LIN 380 Coursebook

Text as Data: An introduction to quantative text analysis and reproducible research with R

Jerid Francom

April 22, 2021 (latest version)

Contents

About **2**
TODOs . . . 2
Build and session information . . . 2

I Welcome **3**

Overview **3**

About this coursebook **3**
Approach . . . 4
Conventions . . . 4

R and RStudio **6**

Git and GitHub **6**

II Foundations **6**

Overview **6**

1 Introduction to text analysis **7**
1.1 Quantitative studies . . . 8
1.2 Quantitative language research . . . 8
Activities . . . 8

III Orientation **8**

Overview **8**

2 Understanding data **8**
2.1 … . . . 8

3 Statistical approaches **8**
3.1 … . . . 8

4 Framing research **8**
4.1 … . . . 9

IV Preparation **9**

Overview **9**

5 Acquire data **9**
5.1 … 9

6 Curate data **9**
6.1 … 9

A … **10**

About

…

TODOs

- Consider creating an R cheat sheet for text analytics in R. templates for creating cheat sheets[1], examples[2]
- …

Build and session information

This coursebook was written in bookdown[3] inside RStudio[4]. The website is hosted with GitHub Pages[5] and the complete source is available from GitHub[6].

This version of the coursebook was built with:

```
#> Finding R package dependencies ... Done!
#>  setting  value
#>  version  R version 4.0.2 (2020-06-22)
#>  os       macOS  10.16
#>  system   x86_64, darwin17.0
#>  ui       X11
#>  language (EN)
#>  collate  en_US.UTF-8
#>  ctype    en_US.UTF-8
#>  tz       America/New_York
#>  date     2021-04-22
```

And depends on these packages:

[1] https://www.rstudio.com/resources/cheatsheets/how-to-contribute-a-cheatsheet/
[2] https://www.rstudio.com/resources/cheatsheets/
[3] http://bookdown.org/
[4] http://www.rstudio.com/ide/
[5] https://pages.github.com/
[6] https://github.com/francojc

Search:

package	loadedversion	date	source
All	All	All	All
base64enc		2015-07-28	CRAN (R 4.0.2)
bookdown	0.22	2021-04-22	CRAN (R 4.0.2)
digest	0.6.27	2020-10-24	CRAN (R 4.0.2)
evaluate	0.14	2019-05-28	CRAN (R 4.0.2)
glue	1.4.2	2020-08-27	CRAN (R 4.0.2)
highr		2021-04-16	CRAN (R 4.0.2)
htmltools	0.5.1.1	2021-01-22	CRAN (R 4.0.2)
jsonlite	1.7.2	2020-12-09	CRAN (R 4.0.2)
knitr	1.32	2021-04-14	CRAN (R 4.0.2)
magrittr	2.0.1	2020-11-17	CRAN (R 4.0.2)
markdown		2019-08-07	CRAN (R 4.0.2)
mime		2021-02-13	CRAN (R 4.0.2)
rlang	0.4.10	2020-12-30	CRAN (R 4.0.2)
rmarkdown	2.7	2021-02-19	CRAN (R 4.0.2)
stringi	1.5.3	2020-09-09	CRAN (R 4.0.2)
stringr	1.4.0	2019-02-10	CRAN (R 4.0.2)
tinytex		2021-03-30	CRAN (R 4.0.2)
xfun	0.22	2021-03-11	CRAN (R 4.0.2)
yaml	2.2.1	2020-02-01	CRAN (R 4.0.2)

\begin{center}

Part I

Welcome

Overview

WELCOME

... overview text

Learning outcomes

- PS (2) demonstrate ability to produce collaborative and reproducible research using R, RStudio, and GitHub

Learning goals

- ...

About this coursebook

In recent years there has been a growing buzz around the term 'Data Science' and related terms; data analytics, data mining, *etc.* In a nutshell data science is the process by which an investigator leverages statistical methods and computational power to uncover insight from large datasets. Driven in large part by the increase in computing power available to the average individual and the increasing amount of electronic data that is now available through the internet, interest in data science has expanded to virtually all fields in academia and areas in the public sector. Data scientists are in high demand and this trend is expected to continue into the foreseeable future, which means that undergraduate and post-graduate students will be increasingly seeking out resources and training in the area.

This coursebook is an introduction to the fundamental concepts and practical programming skills from Data

Science that are increasingly employed in a variety of language-centered fields and sub-fields. It is geared towards advanced undergraduates and graduate students of linguistics and related fields. As quantitative research skills are quickly becoming a core aspect of many language programs, this coursebook aims to provide a fundamental understanding of theoretical concepts, programming skills, and statistical methods for doing quantitative text analysis. Through these skills we will explore topics and replicate previous research in a variety of areas in language research (psycholinguistics, sociolinguistics, translation studies, *etc.*) using common methods and authentic data sources.

No programming knowledge is assumed, either with R or otherwise.

Approach

Many textbooks on doing 'Data Science', even those that have a domain-centric approach, such as text analysis, tend to focus on the basic 'tidy' approach, seen in Figure 1. to analysis and do not tend to encourage readers to lead with research questions. A big part, or perhaps the biggest part of doing quantitative research, and research in general is what is the question to be addressed. Then comes how to orient the research approach to best address this question (or questions). Then we move on to matching data sources, organizing data, modeling data, and finally reporting findings

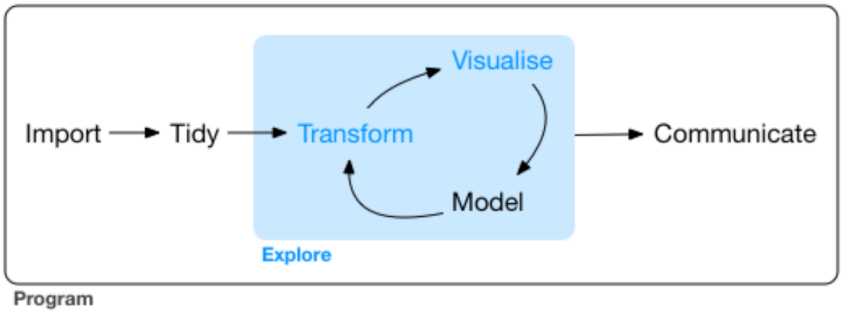

Figure 1: Wickham diagram...

I think a central advantage to this coursebook for language researchers is to thread the project goals without technical implementation in mind first.

Then, after a general idea about what the data should look like, how it should be analyzed, and how the analysis will contribute to knowledge in the field, we can move towards implementing these preliminary formulations in R code. In a way this is the classic separation between content and format –the content of our research should precede the format it should or will take.

Conventions

This coursebook is about the concepts for understanding and the techniques for doing quantitative text analysis with R. Therefore there will be an intermingling of prose and code presented. As such, an attempt to establish consistent conventions throughout the text has been made to signal reader's attention as appropriate. As we explore concepts, R code itself will be incorporated into the text. This may be a unique textbook compared to others you have seen. It has been created using R itself –specifically using an R language package called `bookdown` (Xie, 2021). This R package makes it possible to write, execute ('run'), and display code and results within the text.

For example, the following text block shows actual R code and the results that are generated when running this code. Note that the hashtag `#` signals a **code comment**. The code follows within the same text block and a subsequent text block displays the output of the code.

```r
# Add 1 plus 1
1 + 1
#> [1] 2
```

Inline code will be used when code blocks are short and the results are not needed for display. For example, the same code as above will sometimes appear as `1 + 1`.

When necessary meta-description of code will appear. This is particularly relevant for R Markdown documents.

````
```{r test-code}
1 + 1
```
````

In terms of prose, key concepts will be signaled using ***bold italics***. Terms that appear in this typeface will also appear in the [glossary] at the end of the text. Furthermore, there are four pose text blocks that will be used to signal the reader's attention: *key points*, *notes*, *tips*, and *warnings*.

Key points summarize the main points to be covered in a chapter or a subsection of the text.

In this chapter you will learn:

- the goals of this textbook
- the reasoning for using the R programming language
- important text conventions employed in this textbook

Notes provide a bit more information on the topic or where to find more information.

R is more than a powerful statistical programming language, it also can be used to perform all the necessary steps in a data science project; including reporting. A relatively new addition to the reporting capabilities of R is the `bookdown` package (this textbook was created using this very package). You can find out more here[7].

Tips are used to signal helpful hints that might otherwise be overlooked.

During a the course of an exploratory work session, many R objects are often created to test ideas. At some point inspecting the workspace becomes difficult due to the number of objects displayed using `ls()`.

To remove all objects from the workspace, use `rm(list = ls())`.

Errors will be an inevitable part of learning, but some errors can be avoided. The text will used the warning text block to highlight typical pitfalls and errors.

Hello world!
This is a warning.

Although this is not intended to be a in-depth introduction to statistical techniques, mathematical formulas will be included in the text. These formulas will appear either inline $1 + 1 = 2$ or as block equations.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \, \hat{P}(c) \prod_i \hat{P}(w_i|c) \tag{1}$$

Data analysis leans heavily on graphical representations. Figures will appear numbered, as in Figure 2.

```r
library(ggplot2) # load graphics package
ggplot(mtcars, aes(x = hp, y = mpg)) + # map 'hp' and 'mpg' to coordinate space
  geom_point() + # add points
  geom_smooth(method = "lm") + # draw linear trend line
  labs(x = "Horsepower", # label x axis
       y = "Miles per gallon", # label y axis
       title = "Test plot", # add title
       subtitle = "From mtcars dataset") # add subtitle
```

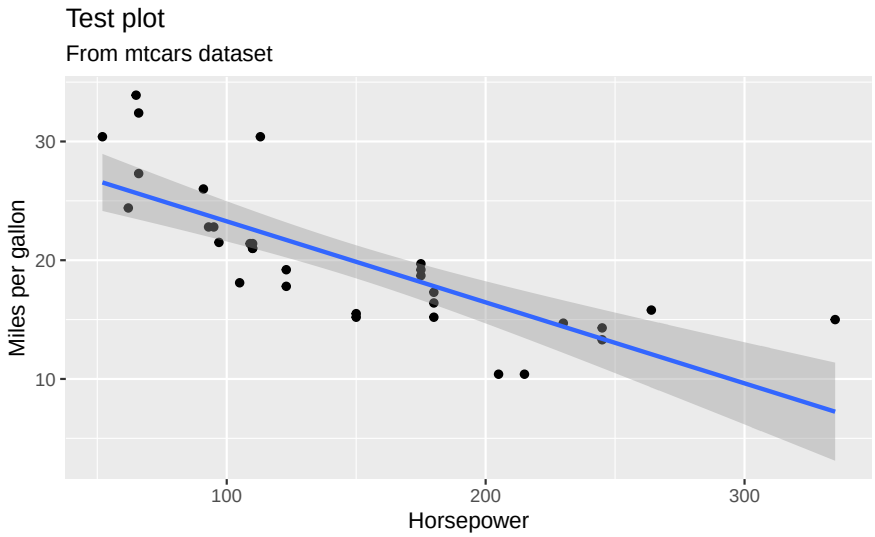

Figure 2: Test plot from mtcars dataset

Tables, such as Table 1 will be numbered separately from figures.

```r
knitr::kable(head(iris, 20), caption = "Here is a nice table!", booktabs = TRUE)
```

R and RStudio

Git and GitHub

(Bryan, 2017)

Part II

Foundations

Overview

FOUNDATIONS

In this section the aim is to (1) provide an overview of quantitative research and their applications, by both highlighting visible applications and notable research in various fields. (2) We will under the hood a bit and consider how quantitative research contributes to language research. (3) I will layout the main types of research and situate quantitative text analysis inside these. Some attention will be given to the historical

Table 1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

background to understand how theory (generative and usage-based grammar) has frame and to some degree continues to frame language research. (4) We will discuss how the programmatic approaches to language, which are fundamental for quantative text analysis, also provide the opportunity to further science through process documentation and research reproducibility.

Learning outcomes

- DL (1) ability to understand and apply data analysis to derive insight from data
- DL (2) ability to understand and apply data knowledge and skills across linguistic and language-related disciplines
- PS (2) demonstrate ability to produce collaborative and reproducible research using R, RStudio, and GitHub
- RS (1) identify an applicable area of investigation in a linguistic or language-related field

Learning goals

-

1 Introduction to text analysis

In this chapter you will learn:

- the goals of this textbook
- the reasoning for using the R programming language
- important text conventions employed in this textbook

1.1 Quantitative studies

Here is a note!

1.2 Quantitative language research

1.2.1 Text analysis

1.2.2 ...

Activities

Here is an activity!

Part III

Orientation

Overview

2 Understanding data

In this chapter you will learn:

- the goals of this textbook
- the reasoning for using the R programming language
- important text conventions employed in this textbook

2.1 ...

(Ackoff, 1989)

3 Statistical approaches

In this chapter you will learn:

- the goals of this textbook
- the reasoning for using the R programming language
- important text conventions employed in this textbook

3.1 ...

4 Framing research

In this chapter you will learn:

- the goals of this textbook

- the reasoning for using the R programming language
- important text conventions employed in this textbook

4.1 ...

Part IV

Preparation

Overview

5 Acquire data

In this chapter you will learn:

- the goals of this textbook
- the reasoning for using the R programming language
- important text conventions employed in this textbook

5.1 ...

5.1.1 Packages

```
library(rvest) # full-fleged web scraping
library(datapasta) # copy/paste approach to HTML tables
```

6 Curate data

In this chapter you will learn:

- the goals of this textbook
- the reasoning for using the R programming language
- important text conventions employed in this textbook

6.1 ...

6.1.1 Packages

```
# ...
```

Data Organization in Spreadsheets (Broman and Woo, 2018). Although based on spreadsheets, many of the best practices discussed apply to good data organization regardless of the technology.

A ...

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16(1):3–9.

Broman, K. W. and Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, 72(1):2–10.

Bryan, J. (2017). Excuse me, do you have a moment to talk about version control? *PeerJ Preprints*, 5:1–23.

Xie, Y. (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.22.