Text Mining

Text analysis of Chinese Paper Daily news article

Eyayaw Teka Beze, David Schulze, Nils Paffen

2020-04-18

Contents

1 Description of Data

The data are collected from the Chinese People's Daily newspaper for year 2019 and 2020. The daily newspapers are published on a consistent structure (we give the details of the structure below) online on the newspaper's website. Despite the slow loading speed of the website, We tapped into its structure and have scraped the **first two pages** of the daily newspapers.

For instance, the 1st page of the issue published on *2020-03-22* can be accessed here.

Figure 1: Example newspaper page published on 2020-03-22 (left), and Section or column 1 of it enlarged (right).

We dissect the different characteristics of this particular article as follows and these properties apply to all other articles.

First of all, the main link to the article is this one: (http://paper.people.com.cn/rmrb/html/2020-03/22/nbs.D110000renmrb_01.htm). This takes us to one page of the newspaper, i.e., to the 1st page of the newspaper in this particular example. However, a lot of content is packed on this single page. There are 10 different sections or columns on this article (see Figure 1 for this example newspaper page published on 2020-03-22).

- The prefix of the URL, i.e., **(http://paper.people.com.cn/rmrb/html)** is the same for every article, regardless of page number or edition.
- Date on which it was published: **2020-03/22**
- Page number of the newspaper: **01**
- Section id prefix: **nbs.D110000renmrb**.

As we can see in Figure (1) there are several different sections crammed or squeezed into the single page of the newspaper and these parts (sections) are clickable, and each has a unique id. A click on each section will redirect to a link where one can access the full content of the section in an enlarged view. For example, the first section out of the 10 sections on this example article is section 1 or (see Figure 1 –right). The section id for this first section is **nw.D110000renmrb_20200322_1-01**.

Accordingly, the id of each section on an article is of this form: **nw.D110000renmrb_yyyymmdd_secti page#**. Each section of the single page newspaper has the following additional characteristics.

- Title (**h1 tag**)
- Subtitle (**h3 tag**)
- Number of paragraphs on each page, and
- Content or body of the news article section.

The maximum number of sections on a page is 15, in an article published on 2019-04-26, and the minimum is 1. On average, there were 5.99 sections per a page of an article, for the newspapers published since January 1st, 2019 regardless of the page number. Looking at the frequencies of articles individually, the tendency is for page 2 to have less issues with very high numbers of articles above around 9.

Therefore, the contents of all the sections (**paragraph or p-tags**) together–on the page of the news article– make up the contents of the entire (single) page–which are compactly placed in a single page of the article. Particularly, we scraped the sections of the newspapers–through their unique ids. The date and the page number of the newspaper uniquely identify a newspaper, where the combination of which forms the ids of the sections–prefixed with **nbs.D110000renmrb**.

Moreover, most of the news articles are bulky in terms of number of paragraphs and text volume. On average, there are 82.64 number of paragraphs per page, and 16.59 numbers of paragraphs per section of a single page.

2 Descriptive Statistics

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

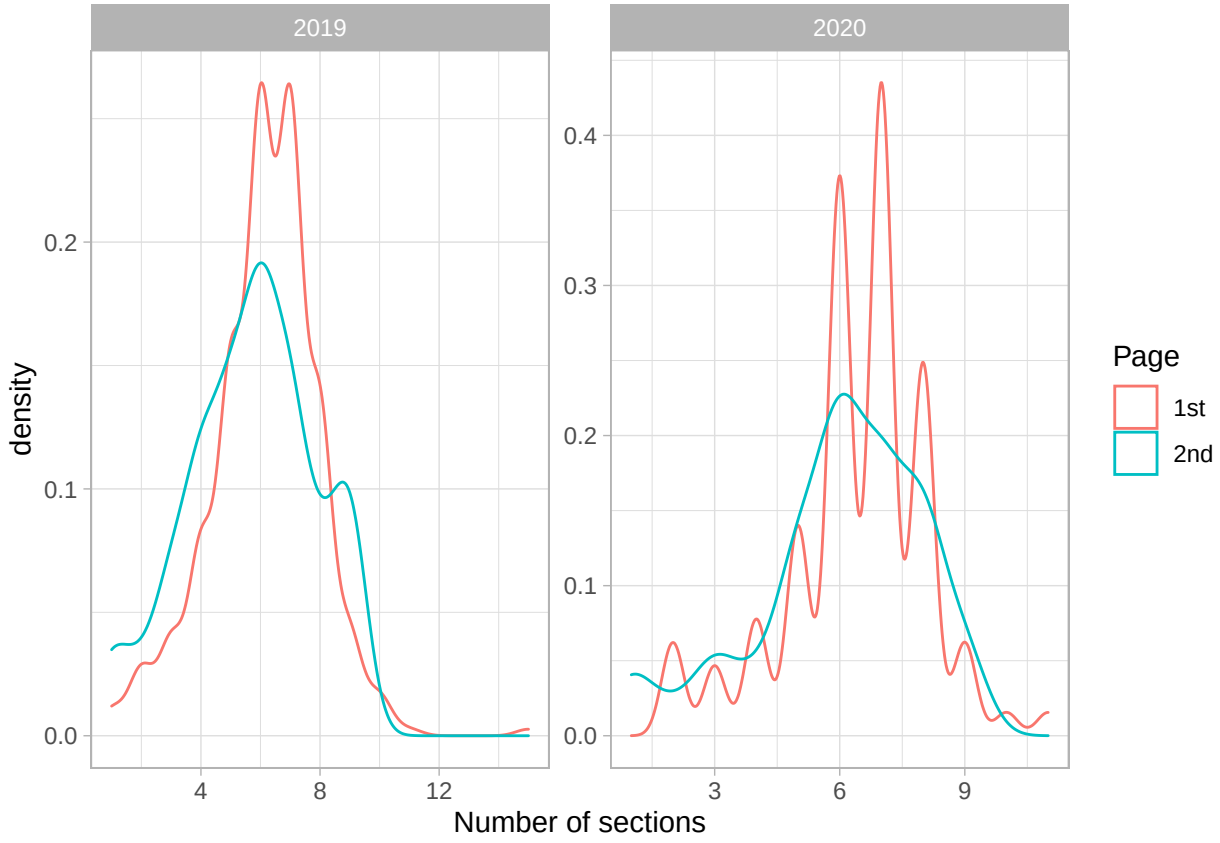

Figure 2: Distribution of **number of sections** in a page of a newspaper

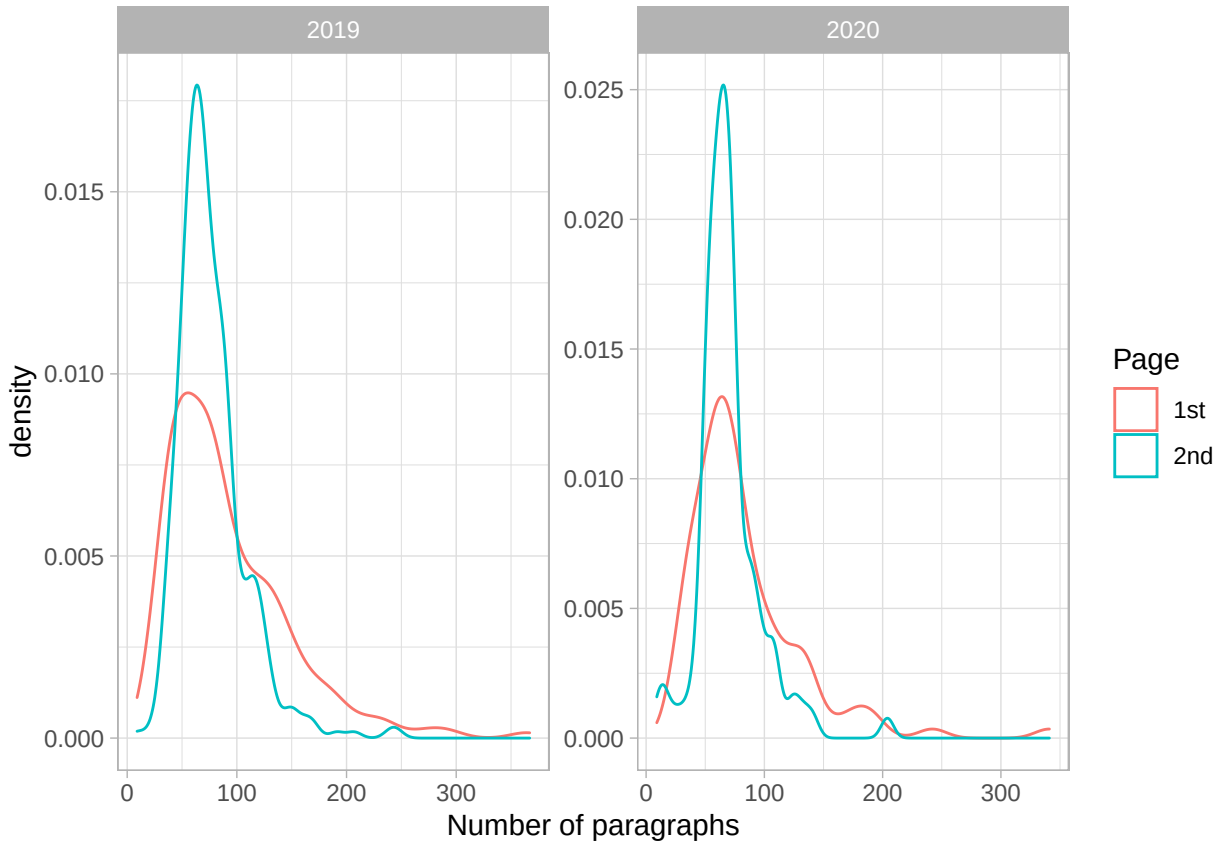

Figure 3: Distribution of Number of Paragraphs in the Newspaper Page

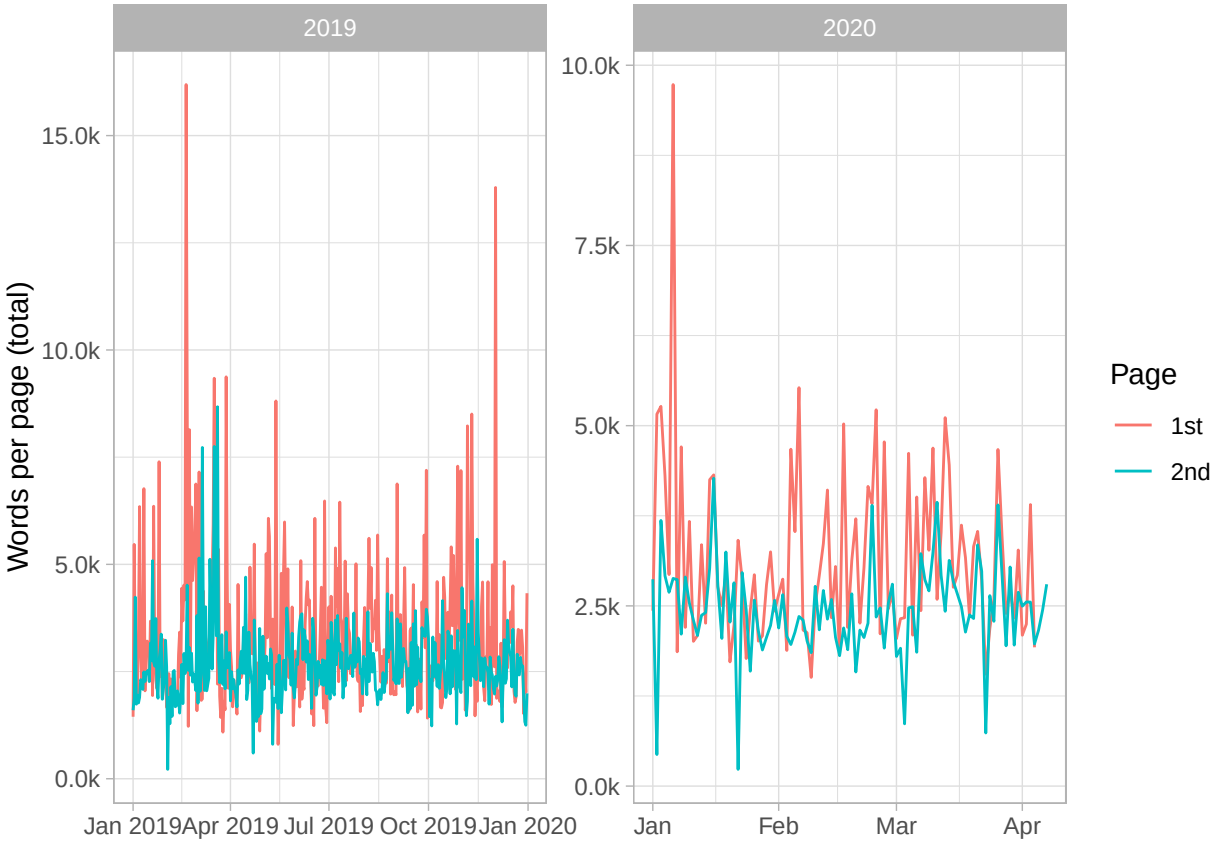

Figure 4: Word counts in the page of the newspaper per day

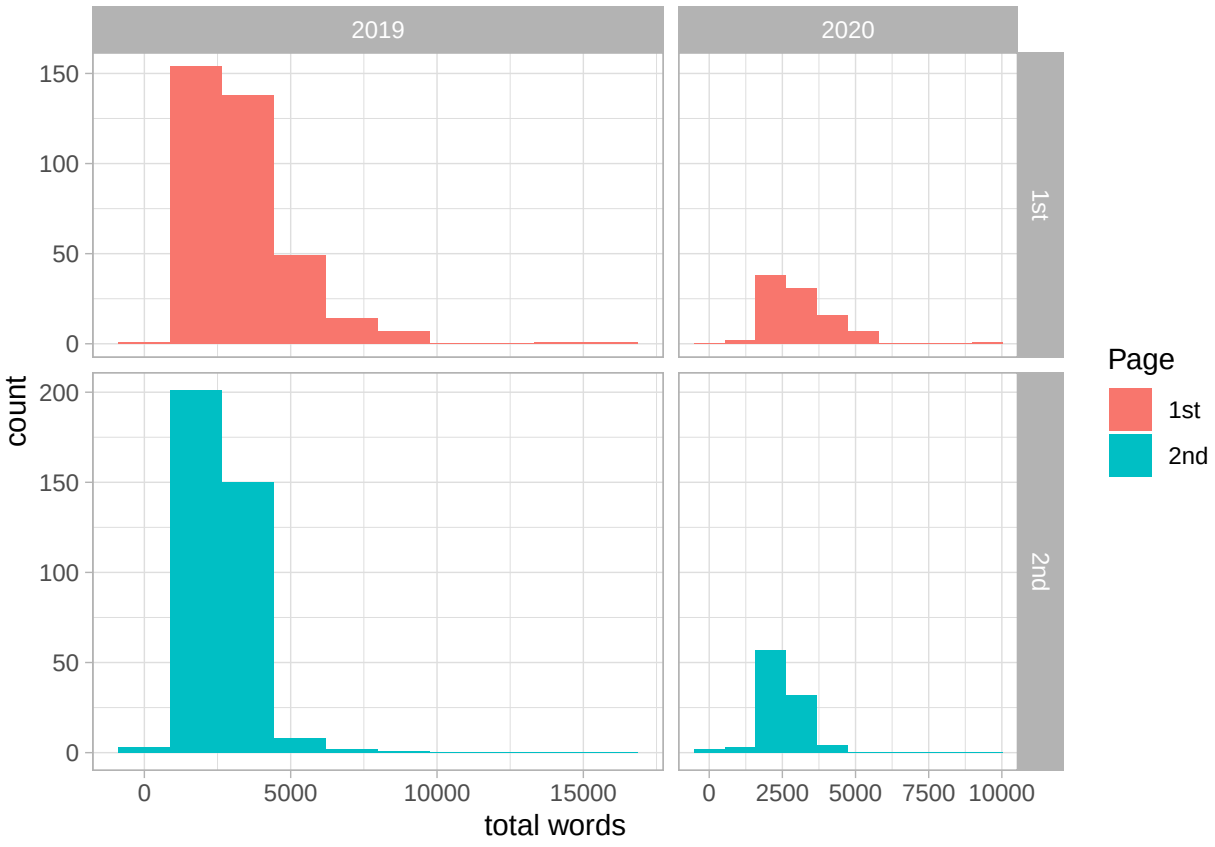

Figure 5: How bulky is a page of a newspaper in terms of word counts?

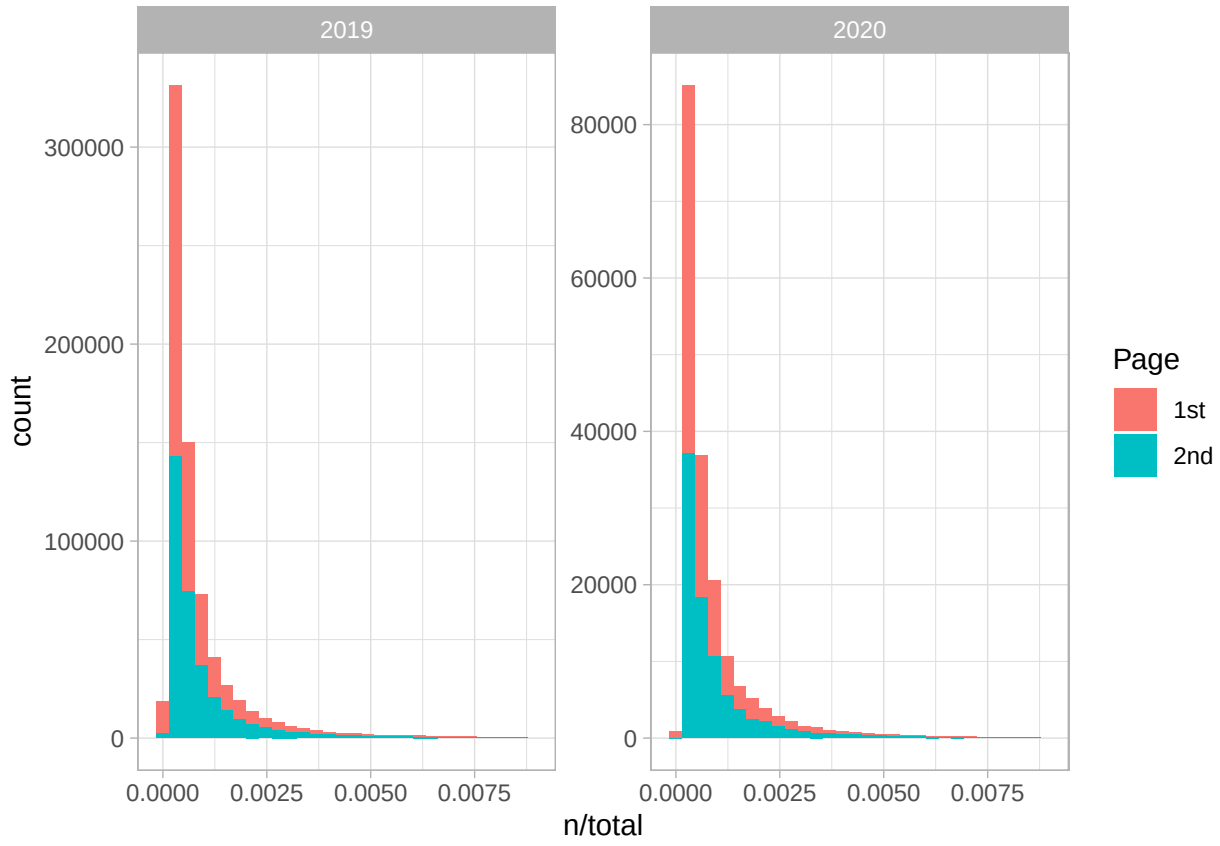

Figure 6: Term Frequency Distribution per page of the newpaper

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