Regular Expressions (part 2)

Stat 133 with Gaston Sanchez

POSIX Classes

Some POSIX character classes

`[[:alpha:]]`	Lower and upper case letter
`[[:lower:]]`	Lower case letter
`[[:upper:]]`	Upper case letter
`[[:digit:]]`	Digit
`[[:alnum:]]`	Letter or digit
`[[:xdigit:]]`	Hexadecimal digit
`[[:punct:]]`	Punctuation symbol
`[[:space:]]`	White space

Character Classes

Character Classes

`\\d`	Any digit
`\\w`	Any word character
`\\s`	Any whitespace character
`\\D`	Any non-digit
`\\W`	Any non-word character
`\\S`	Any non-whitespace character

Anchors

Anchors

^	Beginning of string *e.g. begin with vowel* **`^[aeiou]`**
$	End of string *e.g. end with vowel* **`[aeiou]$`**

Quantifiers

Quantifiers: repetition metacharacters

?	“One optional”	One allowed; none required
*****	“Any amount OK”	Unlim allowed; none required
+	“At least one”	Unlim allowed; one required

They act on the immediately-preceding item

Quantifiers: intervals

`{min,max}`	Interval including min and max
`{min,}`	Min required, no max limit
`{,max}`	No min required, max limit
`{n}`	At least n times

They act on the immediately-preceding item

Regex in R with the package "`stringr`"

Some stringr functions for regex

`str_detect()`	Whether there's a pattern match
`str_match()`	Indicates where the match occurs
`str_locate()`	Positions of the matched pattern
`str_extract()`	Extraction of the matched pattern
`str_replace()`	Replacement of a pattern by another pattern
`str_split()`	Splitting a string based on a pattern

Some basic examples

```
# names containing numbers
starwars$name[str_detect(starwars$name, "[0-9]")]

# names containing dashes
starwars$name[str_detect(starwars$name, "\\w")]

# names containing space
starwars$name[str_detect(starwars$name, "\\s")]

# names containing no whitespaces
starwars$name[!str_detect(starwars$name, "\\s")]
```

```r
# names that don't have vowels
starwars$name[!str_detect(starwars$name, "[aeiou]")]


# names that have 2 or more consecutive vowels
starwars$name[str_detect(starwars$name,
"[aeiou][aeiou]")]


# names that have 3 or more consecutive vowels
starwars$name[str_detect(starwars$name, "[aeiou]{3}")]


# names that have exactly 2 consecutive vowels
starwars$name[str_detect(starwars$name,
                         "[^aeiou][aeiou]{2}[^aeiou]")]
```

```
cols <- colors()

# gray or grey colors
cols[str_detect(cols, "grey")]
cols[str_detect(cols, "gray")]
cols[str_detect(cols, "gray|grey")]
cols[str_detect(cols, "gr[ea]y")]

# grays with various 4's
cols[str_detect(cols, "gr[ea]y4")]
cols[str_detect(cols, "gr[ea]y4[0-9]")]
cols[str_detect(cols, "^gr[ea]y")]
```

```
# letter followed by gra/ey
cols[str_detect(cols, "[a-z]gr[ea]y")]

# at least one letter, then gra/ey
cols[str_detect(cols, "[a-z]+gr[ea]y")]

# zero or one (one optional)
cols[str_detect(cols, "[a-z]?gr[ea]y")]

# zero or none (any amount OK)
cols[str_detect(cols, "[a-z]*gr[ea]y")]
```