LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

25 January 2018

Class outline

- Some new R functions
- Reading in multiple files from a folder
- Storing programs in .R files
- Group exercise
- Reminder: Submit Assignment 1 on time.

Some new R functions

Adding more items to an existing vector

```
days <- c("Monday", "Tuesday")
days <- c("Sunday", days)
days <- c(days,"Wednesday")
days
```

setdiff function

Returns difference between two vectors

```
v1 <- c("This", "is", "a", "small", "example")
v2 <- c("is", "a")
v3 <- setdiff(v1,v2)
```

v3 here will have "This", "small", "example"

Writing "loops" to repeat tasks

- Why?: Tasks that need to be performed repetitively
- E.g., I want to count 10 most frequent words not for one file, but for all files in a folder.

Writing "loops" to repeat tasks

- Why?: Tasks that need to be performed repetitively
- E.g., I want to count 10 most frequent words not for one file, but for all files in a folder.
- R syntax for writing such loops - an example:

```
new <- vector()
for(day in days)
{
 new <- c(new,nchar(day))
}
new
```

Reading a directory of text files - example

Iteratively accessing files in a folder-1

```
file.names <- dir("/home/bangaru/Desktop")
#this path is on my computer!!
for (f in file.names) {
print(f)
}
```

Iteratively accessing files in a folder-2

```
file.names <- dir(getwd(), pattern ="\\.txt")
for (f in file.names) {
print(f)
}
```

Reading the content of all files

```
setwd("my path to the required folder")
file.names <- dir(getwd(), pattern ="\\.txt")
file.names
data <- vector()
for (f in file.names) {
  tempData <- scan(f, what="character", sep="\n")
  tempDataString <- tolower(paste(tempData, collapse = " "))
  data <- c(data,tempDataString)
}
#What will "data" contain now?
length(data)
data[1]
nchar(data[1])
nchar(data)
```

continuing from there...

Let us see if we can get the 10 most frequent words per book:

```
for(item in data)
{
  words <- strsplit(item, "\\W+")
  sorted_freqs <- sort(table(words), decreasing = TRUE)
  print(sorted_freqs[1:10])
}
```

continuing from there...

Let us put everything in one loop:

```
setwd("my path to the required folder")
file.names <- dir(getwd(), pattern ="\\.txt")
for (f in file.names) {
  cat("in this file:",f, "\n")
  tempData <- scan(f, what="character", sep="\n")
  tempDataString <- tolower(paste(tempData, collapse = " "))
   words <- strsplit(tempDataString, "\\W+")
  sorted_freqs <- sort(table(words), decreasing = TRUE)
  print(sorted_freqs[1:10])
}
```

-here, I did not do the part of reading all data into one larger variable - why?
(I uploaded a file called Directory.R which contains these lines of R code)

Note: There are other ways to do this in R.

storing .R files

- Why?: Reusability - it wont go away when you close Rstudio.
- Please note: On lab computers, it will be erased - so upload to box, google drive or some such storage before you leave.
- Using a .R file:
  - You can open in the left top panel.
  - Two options: source, run. Run runs the code on the console, line by line.
  - Source: file is stored and run internally and you just see the output.

storing .R files

- Why?: Reusability - it wont go away when you close Rstudio.
- Please note: On lab computers, it will be erased - so upload to box, google drive or some such storage before you leave.
- Using a .R file:
  - You can open in the left top panel.
  - Two options: source, run. Run runs the code on the console, line by line.
  - Source: file is stored and run internally and you just see the output.
  - If you write your own R functions in your program, source("filename.R") will also allow other R functions you type in the console to use those from this file.

Exercise

- Why are we lower casing, and what happens if we just don't remove punctuations, and just split by white space?
- What sort of frequency tables will you see?
- How do these pre-processing decisions affect our results and analysis we do of them?
- You can work in groups and discuss among yourselves, and post in the discussion forum any code (attach .R files) and comments on the outputs you see.

Next Week

- Continuation and Conclusion of working with different formats (e.g., reading from NYTimes directly from R)
- Recommended reading: `https://cran.r-project.org/web/packages/rtimes/rtimes.pdf`
- Check the first 5-6 pages in this pdf, if possible, try to make it work in R Studio as shown in their example.
- Reminder: Submit Assignment 1

Next Week

- Continuation and Conclusion of working with different formats (e.g., reading from NYTimes directly from R)
- Recommended reading: `https://cran.r-project.org/web/packages/rtimes/rtimes.pdf`
- Check the first 5-6 pages in this pdf, if possible, try to make it work in R Studio as shown in their example.
- Reminder: Submit Assignment 1
- Question: How many of you have Twitter accounts? How many are willing to create one for the class?
- If you are not interested, we won't do it as an exercise - I will just provide a tutorial file for enthusiasts.