Applied Data Science

Xiaoyao Yang
Columbia University

April 10, 2014

1 Finding phone type

```
require(stringr)
```

```
Loading required package: stringr
```

```
# Problem 1 find type
strings ← paste(readLines("problem1.txt"), sep = "", collapse = "\n")
strings ← str_replace_all(string = strings, pattern = "\n", replacement = "")
typep ← "[a-zA-Z]{4}[:punct:][ ]*[1]?[-]?([2-9][0-9]{2})[- .]([0-9]{3})[- .]([0
    -9]{4})"
type ← unlist(str_extract_all(string = strings, pattern = typep))
type ← gsub(pattern = " ", replacement = "", x = type)
type
```

```
 [1] "work:1-266-113-8009" "home:1-465-860-7545" "home:1-707-585-6847"
 [4] "home:1-890-281-7216" "work:1-292-467-4748" "work:1-469-409-0758"
 [7] "work:1-947-564-6985" "work:1-550-914-3267" "home:1-407-441-2266"
[10] "work:1-554-992-6974" "work:1-755-293-8874" "work:1-830-262-2372"
[13] "home:1-833-789-8018" "work:1-377-425-1766" "work:1-270-793-9751"
[16] "home:1-760-711-1858" "home:1-501-370-6447" "home:1-942-466-9544"
[19] "home:1-356-392-8148" "home:1-490-289-5762" "home:1-234-572-1998"
[22] "home:1-448-693-3867" "cell:1-541-479-6934" "work:1-683-871-4920"
[25] "work:1-237-492-8727" "home:1-216-898-8392" "work:1-210-375-7032"
[28] "home:1-591-218-5103" "work:1-742-797-8892" "work:1-237-738-7807"
[31] "work:1-485-575-4135" "work:1-221-769-6071" "home:1-926-815-0922"
[34] "home:1-535-972-9143" "work:1-456-702-1201" "home:1-653-150-4005"
[37] "cell:1-497-935-5888" "work:1-658-362-9597" "work:1-296-703-3683"
[40] "home:1-954-392-0604" "work:1-457-600-6614"
```

```
length(type)
```

```
[1] 41
```

```
type_work ← str_detect(type, "work")
type[type_work]
```

```
 [1] "work:1-266-113-8009" "work:1-292-467-4748" "work:1-469-409-0758"
 [4] "work:1-947-564-6985" "work:1-550-914-3267" "work:1-554-992-6974"
 [7] "work:1-755-293-8874" "work:1-830-262-2372" "work:1-377-425-1766"
[10] "work:1-270-793-9751" "work:1-683-871-4920" "work:1-237-492-8727"
[13] "work:1-210-375-7032" "work:1-742-797-8892" "work:1-237-738-7807"
[16] "work:1-485-575-4135" "work:1-221-769-6071" "work:1-456-702-1201"
[19] "work:1-658-362-9597" "work:1-296-703-3683" "work:1-457-600-6614"
```

```
sum(type_work)
```

```
[1] 21
```

```
type_home ← str_detect(type, "home")
type[type_home]
```

```
 [1] "home:1-465-860-7545" "home:1-707-585-6847" "home:1-890-281-7216"
 [4] "home:1-407-441-2266" "home:1-833-789-8018" "home:1-760-711-1858"
 [7] "home:1-501-370-6447" "home:1-942-466-9544" "home:1-356-392-8148"
[10] "home:1-490-289-5762" "home:1-234-572-1998" "home:1-448-693-3867"
[13] "home:1-216-898-8392" "home:1-591-218-5103" "home:1-926-815-0922"
[16] "home:1-535-972-9143" "home:1-653-150-4005" "home:1-954-392-0604"
```

```
sum(type_home)
```

```
[1] 18
```

```
type_cell ← str_detect(type, "cell")
type[type_cell]
```

```
[1] "cell:1-541-479-6934" "cell:1-497-935-5888"
```

```
sum(type_cell)
```

```
[1] 2
```

2 Extra credit:area code in parenthesis

```
# find area code!!!!!!
areap ← "[\\(]([2-9][0-9]{2})[\\)][ ]*([0-9]{3})[- .]([0-9]{4})"
str_extract_all(string = strings, pattern = areap)
```

```
[[1]]
 [1] "(314) 483-2576" "(821) 474-7064" "(668) 831-0991" "(680) 849-8531"
 [5] "(382) 781-9603" "(930) 375-2196" "(825) 144-2637" "(850) 656-9038"
 [9] "(245) 630-2263" "(564) 477-4993" "(622) 825-3614" "(806) 860-0676"
[13] "(785) 632-5114" "(288) 599-0104" "(808) 169-0296" "(348) 955-0915"
[17] "(849) 583-3586" "(518) 114-5941" "(204) 145-6498" "(764) 381-8888"
[21] "(240) 472-9213" "(772) 921-9459" "(705) 309-9278" "(830) 980-5045"
[25] "(615) 722-7276" "(349) 499-4061" "(776) 549-1154" "(869) 921-3212"
[29] "(948) 122-4463" "(712) 102-2609" "(713) 857-5408" "(948) 767-1338"
[33] "(765) 960-6186" "(268) 362-0185" "(491) 387-9089" "(706) 795-0072"
[37] "(844) 857-6213" "(231) 977-0861" "(549) 832-8823" "(758) 873-5157"
[41] "(742) 755-0657" "(788) 847-6234" "(209) 640-6256" "(975) 715-3150"
[45] "(724) 228-1778" "(417) 663-3639" "(447) 208-5898" "(751) 678-4231"
```

3 Calculate text(also used for extra credit)

```
dat2 ← scan("problem2.txt", character(0))
# only used for single text input 'a+b' or '+-a-b' a, b can be any number
Evaluated ← function(textformula) {
    if (!require(stringr))
        require(stringr)
    # extract number
    test ← textformula
    temp ← unlist(str_split(string = test, pattern = ”\\+|\\−”))
    num ← as.numeric(na.omit(as.numeric(temp)))
    if (length(num) != 2)
        stop(”input error (must be plus or minus)”)
    # determine sign
    temp.sign ← unlist(str_extract_all(string = test, pattern = ”\\+|\\−|\\−−
       |\\++”))
    if (sum(str_detect(temp.sign, ”[−][−]|[+][−]|[−][+]|[+][+]”))) {
        return(test)
    } else {
        for (i in 1:length(temp.sign)) {
            if (temp.sign[i] == ”−−”)
                temp.sign[i] ← ”+”

        }

        if (length(temp.sign) == 1) {
            num[2] ← as.numeric(str_join(temp.sign, num[2]))
            result ← sum(num)
        } else if (length(temp.sign) == 2) {
            num.new ← as.numeric(str_join(temp.sign, num))
            result ← sum(num.new)
        } else {
            stop(”input error”)
        }
        return(result)
    }

}

ans ← vector()
for (j in 1:length(dat2)) {
    ans[j] ← Evaluated(dat2[j])
}
ans
```

```
 [1] "-57.04--57.04" "-8.79"         "14.24"         "2.92"
 [5] "40.37"         "28.85"         "-19.51--19.51" "52.08"
 [9] "25.27"         "-31.95"        "-1.25--1.25"   "47.73"
[13] "97.26"         "-10"           "39"            "92"
[17] "94"            "20"            "-67"           "16"
[21] "-62"           "84"            "-5"            "64"
[25] "87"            "35"            "3"             "40"
[29] "81"            "26"            "36"            "69"
[33] "36"            "66"            "0"             "43"
[37] "-6"            "-43"           "-47.33--47.33" "15.27"
[41] "-88.23"        "65.27"         "-1.38--1.38"   "84.25"
[45] "55.95"         "-53.95--53.95" "-84.30--84.30" "18.38"
[49] "-54.72--54.72" "-34.27--34.27" "20"            "-31"
[53] "12"            "-17"           "-59"           "-31"
[57] "-90"           "-87"           "16"            "42"
[61] "-21"           "12"            "31"            "69"
[65] "-86"           "-58"           "71"            "88"
[69] "-3"            "59"            "-57"           "90"
[73] "77"            "39"            "5"             "85.21"
[77] "-68.82--68.82" "-83.94"        "-70.54--70.54" "23.64"
[81] "-51.71"        "-79.29"        "-83.32"        "-79.92"
[85] "23.38"         "-66.02"        "-38.25--38.25" "-70.9"
```