Assignment 3

Munkhnaran Gankhuyag

September 17, 2017

Question 3

```
library(stringr)
raw.data <- "555-1239Moe Szyslak(636) 555-0113Burns, C. Montgomery555
-6542Rev. Timothy Lovejoy555 8904Ned Flanders,
5553642Dr. Julius Hibbert636-555-3226Simpson, Homer"

name <- unlist(str_extract_all(raw.data, "[[:alpha:]., ]{2,}"))

name
```

```
## [1] "Moe Szyslak"          "Burns, C. Montgomery" "Rev. Timothy Lovejoy"
## [4] "Ned Flanders,"        "Dr. Julius Hibbert"   "Simpson, Homer"
```

name_number <- data.frame(names = name, number = phone) name_number

```
switchnames <- str_split(name, ", ")
switchnames <- sapply(switchnames, function(x) str_replace(paste(x[2], x[1]),"^NA ",""))
titles <-"^(Rev|Dr)\\."
first_last <- str_replace(switchnames, titles,"")

first_last
```

```
## [1] "Moe Szyslak"         "C. Montgomery Burns" " Timothy Lovejoy"
## [4] "Ned Flanders,"       " Julius Hibbert"     "Homer Simpson"
```

```
str_detect(switchnames,titles)
```

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

```
first_last[str_detect(first_last,"\\b[:alnum:]\\b")]
```

```
## [1] "C. Montgomery Burns"
```

Question 4

(a) [0-9]+\$ is used for number of digits followed by the dollar sign.

```
question4a <-c("917$", "data606","data607$", "assignment3$",
"data1234$67")

str_extract(question4a,"[0-9]+\\$")
```

```
## [1] "917$"  NA      "607$"  "3$"    "1234$"
```

(b) \b[a-z]{1,4}\b is used to extract the first lower case word that is between 2-4 characters.

```
question4b <- c('Data 607 hw', 'data 607 hw', 'Data six zero seven hw')

str_extract(question4b, "\\b[a-z]{1,4}\\b")
```

```
## [1] "hw"   "data" "six"
```

(c) .*?\.txt$ is used to extract txt files. It takes any values followed by a .txt.

```
question4c <- c('data607.txt', 'Sept2017.txt', 'data607.csv', 'data.txt')
str_extract(question4c, ".*?\\.txt$")
```

```
## [1] "data607.txt"  "Sept2017.txt" NA             "data.txt"
```

(d) \d{2}/\d{2}/\d{4} is used to extract date values with 2 digit days, 2 digit months and 4 digit years.

```
question4d <- c('9/17/2017', '09/17/2017.txt', '09/17/2017', '09/7/2017', '9/17/17')
str_extract(question4d, "\\d{2}/\\d{2}/\\d{4}")
```

```
## [1] NA           "09/17/2017" "09/17/2017" NA           NA
```

(e) <(.+?)>.+?</\1> is used to extract for values with matching and ending with the same characters.

```
question4e <- c('<Data.607>', '<a>data607</a>', '<a>data607</ab>', 607, '<607>data607.txt</607>')
str_extract(question4e, "<(.+?)>.+?</\\1>")
```

```
## [1] NA                       "<a>data607</a>"
## [3] NA                       NA
## [5] "<607>data607.txt</607>"
```

Question 9

The following code hides a secret message. Crack it with R and regular expressions. Hint: Some of the characters are more revealing than others! The code snippet is also available in the materials at www.r-datacollection.com.

```
Hidden_message <- "clcopCow1zmstc0d87wnkig7OvdicpNuggvhryn92Gjuwczi8hqrfpRxs5Aj5dwpn0TanwoUwisdij7Lj8kp
```

```
str_extract_all(Hidden_message,"[[a-z]]")
```

```
## [[1]]
##   [1] "c" "l" "c" "o" "p" "o" "w" "z" "m" "s" "t" "c" "d" "w" "n" "k" "i"
##  [18] "g" "v" "d" "i" "c" "p" "u" "g" "g" "v" "h" "r" "y" "n" "j" "u" "w"
##  [35] "c" "z" "i" "h" "q" "r" "f" "p" "x" "s" "j" "d" "w" "p" "n" "a" "n"
##  [52] "w" "o" "w" "i" "s" "d" "i" "j" "j" "k" "p" "f" "d" "r" "c" "o" "c"
##  [69] "b" "t" "y" "c" "z" "j" "a" "t" "a" "o" "o" "t" "j" "t" "j" "n" "e"
##  [86] "c" "f" "e" "k" "r" "w" "w" "w" "o" "j" "i" "g" "d" "v" "r" "f" "r"
## [103] "b" "z" "b" "k" "n" "b" "h" "z" "g" "v" "i" "z" "c" "r" "o" "p" "w"
## [120] "g" "n" "b" "q" "o" "f" "a" "o" "t" "f" "b" "w" "m" "k" "t" "s" "z"
## [137] "q" "e" "f" "y" "n" "d" "t" "k" "c" "f" "g" "m" "c" "g" "x" "o" "n"
## [154] "h" "k" "g" "r"
```

```
str_extract_all(Hidden_message,"[[:digit:]]")
```

```
## [[1]]
##  [1] "1" "0" "8" "7" "7" "9" "2" "8" "5" "5" "0" "7" "8" "0" "3" "5" "3"
## [18] "0" "7" "5" "5" "3" "3" "6" "4" "1" "1" "6" "2" "2" "4" "9" "0" "5"
## [35] "6" "5" "1" "7" "2" "4" "6" "3" "9" "5" "8" "9" "6" "5" "9" "4" "9"
## [52] "0" "5" "4" "5"
```

```r
str_extract_all(Hidden_message, "[[A-Z]]")
```

```
## [[1]]
##  [1] "C" "O" "N" "G" "R" "A" "T" "U" "L" "A" "T" "I" "O" "N" "S" "Y" "O"
## [18] "U" "A" "R" "E" "A" "S" "U" "P" "E" "R" "N" "E" "R" "D"
```