HW4 Applied Data Science

Fan Heng fh2294
Columbia University

13 April, 2014

1 Question One: Phonenumber Type

In a text data file that contains phone number in the same format that was given in class (see example 16 in Rlog for lectures 15-16), write code to find all types of phone numbers.

1.1 General Assignment

Read problem1.txt and concatenate all data into one line separated by space.

```
> Phonenumber <- scan("/Users/fanheng/Dropbox/Columbia 2014Spring/W4249-Applied Data Science/Homewor
> Phonenumber <- paste(Phonenumber, collapse = " ")
> Num.pattern<-"([[:alpha:]]+)[:]?[ ]*[1]?[ ]*[-. ]?[2-9][0-9]{2}[- .][0-9]{3}[- .][0-9]{4}"
> require(stringr)
```

Find the data that match Num.pattern and then find the types of them.

```
> Num.type.match <- str_match_all(Phonenumber, Num.pattern)
> Num.type.ls <- matrix(unlist(Num.type.match), ncol=2)[,2]
> Num.type <- unique(tolower(Num.type.ls))
> Num.type
```

```
[1] "work" "home" "cell"
```

The types of phone number are: ”work”, ”home”, ”cell”

1.2 Extra Credit

no parenthesis are expected but you could attempt to find numbers with area code in parenthesis. Please provide separate code for extra credit and clearly mark is as such.

To take into consideration of parenthesis, I redefine the phone number pattern.

```
> Num.pattern<-"([[:alpha:]]+)[:]*[ ]*[\\(]?[1]?[-]?[2-9][0-9]{2}[\\)]?[- .][0-9]{3}[- .][0-9]{4}"
> require(stringr)
```

Find the data that match Num.pattern and then find the types of them.

```
> Num.type.match <- str_match_all(Phonenumber, Num.pattern)
> Num.type.ls <- matrix(unlist(Num.type.match), ncol=2)[,2]
> Num.type <- unique(tolower(Num.type.ls))
> Num.type
```

```
[1] "work" "home" "cell"
```

The types of phone number are: ”work”, ”home”, ”cell”.

2 Question Two: Expression Evaluation

Write code to replace all expression of type âĂIJa+/-bâĂİ, where a and b are whole numbers with the evaluated number.

2.1 General Assignment

Read problem2.txt and concatenate all data into one line separated by space.

```
> Expression <- scan("/Users/fanheng/Dropbox/Columbia 2014Spring/W4249-Applied Data Science/Homework
> Expression <- paste(Expression, collapse = " ")
> exp.pattern<-"[ ]([-]?[0-9]+[ ]*[-+]{1}[ ]*[0-9]+)"
> exp.pattern.match <- str_match_all(Expression, exp.pattern)
> exp.ls <- unlist(exp.pattern.match)
> exp.ls <- parse(text=exp.ls)
> loc.exp <- matrix(unlist(str_locate_all(Expression, exp.pattern)),ncol=2)
> for(i in 1:nrow(loc.exp)){
+   str_sub(Expression, loc.exp[i,1], loc.exp[i,2]) <- str_pad(eval(exp.ls[i]),
+                                    + loc.exp[i,2]-loc.exp[i,1]+1,side="both")
+  }
```

The expression after replacing the a+/-b patten is
> Expression
[1] ”-57.04–57.04 -10.81+2.02 6.79+7.45 8.71-5.79 48.11-7.74 33.77-4.92 -19.51–19.51 49.21+2.87 15.87+9.40 -40.16+8.21 -1.25–1.25 50.79-3.06 94.72+2.54 -10 39 92 94 20 -67 16 -62 84 -5 64 87 35 3 40 81 26 36 69 36 66 0 43 -6 -43 -47.33–47.33 7.73+7.54 -91.41+3.18 67.08-1.81 -1.38–1.38 88.82-4.57 45.98+9.97 -53.95–53.95 -84.30–84.30 23.42-5.04 -54.72–54.72 -34.27–34.27 20 -31 12 -17 -59 -31 -90 -87 16 42 -21 12 31 69 -86 -58 71 88 -3 59 -57 90 77 39 5 82.34+2.87 -68.82–68.82 -88.12+4.18 -70.54–70.54 27.71-4.07 -54.35+2.64 -84.10+4.81 -89.68+6.36 -83.54+3.62 19.68+3.70 -68.48+2.46 -38.25–38.25 -71.01+0.11”

2.2 General Assignment

Sums that involve fractional decimals, e.g. âĂIJ2.57+3âĂİ may be left unchanged.

```
> Expression <- scan("/Users/fanheng/Dropbox/Columbia 2014Spring/W4249-Applied Data Science/Homework
> Expression <- paste(Expression, collapse = " ")
> exp.pattern<-"[ ]([-]?[0-9]*[.]?[0-9]+[ ]*[-+]{1}[ ]*[0-9]*[.]?[0-9]+)"
```

```
> exp.pattern.match <- str_match_all(Expression, exp.pattern)
> exp.ls <- unlist(exp.pattern.match)
> exp.ls <- parse(text=exp.ls)
> loc.exp <- matrix(unlist(str_locate_all(Expression, exp.pattern)),ncol=2)
> for(i in 1:nrow(loc.exp)){
+   str_sub(Expression, loc.exp[i,1], loc.exp[i,2]) <- str_pad(eval(exp.ls[i]),
+                                    + loc.exp[i,2]-loc.exp[i,1]+1,side="both")
+  }
```

The expression after replacing the a+/-b patten is

> Expression

[1] "-57.04–57.04 -8.79 14.24 2.92 40.37 28.85 -19.51–19.51 52.08 25.27 -31.95 -1.25–1.25 47.73 97.26 -10 39 92 94 20 -67 16 -62 84 -5 64 87 35 3 40 81 26 36 69 36 66 0 43 -6 -43 -47.33–47.33 15.27 -88.23 65.27 -1.38–1.38 84.25 55.95 -53.95–53.95 -84.30–84.30 18.38 -54.72–54.72 -34.27–34.27 20 -31 12 -17 -59 -31 -90 -87 16 42 -21 12 31 69 -86 -58 71 88 -3 59 -57 90 77 39 5 85.21 -68.82–68.82 -83.94 -70.54–70.54 23.64 -51.71 -79.29 -83.32 -79.92 23.38 -66.02 -38.25–38.25 -70.9 "