Capturing groups

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Instructor

A regular pattern

```
str_match(
  "payload: 'Adam, 5, 3', headers: 'Auth...'",
  pattern = "[A-Za-z]+, \\d+, \\d+"
)
```

Resulted in:

```
     [,1]
[1,] "Adam, 5, 3"
```

Meet capturing groups

```
str_match(
  "payload: 'Adam, 5, 3', headers: 'Auth...'",
  pattern = "([A-Za-z]+), (\\d+), (\\d+)"
)
```

Results in:

```
     [,1]          [,2]   [,3] [,4]
[1,] "Adam, 5, 3" "Adam" "5"  "3"
```

Replacement

```
str_replace(
  "payload: 'Adam, 5, 3', headers: 'Auth...'",
  pattern = "([A-Za-z]+), (\\d+), (\\d+)",
  replacement = "\\1 tried to log in \\2 times."
)
```

Returns:

```
"payload: 'Adam tried to log in 5 times.', headers: 'Auth...'"`
```

String split

```
str_split(
  "a:b:c:d",
  pattern = ":",
  simplify = FALSE
)
```

```
[[1]]
[1] "a" "b" "c" "d"
```

```
str_split(
  "a:b:c:d",
  pattern = ":",
  simplify = TRUE
)
```

```
     [,1] [,2] [,3] [,4]
[1,] "a"  "b"  "c"  "d"
```

Let's practice!

INTERMEDIATE REGULAR EXPRESSIONS IN R

Tidyr's extract

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Data Journalist

Functions used so far

- `str_match`
- `str_replace`
- `str_match_all`
- `str_replace_all`
- …

Where regular expressions and data frames meet:

```
extract(
    data,
    col,
    into,
    regex = "([[:alnum:]]+)",
    remove = TRUE,
    convert = FALSE,
    ...
 )
```

The arguments of extract

```
extract(
    data,
    col,
    into,
    regex = "([[:alnum:]]+)",
    remove = TRUE,
    convert = FALSE,
    ...
 )
```

- `data`
- `col`
- `into`
- `regex`
- `remove`
- `convert`

Movies data frame

file_source	line
02_11_1	Movie Title Distributor Screens
02_11_1	Karate Kid WDSMP 58
02_11_1	Twilight Saga, The: Eclispe Elite 91
02_11_1	Knight & Day Fox 50
02_11_1	Shrek Forever After (3D) Universal 63
02_11_1	Marmaduke Fox 33
02_11_1	Predators Fox 26
02_11_1	StreetDance (3D) Rialto 11
02_11_1	Robin Hood Universal 9
02_11_2	Micmacs A Tire-Larigot Pathé 4
02_11_2	Sex And the City 2 WB 12
02_11_2	Inception WB 24
02_11_2	Toy Story 3 In Disney Digital 3D WDSMP 25
02_11_2	Shrek Forever After (3D) Universal 22

What we can do with str_match

file_source	line
02_11_1	Movie Title Distributor Screens
02_11_1	Karate Kid WDSMP 58
02_11_1	Twilight Saga, The: Eclispe Elite 91
02_11_1	Knight & Day Fox 50
02_11_1	Shrek Forever After (3D) Universal 63
02_11_1	Marmaduke Fox 33
02_11_1	Predators Fox 26
02_11_1	StreetDance (3D) Rialto 11
02_11_1	Robin Hood Universal 9
02_11_2	Micmacs A Tire-Larigot Pathé 4
02_11_2	Sex And the City 2 WB 12
02_11_2	Inception WB 24
02_11_2	Toy Story 3 In Disney Digital 3D WDSMP 25
02_11_2	Shrek Forever After (3D) Universal 22

```
screens_per_movie %<>%
  mutate(
    is_3d = str_match(line, "3D")
  )
```

What the result of str_match looks like

file_source	line				is_3d
02_11_1	Movie Title	Distributor	Screens		
02_11_1	Karate Kid	WDSMP	58		
02_11_1	Twilight Saga, The: Eclispe	Elite	91		
02_11_1	Knight & Day	Fox	50		
02_11_1	Shrek Forever After (3D)	Universal	63		3D
02_11_1	Marmaduke	Fox	33		
02_11_1	Predators	Fox	26		
02_11_1	StreetDance (3D)	Rialto	11		3D
02_11_1	Robin Hood	Universal	9		
02_11_2	Micmacs A Tire-Larigot	Pathé	4		
02_11_2	Sex And the City 2	WB	12		
02_11_2	Inception	WB	24		
02_11_2	Toy Story 3 In Disney Digital 3D	WDSMP	25		3D
02_11_2	Shrek Forever After (3D)	Universal	22		3D

```
screens_per_movie %<>%
  mutate(
    is_3d = str_match(line, "3D")
  )
```

str_match can only match one information

file_source	line		
02_11_1	Movie Title	Distributor	Screens
02_11_1	Karate Kid	WDSMP	58
02_11_1	Twilight Saga, The: Eclispe	Elite	91
02_11_1	Knight & Day	Fox	50
02_11_1	Shrek Forever After (3D)	Universal	63
02_11_1	Marmaduke	Fox	33
02_11_1	Predators	Fox	26
02_11_1	StreetDance (3D)	Rialto	11
02_11_1	Robin Hood	Universal	9
02_11_2	Micmacs A Tire-Larigot	Pathé	4
02_11_2	Sex And the City 2	WB	12
02_11_2	Inception	WB	24
02_11_2	Toy Story 3 In Disney Digital 3D	WDSMP	25
02_11_2	Shrek Forever After (3D)	Universal	22

file_source	line			is_3d
02_11_1	Movie Title	Distributor	Screens	
02_11_1	Karate Kid	WDSMP	58	
02_11_1	Twilight Saga, The: Eclispe	Elite	91	
02_11_1	Knight & Day	Fox	50	
02_11_1	Shrek Forever After (3D)	Universal	63	3D
02_11_1	Marmaduke	Fox	33	
02_11_1	Predators	Fox	26	
02_11_1	StreetDance (3D)	Rialto	11	3D
02_11_1	Robin Hood	Universal	9	
02_11_2	Micmacs A Tire-Larigot	Pathé	4	
02_11_2	Sex And the City 2	WB	12	
02_11_2	Inception	WB	24	
02_11_2	Toy Story 3 In Disney Digital 3D	WDSMP	25	3D
02_11_2	Shrek Forever After (3D)	Universal	22	3D

This is what extract can do for us

file_source	line
02_11_1	Movie Title Distributor Screens
02_11_1	Karate Kid WDSMP 58
02_11_1	Twilight Saga, The: Eclispe Elite 91
02_11_1	Knight & Day Fox 50
02_11_1	Shrek Forever After (3D) Universal 63
02_11_1	Marmaduke Fox 33
02_11_1	Predators Fox 26
02_11_1	StreetDance (3D) Rialto 11
02_11_1	Robin Hood Universal 9
02_11_2	Micmacs A Tire-Larigot Pathé 4
02_11_2	Sex And the City 2 WB 12
02_11_2	Inception WB 24
02_11_2	Toy Story 3 In Disney Digital 3D WDSMP 25
02_11_2	Shrek Forever After (3D) Universal 22

file_source	line	is_3d	screens
02_11_1	Movie Title Distributor Screens		
02_11_1	Karate Kid WDSMP 58		
02_11_1	Twilight Saga, The: Eclispe Elite 91		
02_11_1	Knight & Day Fox 50		
02_11_1	Shrek Forever After (3D) Universal 63	3D	63
02_11_1	Marmaduke Fox 33		
02_11_1	Predators Fox 26		
02_11_1	StreetDance (3D) Rialto 11	3D	11
02_11_1	Robin Hood Universal 9		
02_11_2	Micmacs A Tire-Larigot Pathé 4		
02_11_2	Sex And the City 2 WB 12		
02_11_2	Inception WB 24		
02_11_2	Toy Story 3 In Disney Digital 3D WDSMP 25	3D	25
02_11_2	Shrek Forever After (3D) Universal 22	3D	22

This is what extract can do for us

file_source	line
02_11_1	Movie Title Distributor Screens
02_11_1	Karate Kid WDSMP 58
02_11_1	Twilight Saga, The: Eclispe Elite 91
02_11_1	Knight & Day Fox 50
02_11_1	Shrek Forever After (3D) Universal 63
02_11_1	Marmaduke Fox 33
02_11_1	Predators Fox 26
02_11_1	StreetDance (3D) Rialto 11
02_11_1	Robin Hood Universal 9
02_11_2	Micmacs A Tire-Larigot Pathé 4
02_11_2	Sex And the City 2 WB 12
02_11_2	Inception WB 24
02_11_2	Toy Story 3 In Disney Digital 3D WDSMP 25
02_11_2	Shrek Forever After (3D) Universal 22

```
extract(
  screens_per_movie,
  col = "line",
  into = c("is_3d", "screens"),
  regex = "(3D).*?(\\d+)$",
  remove = FALSE
 )
```

The result of extract

file_source	line	is_3d	screens
02_11_1	Movie Title Distributor Screens		
02_11_1	Karate Kid WDSMP 58		
02_11_1	Twilight Saga, The: Eclispe Elite 91		
02_11_1	Knight & Day Fox 50		
02_11_1	Shrek Forever After (3D) Universal 63	3D	63
02_11_1	Marmaduke Fox 33		
02_11_1	Predators Fox 26		
02_11_1	StreetDance (3D) Rialto 11	3D	11
02_11_1	Robin Hood Universal 9		
02_11_2	Micmacs A Tire-Larigot Pathé 4		
02_11_2	Sex And the City 2 WB 12		
02_11_2	Inception WB 24		
02_11_2	Toy Story 3 In Disney Digital 3D WDSMP 25	3D	25
02_11_2	Shrek Forever After (3D) Universal 22	3D	22

```
extract(
  screens_per_movie,
  col = "line",
  into = c("is_3d", "screens"),
  regex = "(3D).*?(\\d+)$",
  remove = FALSE
)
```

Let's practice!

INTERMEDIATE REGULAR EXPRESSIONS IN R

Extracting matches and surroundings from a text

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Instructor

Mentions of a company name

"...got to the store. Super smooth and seamless experience. Great value. I would highly recommend **ABC Enterprises** and I will be coming back for sure! Next, we went..."

One word: `(\\w+\\s)` , 0 to 10 words: `(\\w+\\s){0,10}`

```
str_extract_all(
  blog_post,
  pattern = "(\\w+\\s){0,10}ABC Enterprises\\s?(\\w+\\s){0,10}"
)
```

Returns: `"I would highly recommend ABC Enterprises and I will be coming back for"`

Punctuation

"...got to the store. Super smooth and seamless experience. Great value. I would highly recommend **ABC Enterprises** and I will be coming back for sure! Next, we went..."

Extracted: `"I would highly recommend ABC Enterprises and I will be coming back for"`

Replace `\\w+` with `[\\w[:punct:]]+`

Extracted:

```
"smooth and seamless experience. Great value. I would highly recommend ABC Enterprises
and I will be coming back for sure! Next, we "
```

Let's practice!

INTERMEDIATE REGULAR EXPRESSIONS IN R