Лабораторная работа №3

# Задачи работы:

* работа с текстом;
* использование регулярных выражений для извлечения данных;
* переписывание существующего кода;
* ассиметричные распределения.

# Исходные данные

Файл *forbes.htm* содержит список богатейших американцев по версии журнала Форбc.

# Задание 1

Используйте команду readLines для загрузки файла в текстовый вектор html

forbes <- readLines("https://raw.githubusercontent.com/SergeyMirvoda/MD-DA-2017/master/data/forbes.htm")

Сколько строк в файле?

length(forbes)

## [1] 1991

Сколько символов в файле?

sum(nchar(forbes))

## [1] 80380

Напишите шаблон регулярного выражения и используйте функцию grep, чтобы извлечь размер дохода из данных в векторе html. Удостоверьтесь, что полученный вектор номеров строк содержит содержит ровно 100 записей

patterns.profit <- "[$][,0-9]+ [BM]"  
forbes.isprofit <- grep(patterns.profit, forbes)  
  
length(forbes.isprofit) == 100

## [1] TRUE

Напишите код, использовав регулярное выражение из п. 3, и функции regexp и regmatches, чтобы извлечь все данные о доходе

forbes.profits <- regmatches(forbes, regexpr(patterns.profit, forbes))

Должно быть ровно сто значений

length(forbes.profits) == 100

## [1] TRUE

Самый большой доход должен быть доход Билла Гейтса

Такой доход должен быть в списке один раз.

В всписке должна быть цифра, которую мы запомнили для Ларри Эллисона.

Должно быть как минимум два значения, встречающихся несколько раз.

forbes.profits.stats <- table(forbes.profits)  
as.data.frame(forbes.profits.stats)

## forbes.profits Freq  
## 1 $10 B 2  
## 2 $10,2 B 1  
## 3 $10,3 B 1  
## 4 $11,4 B 2  
## 5 $11,7 B 1  
## 6 $12 B 1  
## 7 $12,4 B 1  
## 8 $12,5 B 1  
## 9 $12,9 B 1  
## 10 $13,4 B 1  
## 11 $13,5 B 1  
## 12 $14 B 2  
## 13 $15,8 B 1  
## 14 $15,9 B 1  
## 15 $16,3 B 1  
## 16 $17,2 B 1  
## 17 $17,8 B 1  
## 18 $18 B 1  
## 19 $19 B 1  
## 20 $20 B 1  
## 21 $20,3 B 1  
## 22 $20,5 B 3  
## 23 $24,4 B 1  
## 24 $24,9 B 1  
## 25 $27,2 B 1  
## 26 $28,5 B 1  
## 27 $31 B 1  
## 28 $33,3 B 1  
## 29 $33,5 B 1  
## 30 $33,8 B 1  
## 31 $35,4 B 1  
## 32 $36 B 2  
## 33 $4,6 B 3  
## 34 $4,7 B 3  
## 35 $4,8 B 1  
## 36 $4,9 B 1  
## 37 $41 B 1  
## 38 $5 B 3  
## 39 $5,1 B 2  
## 40 $5,2 B 3  
## 41 $5,3 B 1  
## 42 $5,5 B 7  
## 43 $5,6 B 3  
## 44 $5,7 B 1  
## 45 $5,8 B 2  
## 46 $5,9 B 1  
## 47 $58,5 B 1  
## 48 $6 B 2  
## 49 $6,4 B 3  
## 50 $6,7 B 4  
## 51 $6,8 B 1  
## 52 $6,9 B 1  
## 53 $7,2 B 1  
## 54 $7,5 B 1  
## 55 $7,6 B 1  
## 56 $7,7 B 2  
## 57 $7,9 B 1  
## 58 $72 B 1  
## 59 $8,2 B 1  
## 60 $8,3 B 3  
## 61 $8,5 B 2  
## 62 $8,9 B 1  
## 63 $9 B 1  
## 64 $9,3 B 1  
## 65 $9,4 B 1  
## 66 $9,8 B 1

# Задание 2

В данных доход представлен в формате "$42 B", что означает 42 × 10^9. Преобразуйте этот формат в числовой и сохраните в вектор worths.

M <- 10 \*\* 9  
forbes.profits.locale <- gsub(",", "\\.", forbes.profits)  
worths <- as.double(regmatches(forbes.profits.locale, regexpr("[0-9.]+", forbes.profits.locale))) \* M

worths является вектором и в нём сто занчений типа double.

Все элементы вектора worths больше 1 миллиарда.

Самое большое число это доход Билла Гейтса.

length(worths) == 100

## [1] TRUE

typeof(worths) == "double"

## [1] TRUE

all(worths > 1 \* M)

## [1] TRUE

as.data.frame(head(worths))

## head(worths)  
## 1 7.20e+10  
## 2 5.85e+10  
## 3 4.10e+10  
## 4 3.60e+10  
## 5 3.60e+10  
## 6 3.54e+10

Какова медиана ста этих записей?

median(worths)

## [1] 8.3e+09

Средний доход?

mean(worths)

## [1] 1.293e+10

Как много людей из этого списка имеют доход больше 5млрд., 10, 25?

length(worths[which(worths > 5 \* M)])

## [1] 89

length(worths[which(worths > 10 \* M)])

## [1] 39

length(worths[which(worths > 25 \* M)])

## [1] 12

Какой их общий доход?

sum(worths)

## [1] 1.293e+12

Какую долю от общего дохода, составляет пятёрка самых богатых.

sum(worths[1:5]) \* 100 / sum(worths)

## [1] 18.83217

Какую долю от общего дохода, составляют 20 самых богатых.

sum(worths[1:20]) \* 100 / sum(worths)

## [1] 49.21114

В данных федерального резерва США найдите показатель дохода всех домохозяйств (Household net worth) в соответвующем году, какую долю общего дохода составляют 100 богатейших людей.

Household\_net\_worth <- 96.196 \* M  
Household\_net\_worth \* 100 / sum(worths)

## [1] 7.439753

# Заключение

В ходе выполнения данной работы я получил незабываемый опыт и приобрел навыки, которые пригодятся в дальнейшей профессиональной деятельности. Были изучены и освоены на практике основы манипулирования строковыми данными и регулярные выражения.

С уважением,

студент гр. РИ-440005

Кабанов Евгений