# R for Newspaper Data

*Yann Ryan*

*2019-12-11*

# Contents

# Chapter 1

# Needed

# Chapter 2

# Introduction {intro}

There is a *lot* of newspaper data available now for historical researchers. Once you get hold of it, the rewards can be huge: looking just at English-language users in the last few years, researchers have

## 2.1 Goals

Hopefully, by the end of this book, you will have:

- Know what newspaper data is available, in what format, across a variety of countries and languages.
- Understand something of the various XML formats which make up most newspaper data sources
- Have been introduced to a number of tools which are particulalry useful for large-scale text mining of huge corpora: n-gram counters, topic modelling, text re-use. Including some specific to news, such as the R library *Newsflow*.
- Understand how the tools can be used to answer some basic historical questions (if not provide any answers)

Historians have used newspaper data to do x and y. It can be used as a proxy for public opinion, used to study the movement and genesis of knowledge and information, help to understand the mix of private and public, how power can be advanced through news, and so forth. Some people might not even be interested in the news in its own right:

What do you need in advance?[1]

[1] Bob Nicholson, 'The Victorian Meme Machine: Remixing the Nineteenth-Century Archive', *19: Interdisciplinary Studies in the Long Nineteenth Century*, 2015.21 (2015) <https://doi.org/10.16995/ntn.738>.

## 2.2 Format of the book - bookdown and github

Will try and explain as much as possible, but will take shortcuts[2] Not a programming expert, so it may not be optimised, the best of way doing things. It's just the way I've found that works for me.

There are bits of Python throughout, where I've only managed to work something out using that language. I hope I'll be able to revise at some stage and do everything through one language.

How did I write the book? In R-studio, bookdown and bibdesk ## Why R? Used to be idiosyncratic, is becoming very widely used by data scientist, digital humanities, social scientists. A lot of this is because of developers like Hadley Wickham and R studio - the tidyverse, but also data.table.

### 2.2.1 Who uses R?

### 2.2.2 The Tidyverse

Historians using the language: Sharon Howard,

Writing a book with such a specific goal is a bit of a weird proposition

## 2.3 Sources

### 2.3.1 Country by country

## 2.4 Methods

### 2.4.1 Network analysis of seventeenth century

### 2.4.2 Mapping

### 2.4.3 Geocoding

## 2.5 Text mining

[2]David A. Smith, Ryan Cordell, and Abby Mullen, 'Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers', *American Literary History*, 27.3 (2015), E1–E15 <https://doi.org/10.1093/alh/ajv029>.

# Chapter 3

# Sources {sources}

Sources

Nicholson, Bob, 'The Victorian Meme Machine: Remixing the Nineteenth-Century Archive', *19: Interdisciplinary Studies in the Long Nineteenth Century*, 2015.21 (2015) <https://doi.org/10.16995/ntn.738>

Smith, David A., Ryan Cordell, and Abby Mullen, 'Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers', *American Literary History*, 27.3 (2015), E1–E15 <https://doi.org/10.1093/alh/ajv029>