Lab 3

Ankita Shankhdhar

Due 2 October 2015

Loading the html file

```
richhtml <- readLines("http://people.math.umass.edu/~jstauden/rich.html/rich.html")
```

Problem 1

The html contains:

```
cat("Lines:", length(richhtml)) #lines
```

```
## Lines: 1991
```

```
cat("Charaters",sum(nchar(richhtml))) # chanracters
```

```
## Charaters 80375
```

Problem 2

- ***Bill Gates***

```
bill <- grep("[gG]ates",richhtml)
name.bill <- richhtml[grep("[Bb]ill Gates",richhtml)]
worth.bill<-richhtml[grep("Bill Gates",richhtml)+1]
cat("The text from the lines that has his name:",name.bill)
```

```
## The text from the lines that has his name:           <h3>Bill Gates</h3></a></td>
```

```
cat("The text from the lines that has his net worth:",worth.bill)
```

```
## The text from the lines that has his net worth:      <td class="worth">$72 B</td>
```

- ***Stanley Kroenke***

```
stanley <- grep("[Kk]roenke",richhtml)
name.stanley <- richhtml[grep("[Ss]tanley Kroenke",richhtml)]
worth.stanley<-richhtml[grep("Stanley Kroenke",richhtml)+1]
cat("The text from the lines that has his name:",name.stanley)
```

```
## The text from the lines that has his name:           <h3>Stanley Kroenke</h3></a></td>
```

```
cat("The text from the lines that has his net worth:",worth.stanley)
```

```
## The text from the lines that has his net worth:      <td class="worth">$5,3 B</td>
```

Problem 3

To get the list of names

```
names.all <- richhtml[grep("\\$[0-9,]+\\sB",richhtml)-1]
names.all<-unlist(strsplit(names.all,split="<h3>"))[seq(from=2,to=200,by=2)]
names.all<-gsub("</h3></a></td>","",names.all)
```

To get the net worth

```
# Get all the lines of the net worths
worth.all <- richhtml[grep("\\$[0-9,]+\\sB",richhtml)]
# only need every 2 line to keep those
worth.all <- unlist(strsplit(worth.all,split="\\$"))[seq(from=2,to=200,by=2)]
# get rid of everything before and after it
worth.all <- gsub(" B</td>","", worth.all)
worth.all <- gsub(">","", worth.all)
# set it as a numeric
worth.all <- as.numeric(gsub(",",".", worth.all))
```

The ***stringAsFactors=False*** tells the dataframe not to convert string to factors. The dataframe then keeps the string vector as a character.

```
names.worth <- data.frame(names.all,worth.all,stringsAsFactors=FALSE)
sapply(names.worth, class)
```

```
##   names.all   worth.all
## "character"   "numeric"
```

```
names(names.worth)<-c("Name","Net Worth (Billions $)")
names.worth
```

```
##                          Name Net Worth (Billions $)
## 1                   Bill Gates                  72.0
## 2                Warren Buffett                 58.5
## 3                 Larry Ellison                 41.0
## 4                  Charles Koch                 36.0
## 5                    David Koch                 36.0
## 6       Christy Walton & family                 35.4
## 7                     Jim Walton                 33.8
## 8                   Alice Walton                 33.5
## 9              S. Robson Walton                  33.3
## 10            Michael Bloomberg                 31.0
## 11              Sheldon Adelson                 28.5
## 12                    Jeff Bezos                 27.2
```

```
## 13                   Larry Page                24.9
## 14                  Sergey Brin                24.4
## 15             Forrest Mars, Jr.                20.5
## 16               Jacqueline Mars                20.5
## 17                     John Mars                20.5
## 18                   Carl Icahn                20.3
## 19                  George Soros                20.0
## 20                Mark Zuckerberg               19.0
## 21                  Steve Ballmer               18.0
## 22                  Len Blavatnik               17.8
## 23                Abigail Johnson               17.2
## 24                   Phil Knight                16.3
## 25                  Michael Dell                15.9
## 26                    Paul Allen                15.8
## 27                  Donald Bren                 14.0
## 28                Ronald Perelman               14.0
## 29              Anne Cox Chambers               13.5
## 30       Rupert Murdoch & family                13.4
## 31                    Ray Dalio                 12.9
## 32                  Charles Ergen               12.5
## 33                  Harold Hamm                 12.4
## 34                  James Simons                12.0
## 35  Laurene Powell Jobs & family                11.7
## 36                  John Paulson                11.4
## 37          Jack Taylor & family                11.4
## 38               Philip Anschutz                10.3
## 39                Richard Kinder                10.2
## 40                 George Kaiser                10.0
## 41                 Harold Simmons               10.0
## 42                    Andrew Beal                9.8
## 43                    Steve Cohen                9.4
## 44          Edward Johnson, III.                 9.3
## 45           Patrick Soon-Shiong                 9.0
## 46          Samuel Newhouse, Jr.                 8.9
## 47         Charles Butt & family                 8.5
## 48                Pierre Omidyar                 8.5
## 49      Elaine Marshall & family                 8.3
## 50            Hank & Doug Meijer                 8.3
## 51                  Eric Schmidt                 8.3
## 52                Donald Newhouse                8.2
## 53                   David Tepper                7.9
## 54                   Ralph Lauren                7.7
## 55             Stephen Schwarzman                7.7
## 56                 Leonard Lauder                7.6
## 57               John Menard, Jr.                7.5
## 58                James Goodnight                7.2
## 59                      Eli Broad                6.9
## 60                  Richard DeVos                6.8
## 61                    Jim Kennedy                6.7
## 62                    John Malone                6.7
## 63                     Elon Musk                 6.7
## 64             Blair Parry-Okeden                6.7
## 65                 David Duffield                6.4
## 66  Herbert Kohler, Jr. & family                 6.4
```

```
## 67              Thomas Peterffy                6.4
## 68              S. Truett Cathy                6.0
## 69                 David Geffen                6.0
## 70                 Micky Arison                5.9
## 71              Sumner Redstone                5.8
## 72            Dennis Washington                5.8
## 73                Leslie Wexner                5.7
## 74                 Ray Lee Hunt                5.6
## 75              Charles Johnson                5.6
## 76      Richard LeFrak & family                5.6
## 77                Dannine Avara                5.5
## 78                 Scott Duncan                5.5
## 79                Milane Frantz                5.5
## 80           Jeffery Hildebrand                5.5
## 81          Rupert Johnson, Jr.                5.5
## 82                  Ira Rennert                5.5
## 83               Randa Williams                5.5
## 84              Stanley Kroenke                5.3
## 85                   Leon Black                5.2
## 86                   Gayle Cook                5.2
## 87             Dustin Moskovitz                5.2
## 88             Patrick McGovern                5.1
## 89               Charles Schwab                5.1
## 90       Jin Sook & Do Won Chang                5.0
## 91   Thomas Frist, Jr. & family                5.0
## 92                 David Green                5.0
## 93              Robert Rowling                4.9
## 94                Stephen Ross                4.8
## 95                Bruce Kovner                4.7
## 96                Henry Kravis                4.7
## 97           Ann Walton Kroenke                4.7
## 98                 Gordon Moore                4.6
## 99                  Daniel Ziff                4.6
## 100                   Dirk Ziff                4.6
```