt]{2}[Ll][Ee][^A-Za-z0-9][Pp][OOo][Nn][Dd]
grepl(regex, ponds)
```

```
## [1]  TRUE  TRUE  TRUE  TRUE FALSE
```

```
# Simpler
regex <- " *little[^A-Za-z0-9]pond *"
grepl(regex, ponds, ignore.case=T)
```

```
## [1]  TRUE  TRUE  TRUE  TRUE FALSE
```

Note: if we were using grep on a file, we can ignore case with the -i option

Missing values (question)

Handle missing values in a csv formatted file:

1. Find lines with missing entries
2. Remove any missing entry labels

How do we identify missing values in a csv file?

- Special values (“NA”, “NaN”, etc.)
- Empty field (“,,”)

Missing values (solution)

```
content <- c(
  "H,1,1,1,1",
  "He,2,NA,10,20",
  "Na,2,1,10,#VALUE",
  "Cu,,2,5,10",
  ",5,10,2,1"
)

regex <- "(N[Aa]N)|NA|(#VALUE)|(,,)|(^,)|(,$)"
grepl(regex, content)
```

```
## [1] FALSE  TRUE  TRUE  TRUE  TRUE
```

Note: Careful not to match Na (since in this case, it stands for Sodium)

Missing values (solution)

Let's try to do question 2) directly

```
for (line in content) {
  print(sprintf(
    "%s -> %s", line, gsub(regex, "", line)
  ))
}
```

```
## [1] "H,1,1,1,1 -> H,1,1,1,1"
## [1] "He,2,NA,10,20 -> He,2,,10,20"
## [1] "Na,2,1,10,#VALUE -> Na,2,1,10,"
## [1] "Cu,,2,5,10 -> Cu2,5,10"
## [1] ",5,10,2,1 -> 5,10,2,1"
```

Solution: just don't match empty fields

```
gsub("(N[Aa]N)|NA|(#VALUE)", "", content)
```

```
## [1] "H,1,1,1,1"   "He,2,,10,20" "Na,2,1,10,"  "Cu,,2,5,10"  "
```

Phone numbers (question)

How would you find US phone numbers in a text?

Phone numbers (answer)

How would you find US phone numbers in a text?

```
numbers <- c(
  "(808)-000-1111",
  "000-111-2222",
  "0001112222",
  "+33 (0)6 11 22 33 44",
  "808-11-1111"
)

country_code <- "(\\+[0-9]{1,3})?" # optional
area_code <- "\\(?[0-9]{3}\\)?"
regex <- paste(country_code, area_code, "[0-9]{3}", "[0-9]{4}",
               sep="[- ]?")

grepl(regex, numbers)
```

```
## [1]  TRUE  TRUE  TRUE FALSE FALSE
```

Passwords (question)

- Check if a password contains at least one capital letter
- Check if a password contains at least one special character
- Both?