



STRING MANIPULATION WITH STRINGR

Case study



“The truth is rarely pure and never simple.”

*The Importance of Being Earnest,
A Trivial Comedy for Serious People* by Oscar Wilde

Your task: Read the play and count the number of lines each character has.

readLines()

```
> old_mac <- readLines("old_mac.txt")

> str(old_mac)
chr [1:7] "Old MacDonald had a farm" ...

> old_mac[1:2]
[1] "Old MacDonald had a farm" "E-I-E-I-O"

> str_detect(old_mac, "moo")
[1] FALSE FALSE FALSE FALSE TRUE FALSE FALSE

> which(str_detect(old_mac, "moo"))
[1] 5
```

old_mac.txt

```
Old MacDonald had a farm
E-I-E-I-O
And on his farm he had a cow
E-I-E-I-O
Here a moo, There a moo,
Everywhere a moo-moo
Old MacDonald had a farm
E-I-E-I-O
```

Alternatively: `stringi::stri_read_lines()`



String Manipulation in R with stringr

Let's practice!



STRING MANIPULATION WITH STRINGR

A case study on case

regex are case sensitive

- "dog" won't match "Dog"
- Accidents involving cats: catcidents

```
"87YOF TRIPPED OVER CAT, HIT LEG ON STEP. DX LOWER LEG  
CONTUSION "
```

```
"unhelmeted 14yof riding her bike with her dog when she  
saw a cat and sw erved c/o head/shoulder/elbow pain.dx:  
minor head injury,left shoulder"
```

```
"44Yof Walking Dog And The Dof Took Off After A Cat And  
Pulled Pt Down B Y The Leash Strained Neck"
```

```
"LEFT KNEE cOntusioN.78YOf triPPed OVEr CaT aND fell  
and hIt knEE ON the fLoOr."
```

Case sensitive matching

```
> str_subset(catcidents, "food")
[1] "3Yof-foot lac-cut on cat food can-@ home "
[2] "4 Yom was cut on cat food can. Dx:  r index lac 1 cm."

> str_subset(catcidents, "Food")
[1] "17Yof Cut Right Hand On A Cat Food Can - Laceration "
[2] "Pt Lifted Bag Of Cat Food. Dx:  Low Back Px, Hx Arthritic
Spine."

> str_subset(catcidents, "f00d")
[1] "LaC FInGer oN a meTAL Cat f00d CaN "
```

Change case of input

```
> str_subset(str_to_lower(catcidents), "food")

[1] "21 yof reports sus laceration of her left hand when she ..."
[2] "3yof-foot lac-cut on cat food can-@ home "
[3] "15 mo m cut finger on cat food can lid. dx:  r index lac..."
[4] "accidentally cut finger while opening a cat food can,..."
[5] "4 yom was cut on cat food can. dx:  r index lac 1 cm."
[6] "17yof cut right hand on a cat food can - laceration "
[7] "50yof cut finger on cat food can lid.  dx: lt ring ..."
[8] "lac finger on a metal cat food can "
[9] "10 yo female opening a can of cat food.  dx hand ..."
[10] "pt lifted bag of cat food. dx:  low back px, hx ..."
```


Use case insensitive matching

```
> str_subset(catcidents, regex("food", ignore_case = TRUE))

[1] "21 YOF REPORTS SUS LACERATION OF HER LEFT HAND WHEN SHE ..."
[2] "3Yof-foot lac-cut on cat food can-@ home "
[3] "15 m0 m cut FinGer ON cAT FoOd CAn LID. Dx:  r INDeX laC..."
[4] "ACCIDENTALLY CUT FINGER WHILE OPENING A CAT FOOD CAN, ..."
[5] "4 Yom was cut on cat food can. Dx:  r index lac 1 cm."
[6] "17Yof Cut Right Hand On A Cat Food Can - Laceration "
[7] "50YOF CUT FINGER ON CAT FOOD CAN LID.  DX: LT RING ..."
[8] "LaC FInGer oN a meTAL Cat f00d CaN "
[9] "10 YO FEMALE OPENING A CAN OF CAT FOOD.  DX HAND ..."
[10] "Pt Lifted Bag Of Cat Food. Dx:  Low Back Px, Hx ..."
```



STRING MANIPULATION WITH STRINGR

Let's practice!



STRING MANIPULATION WITH STRINGR

Wrapping up

Next steps

- Look at other `stringr` functions
 - <http://www.rdocumentation.org/packages/stringr>
- Look to `stringi` when `stringr` doesn't solve your problem
 - Functions start with `stri_`
- Regular expressions
 - <http://www.regular-expressions.info>
 - Mastering Regular Expressions by Jeffrey Friedl
 - <http://r4ds.had.co.nz/strings.html#matching-patterns-with-regular-expressions>

Next steps

- *Text Mining: Bag of Words* course
- Analyze text for insights



STRING MANIPULATION WITH STRINGR

Thanks!