LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

30 January 2018

Class Outline

1. Quick recap of last week
2. Assignment 1 Discussion
3. Some new functions
4. R library to work with NYT data
5. Assigment 2 description
6. Quick note on using twitter data

Recap of last week

Topics

1. Working with file formats (.txt., .docx, .pdf, .html, .xml)
2. Reading all files (or files that match a pattern such as "ending in .txt") in a folder
3. Storing .R files
4. New R stuff we learnt:
   - libraries: pdftools, qdapTools, XML
   - functions: setdiff(), dir()
   - others: Writing a for loop, adding new items to an existing vector

Solution to last class' exercise

- What happens if we remove lower casing, don't remove punctuations, and just split by whitespace, and then look for 10 most frequent words?
- How do we remove lower casing? - remove tolower() function call.
- How do we split on whitespace? - instead of splitting with "\\W+", just split using " " (space).

Solution to last class' exercise

- What happens if we remove lower casing, don't remove punctuations, and just split by whitespace, and then look for 10 most frequent words?
- How do we remove lower casing? - remove tolower() function call.
- How do we split on whitespace? - instead of splitting with "\\W+", just split using " " (space).
- What happens as a result?

Solution to last class' exercise

- What happens if we remove lower casing, don't remove punctuations, and just split by whitespace, and then look for 10 most frequent words?
- How do we remove lower casing? - remove tolower() function call.
- How do we split on whitespace? - instead of splitting with ”\\W+”, just split using ” ” (space).
- What happens as a result?
- Punctuations remain. Case differences remain. So - being; Being; being,; being. - all will be considered different. Your word frequency list changes accordingly!

Assignment 1 discussion

Question 1: Journalism and Mass communication; Business; Sociology

- Working with large amounts of public information, and performing content analysis efficiently
- Analyzing social media to identify trends
- Identify consumer reactions on social media (by companies)
- Choosing the right message strategy to reach consumers (what kind of messages about products, how to send? etc)
- Crisis management: spreading information about disasters etc
- Negative vs positive news identification

Question 1: Linguistics

- Approximate translation in a large scale
- Identifying function vs content words in text, how many words for each category etc
- study language development
- create vocabulary lists
- automatic writing evaluation
- study language variation
- in tools such as alexa, siri etc

Question 1: Literature

- Analyze overuse or underuse of words by authors
- identifying patterns of speech used in greeting people
- quick editing/proofreading of documents
- how many times a character is mentioned

Question 2: Solutions

```
> cyclones <- "that string I gave in quotes"
> nchar(cyclones)
> cyclones_upper <- toupper(cyclones)
> gsub("(\\d{2}\\.\\d{1})","NUM", cyclones)
(or, to be elaborate: )
> gsub("(\\D\\d{2}\\.\\d{1}\\D)","NUM", cyclones)
(but, this second expression also substitutes parantheses before numbers)
> strsplit(cyclones, ".", fixed = TRUE) or strsplit(cyclones, "\\.")
(what does fixed=TRUE do?)
> str_to_title(cyclones)
```

General Comments on Assignment 1

- Submit in the format I asked for in the Assignment description.
- Read any supporting materials provided carefully - I won't ask you to do anything that you will not be able to do at that point in course work.
- I ask for Zip files, so that I can download one file for person (programs cannot be evaluated on canvas in browser!

Some new stuff about R vectors, and lists

Vectors and Lists

- Vectors: a collection of objects of same kind (numbers, characters, logical values etc)

  ```
  > vector1 <- c(1,2,3,4)
  > vector2 <- c("English", "German", "French", "Italian", "Chinese")
  > vector3 <- c(TRUE, FALSE, FALSE, TRUE)
  ```

  ... and so on
- Lists: collection of objects of different kind.

  ```
  > list1 <- list(1,"a",TRUE,4)
  (This list has two numbers, one string and a boolean value)
  > list2 <- list(1,"a",c(1,2,3),4)
  > list3 <- list(1,"a",list(1,2,3),4)
  ```

More examples of vectors and lists

We can also have named lists and vectors like this:

```
list4 <- list(first="Sowmya", course=410, office=331, address="Ross")
vector4 <- c(first="Sowmya", course=410, office=331, address="Ross")
(R coerces numbers into strings in above vector1)
names(list4); names(vector4) gives you -
name, course, office, address
```

accessing individual elements of vectors and lists

- if I have vector4 $<-$ c(1,4,7,15), vector4[1] gives me 1, vector4[2] gives me 4 and so on.

accessing individual elements of vectors and lists

- if I have vector4 $< -$ c(1,4,7,15), vector4[1] gives me 1, vector4[2] gives me 4 and so on.
- The way you access elements of a list is slightly different from this in R. It is just the syntax of that language - nothing very logical about it.
- Let us take the list from previous slide:
- list4 $< -$ list(name="Sowmya", courseNum=410, office=331, address="Ross")
- To access the first element in this, I use [[]] instead of [].
- list4[["name"]] or list4[[1]] will give me "Sowmya".

accessing individual elements of vectors and lists

- if I have vector4 $< -$ c(1,4,7,15), vector4[1] gives me 1, vector4[2] gives me 4 and so on.
- The way you access elements of a list is slightly different from this in R. It is just the syntax of that language - nothing very logical about it.
- Let us take the list from previous slide:
- list4 $< -$ list(name="Sowmya", courseNum=410, office=331, address="Ross")
- To access the first element in this, I use [[]] instead of [].
- list4[["name"]] or list4[[1]] will give me "Sowmya".
- list4[1] will give me:
  $name
  [1]"*Sowmya*"

How do we know whether something is a list or vector

Apart from visual inspection, is.vector(some_variable), is.list(some_variable) are two functions we can use to find out whether something is a vector or a list.

Two more

- write.csv(some_variable, "filename.csv") - creates a comma separated value file (which can be read as a spreadsheet)
- data.frame(col1,col2) - takes two vectors col1, col2 (equal length) and puts them into a table like format, as two columns (we can put any number of columns we want)

working with R libraries: Example with NYT

R libraries for specific data collections

- There are custom R libraries for specific data collections (such as NYT, Guardian, Gutenberg, Wikipedia etc)

R libraries for specific data collections

- There are custom R libraries for specific data collections (such as NYT, Guardian, Gutenberg, Wikipedia etc)
- We can always access those websites as if they are any other website, use XML library and work with HTML format.
- However, these libraries make our job easier by providing some custom functions to access data from these websites.

R libraries for specific data collections

- There are custom R libraries for specific data collections (such as NYT, Guardian, Gutenberg, Wikipedia etc)
- We can always access those websites as if they are any other website, use XML library and work with HTML format.
- However, these libraries make our job easier by providing some custom functions to access data from these websites.
- I am taking NYT as an example. Your Assignment 2 will require you to use Guardian library.
- We cannot exhaustively do for all websites in internet world.

Analyzing NYT data - example

needs: rtimes package

needs: NY Times ”key”

`http://developer.nytimes.com/apps/register`

Example Usage

```
library(rtimes)
Sys.setenv(NYTIMES_AS_KEY = "THE KEY YOU GET AFTER REGISTERING ")
res1 <- as_search(q="artificial intelligence", begin_date = "20081001", end_date = "20081201")
res2 <- as_search(q="artificial intelligence", begin_date = "20180101", end_date = "20180120")
res3 <- as_search(q = "money", fq = 'news_desk:("Sports" "Foreign")') #search within categories
res4 <- as_search("iowa caucus")
names(res1)
```

References:

- https://cran.r-project.org/web/packages/rtimes/vignettes/rtimes_vignette.html
- https://cran.r-project.org/web/packages/rtimes/rtimes.pdf
- I am following their guidelines for date formats, query format etc.

How does the output look like?

- seems like a big list.
- res1$data$snippet - gives me snippets for retrieved news items from 2008.
- res2$data$snippet - gives me snippets for retrieved news items from 2008.
- These seem to be vectors. is_vector(res1$data$snippet) gives TRUE.
- We can do other stuff we did before with this. Example:

  ```
  > snippets_2008 <-  res1$data$snippet
  > for (snippet in snippets_2008) {
     print(tolower(snippet))
  }
  We can also write specific columns into a new file
  > df <- data.frame(res1$data$snippet,res1$data$pub_date)
  > write.csv(df,"temp.csv")
  ```

- We can do analyses such as: what are people talking about in 2008 vs 2018 on AI etc.

Things to keep in mind when working with such libraries

- Always check the documentation for how to use different functions, what values they return to you etc.
- Some libraries change formats between versions: so the same code may not work 5 years later, if your library is updated
- It will work ofcourse, if you did not update your R version, R libraries etc.
- Example: I talked about the same NYT library last year too, but results (same information) was shown in a different format (not as a list).

Assignment 2 Description

Assignment 2

1. 2 Questions, 10% of your grade in total (5% for each question)
2. Deadline: 10th February 2018
3. First question: Very easy, but you should learn to use something I did not discuss in class (use ?readLines and figure out!)
4. Second question: Use GuardianR library (not NYTimes) and answer few questions. You should look at the GuardianR package documentation on R website and understand how to use it.

working with Twitter: Quick introduction

(Note: I will not do this in the class, as not everyone wants a twitter account. But I strongly encourage you to learn to scrape data from twitter atleast during your course projects. I can do an additional tutorial session for those who are interested, perhaps in the week after spring break.)

why care about twitter?

- Twitter (and other such social media) is widely used these days.
- Millions of people tweet every day.
- This includes government agencies and people who run the country.
- This means social media is a useful source to analyze current trends and thoughts
- Tweets are textual data too! lot of it!

What can we study on Twitter

- how information spreads across geographical locations
- how are people reacting to the release of the new iphone version?
- what is white house communicating with its citizens and foreigners?
- What are the political views of a person?

Twitter in R

- twitteR and streamR libraries are commonly used.
- twitteR is more about doing search for keywords, hashtags, users, followers.
- streamR will also do location based sorting of tweets, you can access tweets in real time (as they get tweeted, almost) etc.
- There are also such APIs for facebook, instagram etc, if you want to explore.

What do you need before starting to work?

- a twitter account (it asks for your phone number - this is why I am not making it mandatory)
- Through twitter account: API Key, API secret; access token, access token secret
- install required libraries as needed: ROAuth, twitteR, streamR, rTweet, tweetscores etc
- Use existing documentation: e.g., you can look at the documentation for twitteR and understand what you can do with it. `https://cran.r-project.org/web/packages/twitteR/twitteR.pdf`

Free course materials online on using Twitter data in R

- New York university has a 3 day crash course on "Data Science and Social Science".
- Their materials are online: `https://github.com/pablobarbera/data-science-workshop`
- All their course slides and R code are free! So, you can take a look if you want to work with some social media data for course projects!

Free course materials online on using Twitter data in R

- New York university has a 3 day crash course on "Data Science and Social Science".
- Their materials are online: `https://github.com/pablobarbera/data-science-workshop`
- All their course slides and R code are free! So, you can take a look if you want to work with some social media data for course projects!
- A workshop: "Collecting and Analyzing Social Media data with R" happened last week, and all its materials are also free and publicly shared! `https://github.com/pablobarbera/social-media-workshop`

Free course materials online on using Twitter data in R

- New York university has a 3 day crash course on "Data Science and Social Science".
- Their materials are online: `https://github.com/pablobarbera/data-science-workshop`
- All their course slides and R code are free! So, you can take a look if you want to work with some social media data for course projects!
- A workshop: "Collecting and Analyzing Social Media data with R" happened last week, and all its materials are also free and publicly shared! `https://github.com/pablobarbera/social-media-workshop`
- Initial part of this article: (`https://goo.gl/ojPsYU`) also gives an overview of what you need to setup twitter and R to work together.
- You can look for other online tutorials, but look for recent ones (may be after 2015).

Next Class

- Back to corpus analysis, where we left in Week 2.
- Read: Chapter 4 in the textbook
- If possible: Take a look at the WordFreq.R code from last week, to remind yourself what we did in the past
- I posted a question on the forum for today - answer that question before next class