IS 607: WEEK 4 ASSIGNMENT SOLUTION

MUSA T. GANIYU

February 20, 2016

3. We load the data from example given in chapter 8 of Automated Data Collection with R (page 196).

```
data <- "555-123Moe Szyslak (636) 555-0113Burns, C. Montgomery555-6542Rev. Timothy Lovejoy555 8904Ned Fl

library(stringr);

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

name;
```

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

```
 # Rearrange the vector to so that all element conform to the standard first_name, last_name.

sort(name, partial = NULL, na.last = NA, decreasing = FALSE,
     method = c("first_name", "last_name"), index.return = FALSE);
```

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

```
# Vector indicating wether a character has a title ( i.e Rev. and Dr.)

str_extract(name, ("Dr.|Rev."));
```

```
## [1] NA     NA     "Rev." NA     NA     "Dr."
```

```
str_detect(name, ("Dr.|Rev."));
```

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

```
# Vector indicating wether a character has a second name.

str_detect(name, ("second name"));
```

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

4. Consider the string < title>+++BREAKING NEWS+++
   . We would like to extract the first HTML tag. To do so we write the regular expression <.+>. Explain why this fail and correct the expression.

```
# note that this is HTML with + as COMMON QUANTIFICATION OPERATOR, "." as character to extract order in



html_tag <- "< title>+++BREAKING NEWS+++</title>";
str_extract(html_tag, "<.+>");
```

```
## [1] "< title>+++BREAKING NEWS+++</title>"
```

```
# This is a Greedy Quantification; We Correct this by adding the operator "?" after operator "+".

str_extract(html_tag, "<.+?>");
```

```
## [1] "< title>"
```

8. Consider the string (5-3)$^{2=5}$2-2*5*3+3 conforms to the binomial theorem. We would like to extract the formula in the string. To do so we write the regular expression [ˆ0-9=+*()] +.Explain why this fails and correct the expression.

```
data2 <- "(5-3)^2=5^2-2*5*3+3 conforms to the binomial theorem.";

str_extract(data2, "[^0-9=+*()]+");
```

```
## [1] "-"
```

```
# The "^" raises all the characters at its end, and the "-" makes an inclusion in the character class.

str_extract(data2, "[0-9=+*()^]+");
```

```
## [1] "(5"
```

```
str_extract(data2, "[0-9=+*()^-]+")
```

```
## [1] "(5-3)^2=5^2-2*5*3+3"
```