Regular Expressions

CHINDU

The dataset used is from Kaggle (Women's E-Commerce Clothing Reviews). The dataset contains 23000 customer reviews and ratings.

```r
library(stringr)
data<- read.csv("Womens Clothing E-Commerce Reviews.csv")
```

Using Regexp to get some understanding about our data and to replace, extract or count words.

```r
# Print data which contains a number (by doing this we can filter only reviews were customer mention ab
head(grep(pattern="\\d",x=data$Review.Text,value=TRUE))
```

```
## [1] "Love this dress!  it's sooo pretty.  i happened to find it in a store, and i'm glad i did bc i r
## [2] "I love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tall
## [3] "I aded this in my basket at hte last mintue to see what it would look like in person. (store pic
## [4] "I'm 5\"5' and 125 lbs. i ordered the s petite to make sure the length wasn't too long. i typical
## [5] "Bought the black xs to go under the larkspur midi dress because they didn't bother lining the sh
## [6] "This is a nice choice for holiday gatherings. i like that the length grazes the knee so it is co
```

```r
# Find all items with a number followed by a space
head(grep(pattern = "\\d\\s", x = data$Review.Text))
```

```
## [1]  6  7 10 14 15 17
```

```r
# How many times was the word 'favorite' used?
length(grep(pattern = "favorite", x = data$Review.Text))
```

```
## [1] 569
```

```r
# Replacing words/punctuations, example ! with .
head(gsub(pattern = '!', replacement = '.', x = data$Review.Text))
```

```
## [1] "Absolutely wonderful - silky and sexy and comfortable"
## [2] "Love this dress.  it's sooo pretty.  i happened to find it in a store, and i'm glad i did bc i r
## [3] "I had such high hopes for this dress and really wanted it to work for me. i initially ordered th
## [4] "I love, love, love this jumpsuit. it's fun, flirty, and fabulous. every time i wear it, i get ne
## [5] "This shirt is very flattering to all due to the adjustable front tie. it is the perfect length t
## [6] "I love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tall
```

```r
# Replace all occurences of "it" with 'Dress '.
head(gsub(pattern = 'it\\s', replacement = 'the dress ', x =data$Review.Text ))
```

```
## [1] "Absolutely wonderful - silky and sexy and comfortable"
## [2] "Love this dress!  it's sooo pretty.  i happened to find the dress in a store, and i'm glad i di
## [3] "I had such high hopes for this dress and really wanted the dress to work for me. i initially or
## [4] "I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get n
## [5] "This shirt is very flattering to all due to the adjustable front tie. the dress is the perfect
## [6] "I love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tal
```

```r
# Replace all occurences of "it's" with 'It is'
head(gsub(pattern = "it\\'s", replacement = 'it is ', x = data$Review.Text))
```

```
## [1] "Absolutely wonderful - silky and sexy and comfortable"
## [2] "Love this dress!  it is  sooo pretty.  i happened to find it in a store, and i'm glad i did bc
## [3] "I had such high hopes for this dress and really wanted it to work for me. i initially ordered t
## [4] "I love, love, love this jumpsuit. it is  fun, flirty, and fabulous! every time i wear it, i get
## [5] "This shirt is very flattering to all due to the adjustable front tie. it is the perfect length
## [6] "I love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tal
```

```r
# Convert to lower (conver all to lower letter)
head(tolower(data$Review.Text))
```

```
## [1] "absolutely wonderful - silky and sexy and comfortable"
## [2] "love this dress!  it's sooo pretty.  i happened to find it in a store, and i'm glad i did bc i
## [3] "i had such high hopes for this dress and really wanted it to work for me. i initially ordered t
## [4] "i love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get n
## [5] "this shirt is very flattering to all due to the adjustable front tie. it is the perfect length
## [6] "i love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tal
```

```r
# Extract parts of string
head(substr(x=data$Review.Text, start=1,stop=10))
```

```
## [1] "Absolutely" "Love this " "I had such" "I love, lo" "This shirt"
## [6] "I love tra"
```

```r
# Find and replace first match
head(sub(pattern = "L",replacement = "B",x = data$Review.Text,ignore.case = T))
```

```
## [1] "AbsoButely wonderful - silky and sexy and comfortable"
## [2] "Bove this dress!  it's sooo pretty.  i happened to find it in a store, and i'm glad i did bc i
## [3] "I had such high hopes for this dress and reaBly wanted it to work for me. i initially ordered t
## [4] "I Bove, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get n
## [5] "This shirt is very fBattering to all due to the adjustable front tie. it is the perfect length
## [6] "I Bove tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tal
```

```r
# Find and replace all matches
head(gsub(pattern = "Lo",replacement = "Ha",x = data$Review.Text,ignore.case = T))
```

```
## [1] "Absolutely wonderful - silky and sexy and comfortable"
## [2] "Have this dress!  it's sooo pretty.  i happened to find it in a store, and i'm glad i did bc i r
## [3] "I had such high hopes for this dress and really wanted it to work for me. i initially ordered th
## [4] "I Have, Have, Have this jumpsuit. it's fun, flirty, and fabuHaus! every time i wear it, i get no
## [5] "This shirt is very flattering to all due to the adjustable front tie. it is the perfect length f
## [6] "I Have tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tall
```

```r
# All reviews with the word love one or more times
head(grep(pattern="love+",data$Review.Text,value = T))
```

```
## [1] "Love this dress!  it's sooo pretty.  i happened to find it in a store, and i'm glad i did bc i r
## [2] "I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get no
## [3] "This shirt is very flattering to all due to the adjustable front tie. it is the perfect length f
## [4] "I love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tall
## [5] "I love this dress. i usually get an xs but it runs a little snug in bust so i ordered up a size
## [6] "I'm 5\"5' and 125 lbs. i ordered the s petite to make sure the length wasn't too long. i typical
```

```r
# All reviews with the word love excatly 2 times
head(grep(pattern="love{2}",x=data$Review.Text,value=T))
```

```
## [1] "I loveeeeeeee this dress! i saw it online and immediately loved the blue/ orange large floral pi
## [2] "This dress turned out to be a huge disappointment!!! i was in loveee with the picture! in real l
```

Simple examples to understand Regexp

```r
string <- "There are 20 sweets in the bag, 5 are for John"
# Replace numbers by _
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gsub(pattern = "\\d+",replacement = "_",x = string)
```

```
## [1] "There are _ sweets in the bag, _ are for John"
```

```r
# Extract the first number from a string
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
regmatches(string,regexpr(pattern = "\\d+",text = string))
```

```
## [1] "20"
```

```r
# Extract all numbers
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
regmatches(x = string,gregexpr("[0-9]+",text = string))
```

```
## [[1]]
## [1] "20" "5"
```

```r
# Get digits
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
unlist(regmatches(string,gregexpr("[[:digit:]]+",text = string)))
```

```
## [1] "20" "5"
```

```r
# Match a space - returns positions
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gregexpr(pattern = "\\s+",text = string)
```

```
## [[1]]
##  [1]  6 10 13 20 23 27 32 34 38 42
## attr(,"match.length")
##  [1] 1 1 1 1 1 1 1 1 1 1
## attr(,"index.type")
## [1] "chars"
## attr(,"useBytes")
## [1] TRUE
```

```r
# Match a non space
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gsub(pattern = "\\S+",replacement = "app",x = string)
```

```
## [1] "app app app app app app app app app app app"
```

```r
# Match a word character
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gsub(pattern = "\\w",replacement = "k",x = string)
```

```
## [1] "kkkkk kkk kk kkkkkk kk kkk kkk, k kkk kkk kkkk"
```

```r
# Match a non-word character
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gsub(pattern = "\\W",replacement = "k",x = string)
```

```
## [1] "Therekarek20ksweetskinkthekbagkk5karekforkJohn"
```

```r
# Extract without digits
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
regmatches(x = string,gregexpr("[^0-9]+",text = string))
```

```
## [[1]]
## [1] "There are "           " sweets in the bag, " " are for John"
```

```r
# Remove punctuations
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gsub(pattern = "[[:punct:]]+",replacement = "",x = string)
```

```
## [1] "There are 20 sweets in the bag 5 are for John"
```

```r
# Remove spaces
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gsub(pattern = "[[:blank:]]",replacement = "-",x = string)
```

```
## [1] "There-are-20-sweets-in-the-bag,-5-are-for-John"
```

```r
# Remove non graphical characters
string
```

```
## [1] "There are 20 sweets in the bag, 5 are for John"
```

```r
gsub(pattern = "[^[:graph:]]+",replacement = "",x = string)
```

```
## [1] "Thereare20sweetsinthebag,5areforJohn"
```

```
# Extract email addresses from a given string
string <- c("My email address is CHINDU@hotmail.com","address is john@hotmail.com","aescher koeif",
            "paul renne","randomguy@gmail.com")
string
```

```
## [1] "My email address is CHINDU@hotmail.com"
## [2] "address is john@hotmail.com"
## [3] "aescher koeif"
## [4] "paul renne"
## [5] "randomguy@gmail.com"
```

```
unlist(regmatches(x = string, gregexpr(pattern = "[[:alnum:]]+\\@[[:alpha:]]+\\.com",text = string)))
```

```
## [1] "CHINDU@hotmail.com"  "john@hotmail.com"    "randomguy@gmail.com"
```

```
# Extract the minimum number from each range
x <- c("15 to 30", "31 to 45", "46 to 80")
x
```

```
## [1] "15 to 30" "31 to 45" "46 to 80"
```

```
gsub(" .*\\d+", "", x)
```

```
## [1] "15" "31" "46"
```

```
# Extract information inside brackets in a string
string <- "This is an important message (Call me ASAP)"
string
```

```
## [1] "This is an important message (Call me ASAP)"
```

```
gsub("[\\(\\)]","",regmatches(string, gregexpr("\\(.*?\\)", string))[[1]])
```

```
## [1] "Call me ASAP"
```

```
# Remove digits from a string which contains alphanumeric characters
c2 <- "In the competition held on the 2nd of April 02042020, John came in 1st. "
c2
```

```
## [1] "In the competition held on the 2nd of April 02042020, John came in 1st. "
```

```
gsub(pattern = "\\b\\d+\\b",replacement = "",x = c2)
```

```
## [1] "In the competition held on the 2nd of April , John came in 1st. "
```

```
# Remove punctuation from a line of text
going <- "Hey! what are you doing? It's crazy here."
going
```

```
## [1] "Hey! what are you doing? It's crazy here."
```

```r
gsub(pattern = "[[:punct:]]+",replacement = "",x = going)
```

```
## [1] "Hey what are you doing Its crazy here"
```

```r
# In a key value pair, extract the values
string = c("G1:E001", "G2:E002", "G3:E003")
string
```

```
## [1] "G1:E001" "G2:E002" "G3:E003"
```

```r
gsub(pattern = ".*:",replacement = "",x = string)
```

```
## [1] "E001" "E002" "E003"
```

```r
# Extract strings which are available in key value pairs
d <- c("(monday :: 0.1231313213)","tomorrow","(tuesday :: 0.1434343412)")
d
```

```
## [1] "(monday :: 0.1231313213)"  "tomorrow"
## [3] "(tuesday :: 0.1434343412)"
```

```r
grep(pattern = "\\([a-z]+ :: (0\\.[0-9]+)\\)",x = d,value = T)
```

```
## [1] "(monday :: 0.1231313213)"  "(tuesday :: 0.1434343412)"
```