Web Scraping, Data Wrangling and Visualization

Eyayaw Teka Beze

April 18, 2020

Structure of the Newspaper page, and the scraping work

- ☒ We scraped the first and the second pages of the People's Daily Newspaper, from January 1st, 2019 onwards.
- The newspaper's website structure
  - page-01 (http://paper.people.com.cn/rmrb/html/2020-04/16/nbs.D110000renmrb_01.htm)
  - section-1 (http://paper.people.com.cn/rmrb/html/2020-04/16/nw.D110000renmrb_20200416_1-01.htm)
- ☒ Steps: Article → pages(01&02) → sections/columns → paragraphs

The functions for the scraping

1. ***make_dates(year, mon, from_day, to_day, all_dates)***
2. ***generate_article_url(date, page_num)***
3. get_article_contents.R
   - ***get_article_data(article_urls)***
     $\rightarrow$ tbl of [title, subtitle, content, num_paraghs]
4. ***scrape_article(page_num, dates = NULL, ...)***

   - returns a tbl of successful requests.

5. download_article_data.R

   - With page_nums, years or months we get article data.

6. ***update_article_data(year, page_num, write_to_disk)***

- Data cleaning
  - our data come in a tidy form, i.e. one-section-per-row
  - paste contents of sections together to form a page of newsarticle
  - we then get one-row-per-page-per-day
  - then unnest the contents into one-token-per-row
- Text mining work, tidytext package [Robinson and Silge, 2020]
  - term-frequency(tf)
  - term-frequncy-inverse-document-frequency(tf-idf)
  - n-grams

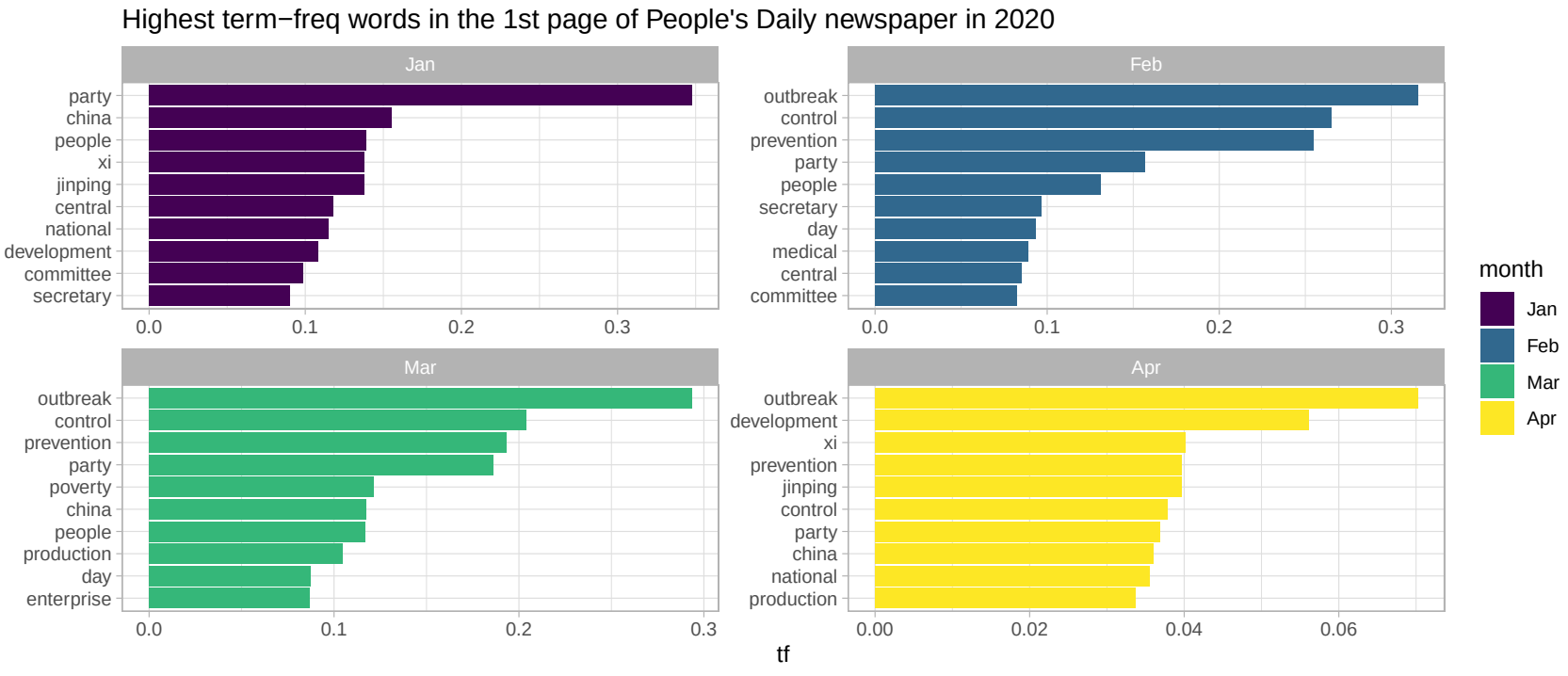
please refer to the rmd file here[1]

[1] https://github.com/Npaffen/Advanced_R_Project/blob/master/analysis/text_analysis.Rmd

Table 1: Descriptive Statistics: How bulky is the daily newspaper?

	Page 1					***Page 2***				
Variable	median	mean	sd	max	min	median	mean	sd	max	min
2019										
num_of_paragraphs	77.0	91.9	55.8	367.0	20.0	70.0	76.6	30.5	246	9.0
num_of_sections	6.0	6.2	1.8	15.0	1.0	6.0	5.7	2.1	9	1.0
paragraph_per_section	12.7	15.6	9.4	65.5	4.1	12.7	18.5	23.5	246	4.2
words	2898.0	3326.3	1756.0	16190.0	803.0	2537.0	2661.3	888.1	8675	219.0
2020										
num_of_paragraphs	69.0	83.1	48.8	341.0	30.0	66.0	70.0	26.5	204	9.0
num_of_sections	7.0	6.4	1.7	11.0	2.0	6.0	6.0	2.0	9	1.0
paragraph_per_section	11.3	13.9	8.9	56.8	4.2	10.0	15.7	22.4	204	4.0
words	2871.0	3137.7	1212.0	9732.0	1508.0	2444.0	2439.9	643.3	4270	232.0

term-frequency, bigrams and trigrams



- We looked not only at the most frequent words but also at bigrams and trigrams.

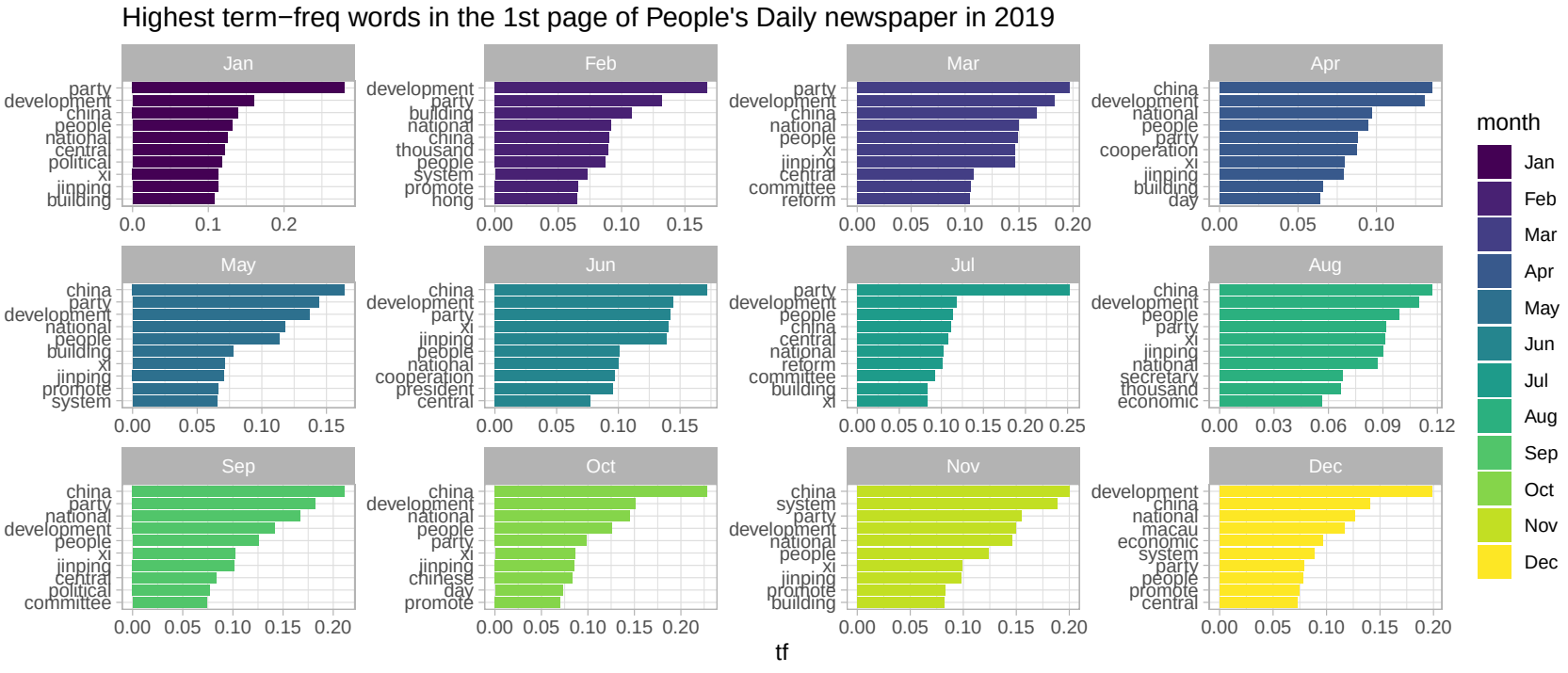
$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

Table 2: The 12 bigrams with the highest tf_idf in 2020

page_num	month	bigram	n	tf	idf	tf_idf
1st	Feb	outbreak prevention	1016	0.01127	0.40547	0.00457
1st	Jan	xi jinping	919	0.00922	0.40547	0.00374
1st	Feb	spotter sue	111	0.00123	2.48491	0.00306
2nd	Apr	zhao pei	17	0.00105	2.48491	0.00261
2nd	Apr	outbreak prevention	89	0.00550	0.40547	0.00223
1st	Apr	basic law	19	0.00117	1.79176	0.00210
1st	Mar	labor education	75	0.00072	2.48491	0.00179
1st	Feb	education supervision	88	0.00098	1.79176	0.00175
1st	Mar	crown pneumonia	260	0.00249	0.69315	0.00173
2nd	Jan	xi jinping	413	0.00414	0.40547	0.00168
1st	Jan	theme education	187	0.00188	0.87547	0.00164
2nd	Mar	crown pneumonia	222	0.00213	0.69315	0.00148

Table 3: The 12 trigrams with the highest tf_idf in 2020

page_num	month	trigram	n	tf	idf	tf_idf
2nd	Apr	zhao pei yu	17	0.00171	2.48491	0.00426
2nd	Apr	ningbo zhoushan port	16	0.00161	1.79176	0.00289
1st	Mar	crown pneumonia outbreak	202	0.00320	0.69315	0.00222
2nd	Apr	resume production complex	52	0.00524	0.40547	0.00213
1st	Mar	resume production complex	294	0.00466	0.40547	0.00189
1st	Jan	party central committee	259	0.00445	0.40547	0.00180
2nd	Feb	health care professionals	134	0.00255	0.53900	0.00137
1st	Jan	china features socialist	189	0.00325	0.40547	0.00132
1st	Jan	era china features	108	0.00185	0.69315	0.00129
2nd	Feb	traditional chinese medicine	87	0.00166	0.69315	0.00115
1st	Feb	education supervision mechanism	24	0.00046	2.48491	0.00113
1st	Mar	rural community workers	43	0.00068	1.38629	0.00094


Highest term−freq words in the 1st page of People's Daily newspaper in 2019
Jan
party
development
china
people
national
central
political
xi
jinping
building
0.0
0.1
0.2
Feb
development
party
building
national
china
thousand
people
system
promote
hong
0.00
0.05
0.10
0.15
Mar
party
development
china
national
people
xi
jinping
central
committee
reform
0.00
0.05
0.10
0.15
0.20
Apr
china
development
national
people
party
cooperation
xi
jinping
building
day
0.00
0.05
0.10
May
china
party
development
national
people
building
xi
jinping
promote
system
0.00
0.05
0.10
0.15
Jun
china
development
party
xi
jinping
people
national
cooperation
president
central
0.00
0.05
0.10
0.15
Jul
party
development
people
china
central
national
reform
committee
building
xi
0.00
0.05
0.10
0.15
0.20
0.25
Aug
china
development
people
party
xi
jinping
national
secretary
thousand
economic
0.00
0.03
0.06
0.09
0.12
Sep
china
party
national
development
people
xi
jinping
central
political
committee
0.00
0.05
0.10
0.15
0.20
Oct
china
development
national
people
party
xi
jinping
chinese
day
promote
0.00
0.05
0.10
0.15
0.20
Nov
china
system
party
development
national
people
xi
jinping
promote
building
0.00
0.05
0.10
0.15
0.20
Dec
development
china
national
macau
economic
system
party
people
promote
central
0.00
0.05
0.10
0.15
0.20
tf
month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

David Robinson and Julia Silge. *tidytext: Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools*, 2020. URL https://CRAN.R-project.org/package=tidytext. R package version 0.2.3.

Vitalie Spinu, Garrett Grolemund, and Hadley Wickham. *lubridate: Make Dealing with Dates a Little Easier*, 2020. URL https://CRAN.R-project.org/package=lubridate. R package version 1.7.8.

Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2019. URL https://CRAN.R-project.org/package=tidyverse. R package version 1.3.0.

Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2019. URL https://CRAN.R-project.org/package=kableExtra. R package version 1.1.0.