Assignment 4

Jack Scherer

2024-10-21

Exercise 1

For the following regular expression, explain in words what it matches on. Then add test strings to demonstrate that it in fact does match on the pattern you claim it does. Make sure that your test set of strings has several examples that match as well as several that do not. Show at least two examples that return TRUE and two examples that return FALSE. *If you copy the Rmarkdown code for these exercises directly from my source pages, make sure to remove the* `eval=FALSE` *from the R-chunk headers.*

Here is an example of what a solution might look like.

q) This regular expression matches:

Any string that contains the lower-case letter "a".

```
strings <- c('Adel', 'Mathematics', 'able', 'cheese')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, 'a') )
```

```
##        string result
## 1        Adel  FALSE
## 2 Mathematics   TRUE
## 3        able   TRUE
## 4      cheese  FALSE
```

Please complete the questions below.

a) This regular expression matches:

Any string that contains a lower-case "a" followed by a lower case "b".

```
strings <- c('ability','blame','arbitrary','dab')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, 'ab') )
```

b) This regular expression matches:

Any string that contains "a" or "b".

```
strings <- c('bed','computer','water','note','Arnold')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, '[ab]') )
```

c) This regular expression matches:

```
strings <- c('Betsy','betsy','tame','apple')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, '^[ab]') )
```

d) This regular expression matches:

All strings with one or more digits followed by a white space followed by either a lower-case or upper-case "a".

```
strings <- c('42 apples','42apples','42apples today','I ate 3 Apples')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, '\\d+\\s[aA]') )
```

e) This regular expression matches:

All strings with one or more digits followed by an optional white space followed by either a lower-case or upper-case "a".

```
strings <- c('42 apples','42apples','42apples today','I ate 3 Apples','I ate 3 Bananas','apples 42')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, '\\d+\\s*[aA]') )
```

f) This regular expression matches:

Any string with zero or more characters.

```
strings <- c('apple','banana','k','3',NA)
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, '.*') )
```

g) This regular expression matches:

All strings that start with 2 alphanumeric characters followed by "bar".

```
strings <- c('fubar','FUBAR','33bar','#3bar','jacob','barstool')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, '^\\w{2}bar') )
```

h) This regular expression matches:

Any string that starts with "foo", followed by ".", followed by "bar" OR any string starting with two alphanumeric characters followed by "bar".

```
strings <- c('foo.bar','foo bar','fubar','foo-bar')
data.frame( string = strings ) %>%
  mutate( result = str_detect(string, '(foo\\.bar)|(^\\w{2}bar)') )
```

Exercise 2 {-}

The following file names were used in a camera trap study. The S number represents the site, P is the plot within a site, C is the camera number within the plot, the first string of numbers is the YearMonthDay and the second string of numbers is the HourMinuteSecond.

```
file.names <- c( 'S123.P2.C10_20120621_213422.jpg',
                 'S10.P1.C1_20120622_050148.jpg',
                 'S187.P2.C2_20120702_023501.jpg')
```

Produce a data frame with columns corresponding to the `site`, `plot`, `camera`, `year`, `month`, `day`, `hour`, `minute`, and `second` for these three file names. So we want to produce code that will create the data frame:

```
 Site Plot Camera Year Month Day Hour Minute Second
 S123   P2    C10 2012    06  21   21     34     22
  S10   P1     C1 2012    06  22   05     01     48
 S187   P2     C2 2012    07  02   02     35     01
```

```
file.names <-
  file.names %>% str_replace_all( pattern='_', replacement='.' )

split.list <-
  file.names %>% str_split_fixed( n=6, pattern='\\.' )

final <- data.frame( string=file.names ) %>%
  mutate( Site = split.list[,1] ) %>%
  mutate( Plot = split.list[,2] ) %>%
  mutate( Camera = split.list[,3] ) %>%
  mutate( Year = split.list[,4] %>% str_sub(start=1, end=4) ) %>%
  mutate( Month = split.list[,4] %>% str_sub(start=5, end=6) ) %>%
  mutate( Day = split.list[,4] %>% str_sub(start=7, end=8) ) %>%
  mutate( Hour = split.list[,5] %>% str_sub(start=1, end=2) ) %>%
  mutate( Minute = split.list[,5] %>% str_sub(start=3, end=4) ) %>%
  mutate( Second = split.list[,5] %>% str_sub(start=5, end=6) )

final
```

```
##                              string Site Plot Camera Year Month Day Hour Minute
## 1 S123.P2.C10.20120621.213422.jpg S123   P2    C10 2012    06  21   21     34
## 2   S10.P1.C1.20120622.050148.jpg  S10   P1     C1 2012    06  22   05     01
## 3  S187.P2.C2.20120702.023501.jpg S187   P2     C2 2012    07  02   02     35
##   Second
## 1     22
## 2     48
## 3     01
```

Exercise 3

The full text from Lincoln's Gettysburg Address is given below. It has been provided in a form that includes lots of different types of white space. Your goal is to calculate the mean word length of Lincoln's Gettysburg Address! *Note: you may consider 'battle-field' as one word with 11 letters or as two words 'battle' and 'field'. The first option a bit more difficult and technical!.*

```
Gettysburg <- 'Four score and seven years ago our fathers brought forth on this
continent, a new nation, conceived in Liberty, and dedicated to the proposition
that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any
```

```
nation so conceived and so dedicated, can long endure. We are met on a great
battle-field of that war. We have come to dedicate a portion of that field, as
a final resting place for those who here gave their lives that that nation might
live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate -- we can not consecrate -- we can
not hallow -- this ground. The brave men, living and dead, who struggled here,
have consecrated it, far above our poor power to add or detract. The world will
little note, nor long remember what we say here, but it can never forget what
they did here. It is for us the living, rather, to be dedicated here to the
unfinished work which they who fought here have thus far so nobly advanced. It
is rather for us to be here dedicated to the great task remaining before us --
that from these honored dead we take increased devotion to that cause for which
they gave the last full measure of devotion -- that we here highly resolve that
these dead shall not have died in vain -- that this nation, under God, shall
have a new birth of freedom -- and that government of the people, by the people,
for the people, shall not perish from the earth.'
```

```
Gettysburg <- Gettysburg %>%
  str_replace_all( pattern='\\-', replacement='' ) %>%
  str_replace_all( pattern='\\,', replacement='' ) %>%
  str_replace_all( pattern='\\.', replacement='' )

words <- str_split( Gettysburg, pattern=' ' )
lengths <- str_length( words[[1]] )
mean( lengths )
```

```
## [1] 4.23913
```