Topic Modeling of New York Times Article Headlines During the Spanish Flu & COVID-19 Pandemic

Term paper SURV727

Anthony Garove and Ujjayini Das

2022-12-09

Contents

Introduction

The overarching purpose of this term paper was to obtain insights about the similarities in major societal issues caused by pandemics, as well as to examine time-era specific differences across pandemics from different points in history. As researchers interested in history, we sought to make historical comparisons across different pandemics using text as data.

We specifically chose to compare the Spanish Flu Pandemic of 1918 to the modern COVID-19 pandemic. We were curious to see how these two pandemics from different centuries compared in terms of the areas of life they affected. We hypothesized that, since each pandemic caused widespread disease of different classes (i.e., an influenza versus coronavirus pandemic), we might observe differences in how each pandemic affected society. Searching the historical archives of printed media could theoretically provide insights to this research question.

We operationalized this research question by asking, "can we observe differences in the topics that emerge from published New York Times article headlines during the first respective month of the Spanish Flu and COVID-19 pandemic using Latent Dirichlet Allocation?"

Data

We used the New York Times Archive API [3] to access N=12,028 article headlines published during the first month of each pandemic. The Spanish Flu was officially declared a pandemic around March 4, 1918 [4]. The COVID-19 public health crisis was officially declared a pandemic on 11th March, 2020 [2].

We used the specific endpoints, month and year, while requesting for published article headlines from the New York Times Archive API with an authorized key. The following code was used to retrieve the data:

```
# Building URLs
base_url1 <-
'https://api.nytimes.com/svc/archive/v1/1918/3.json?api-key=SVyAXhJrhFVCMF4tvGU2GZY3jm1greFU'
request1 <- GET(base_url1)
base_url2 <-
'https://api.nytimes.com/svc/archive/v1/2020/3.json?api-key=SVyAXhJrhFVCMF4tvGU2GZY3jm1greFU'
request2 <- GET(base_url2)

# Getting the Contents as Data Frames
response1 <- content(request1, as = "text", encoding = "UTF-8")
spanishfludata <- fromJSON(response1, flatten = TRUE) %>% data.frame()
response2 <- content(request2, as = "text", encoding = "UTF-8")
coviddata <- fromJSON(response2, flatten = TRUE) %>% data.frame()

# Extract the Subsetted Data
headline_spanish <- spanishfludata[,19]
headline_covid <- coviddata[,20]
```

We first formatted the JSON objects into two data frames — one for the Spanish Flu and another for the COVID-19. After that, we extracted the headlines from both of the data frames and used those extracted objects for the remaining analysis.

Results

Pre Processing

Next, we pre-processed the headline texts using standard text mining procedures in order to prepare the data for topic modelling [5]. These processes included tokenization, lemmatization and removing stopwords.Tokenization is a method used in natural language processing that separates text data into units called "tokens". Tokenization can be done to break up data in to words, characters, and subwords. We used n=1 gram tokenization, which separates our headline data into one-word texts [5].

Lemmatization is a process that normalizes text such that words are substituted with their base root. This step is used to categorize words into groups of root forms with similar meanings. Stopwords are words that are frequently used in a given language, and these words are removed prior to topic modelling. This step is necessary because stopwords contribute noise when performing Latent Dirichlet Allocation, and removing stopwords allows researchers to interpret the topics extracted from text with more coherence. For both lemmatization and removal of stopwords, we used the standard English dictionary lemmas and stopwords [5].

After performing these pre-processing techniques, two corpora of texts were produced; one for Spanish Flu headlines and another for COVID-19 headlines. Based on the two corpora, we then created two document term matrices—where an individual row represents a single document, a column represents a single term, and each value in a matrix contains the frequency of that term in the document [1,5]. Document-term-matrices are used as inputs for topic modelling with the `topicmodels` package. The code below was used for tokenization, lemmatization, removal of stopwards, and creation of document-term-matrices:

```
# Tokenization
headline_spanish_gsub <- gsub("'","",headline_spanish)
headline_covid_gsub <- gsub("’","",headline_covid)
token_spanish <- tokens(headline_spanish_gsub,
                        remove_punct = TRUE,
                        remove_numbers = TRUE,
```

```
                        remove_symbols = TRUE) %>%
  tokens_tolower()
token_covid <- tokens(headline_covid_gsub,
                        remove_punct = TRUE,
                        remove_numbers = TRUE,
                        remove_symbols = TRUE) %>%
  tokens_tolower()

# Lemmatization
library(haven)
lemmaData <- read_sav("./lemma_spss.sav")
corpus_spanish <-  tokens_replace(token_spanish,
                                  lemmaData$inflected_form,
                                  lemmaData$lemma,
                                  valuetype = "fixed")
corpus_covid <-  tokens_replace(token_covid,
                                  lemmaData$inflected_form,
                                  lemmaData$lemma,
                                  valuetype = "fixed")

# Removing Stopwords
corpus_spanish <- corpus_spanish %>%
                            tokens_remove(stopwords("english")) %>%
                            tokens_ngrams(1)
corpus_covid <- corpus_covid %>%
                            tokens_remove(stopwords("english")) %>%
                            tokens_ngrams(1)

#Create document term matrices
DTM_spanish <- dfm(corpus_spanish)
minimumFrequency <- 10
DTM_spanish <- dfm_trim(DTM_spanish,
                min_docfreq = minimumFrequency,
                max_docfreq = 1000000)
DTM_spanish  <- dfm_select(DTM_spanish,
                   pattern = "[a-z]",
                   valuetype = "regex",
                   selection = 'keep')
colnames(DTM_spanish) <- stringi::stri_replace_all_regex(colnames(DTM_spanish),
                                                         "[^_a-z]","")
DTM_spanish <- dfm_compress(DTM_spanish, "features")
sel_idx_spanish <- rowSums(DTM_spanish) > 0
DTM_spanish <- DTM_spanish[sel_idx_spanish, ]
corpus_spanish <- corpus_spanish[sel_idx_spanish, ]
DTM_covid <- dfm(corpus_covid)
minimumFrequency <- 10
DTM_covid <- dfm_trim(DTM_covid,
                min_docfreq = minimumFrequency,
                max_docfreq = 1000000)
DTM_covid  <- dfm_select(DTM_covid,
                   pattern = "[a-z]",
                   valuetype = "regex",
                   selection = 'keep')
```

```
colnames(DTM_covid) <- stringi::stri_replace_all_regex(colnames(DTM_covid),
                                                       "[^_a-z]","")
DTM_covid <- dfm_compress(DTM_covid, "features")
sel_idx_covid <- rowSums(DTM_covid) > 0
DTM_covid <- DTM_covid[sel_idx_covid, ]
corpus_covid <- corpus_covid[sel_idx_covid, ]
```

Topic Modeling Using Latent Dirichlet Allocation

We used Latent Dirichlet Allocation (LDA) to examine topics from our two corpora of headlines. LDA is a frequently used algorithm for fitting a topic model [1,5]. LDA is based on an assumption that documents used in your analysis are comprised of a mixture of topics, and that every topic is an amalgamation of words. LDA simultaneously finds the words associated with each topic while extracting the topics contained in each document [1,5].

LDA for Spanish Flu Headlines: Our initial LDA model was set to extract 20 topics with $\alpha = 3.33$ from the corpus containing New York Times article headlines during the first month of the Spanish Flu Pandemic. After examining the initial results, we reset our model to extract 15 topics with α to 0.2 to make the topics more comprehensible.

Next we observed the probability distributions of the 15 topics over all the headlines, given by `theta_spanish2` as well as the probability distribution of the words over all the 15 topics, given by `beta_spanish2`. To have a better understanding of the topics, we specifically looked into the five terms with the highest per-topic-per word probabilities within each topic.

```
K_spanish <- 15
set.seed(1234)
topicModel_spanish2 <- LDA(DTM_spanish,
                   K_spanish,
                   method="Gibbs",
                   control=list(iter = 500,
                                verbose = 25,
                                alpha = 0.2))
tmResult_spanish2 <- posterior(topicModel_spanish2)
theta_spanish2 <- tmResult_spanish2$topics
beta_spanish2 <- tmResult_spanish2$terms
top5termsPerTopic_spanish2 <- terms(topicModel_spanish2,
                           5)
topicNames_spanish2 <- apply(top5termsPerTopic_spanish2,
                     2,
                     paste,
                     collapse=" ")
topicProportions_spanish2 <- colSums(theta_spanish2) / nrow(DTM_spanish)
names(topicProportions_spanish2) <- topicNames_spanish2
sort(topicProportions_spanish2, decreasing = TRUE)
topicNames_spanish2 <- apply(terms(topicModel_spanish2, 5), 2, paste, collapse = " ")
```

LDA for COVID-19 Headlines: Our initial LDA model was set to extract 20 topics with $\alpha = 5$ from the corpus containing New York Times article headlines during the first month of the COVID-19 Pandemic. After examining the initial results, we reset our model to extract 10 topics with α to 0.15 to make the topics more interpretable. At this point, we observed that the word `Coronavirus` appeared in all 10 topics with the highest β– suggesting that `Coronavirus` is potentially a stopword specific to this corpus of text.

Next we removed the term `Coronavirus` from the corpus and fit the LDA model again. After observing the results from this revised model, we ultimately specified the model to extract 15 topics with $\alpha = 0.15$. This was done because while observing the top five terms for each topic, different combinations of `K` topics and α values repeatedly produced more than 5 terms with equal β-s for many topics. Our final specifications seemed to produce a more parsimonious topic model.

```
headline_covid_gsub2 <- gsub("Coronavirus","",headline_covid_gsub)
token_covid_gsub2 <- tokens(headline_covid_gsub2,
                        remove_punct = TRUE,
                        remove_numbers = TRUE,
                        remove_symbols = TRUE) %>%
  tokens_tolower()
```

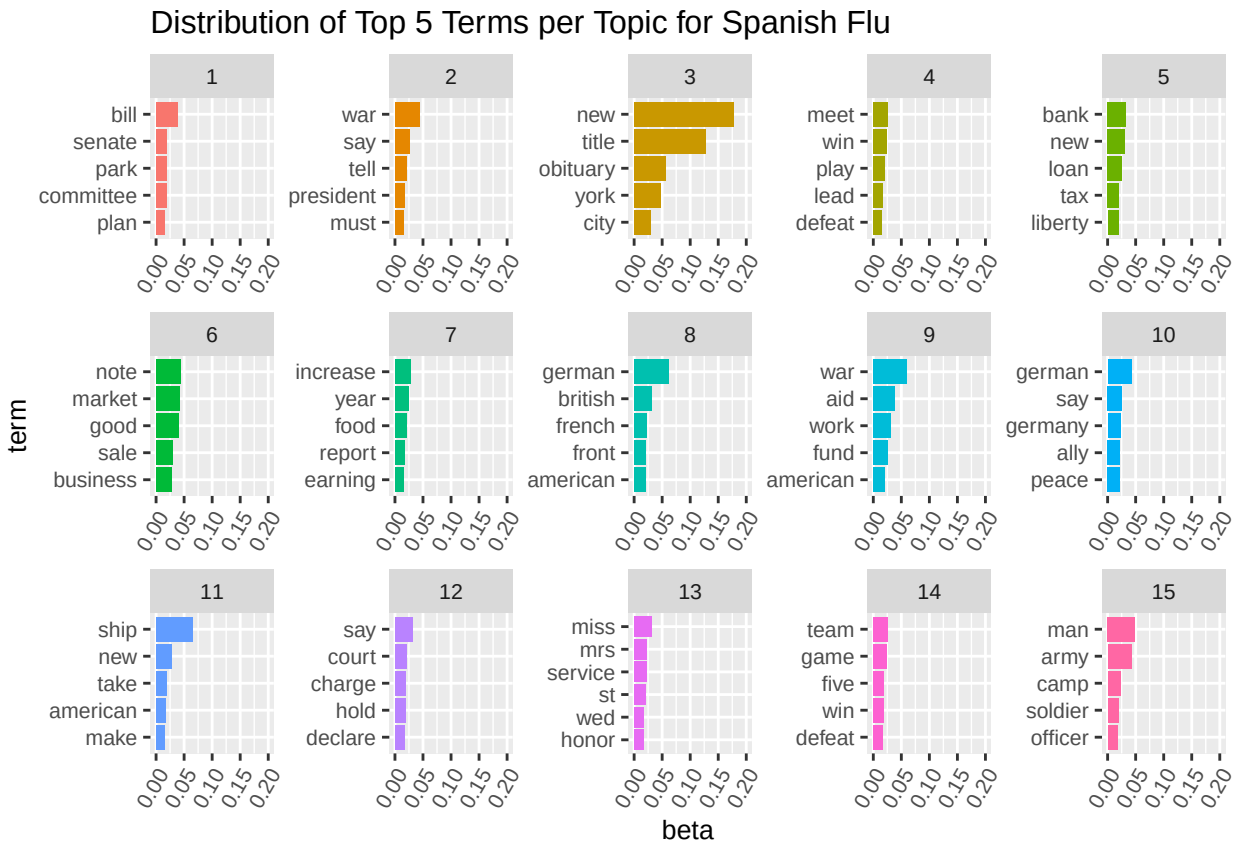
```
K_covid_gsub2 <- 15
set.seed(1234)
topicModel_covid_gsub2 <- LDA(DTM_covid_gsub2,
                  K_covid_gsub2,
                  method="Gibbs",
                  control=list(iter = 500,
                               verbose = 25,
                               alpha = .15))
tmResult_covid_gsub2 <- posterior(topicModel_covid_gsub2)
beta_covid_gsub2 <- tmResult_covid_gsub2$terms
glimpse(beta_covid_gsub2)
theta_covid_gsub2 <- tmResult_covid_gsub2$topics
glimpse(theta_covid_gsub2)
terms(topicModel_covid_gsub2, 10)
top5termsPerTopic_covid_gsub2 <- terms(topicModel_covid_gsub2,
                           5)
topicNames_covid_gsub2 <- apply(top5termsPerTopic_covid_gsub2,
                    2,
                    paste,
                    collapse=" ")
```

Visualizations

To assist with the interpretation of our topic models, we visualized the per-topic-per-term probabilities for both the Spanish Flu and the COVID-19 corpora. We created horizontal bar graphs for each topic [6]; the bars indicating the probabilities of those five words that were most probable to appear within each topic.

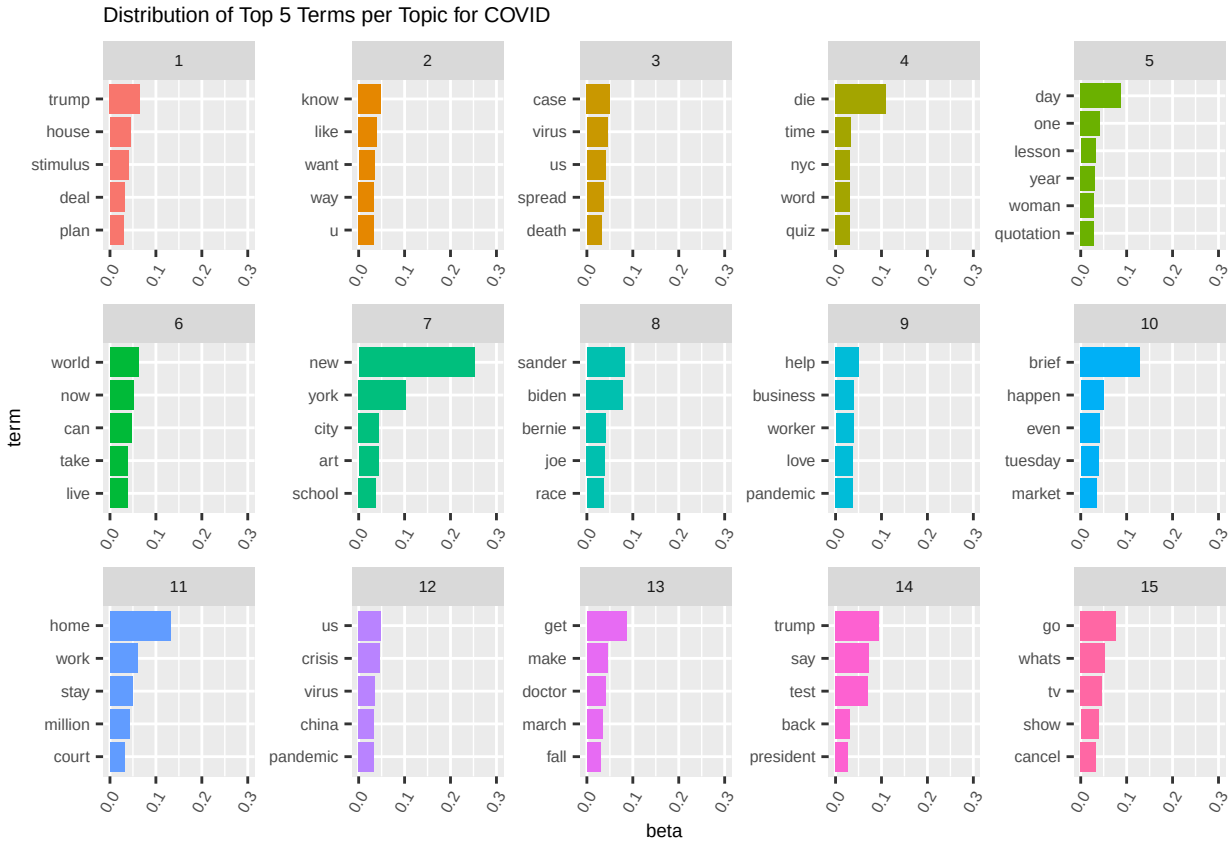
```
### Spanish Flu
## Per Topic Term Distribution
spanishflu_topics <- tidy(topicModel_spanish2,matrix = "beta")
top_terms_spanish <- spanishflu_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 5) %>%
  ungroup() %>%
  arrange(topic, -beta)
top_terms_spanish_graph <- top_terms_spanish %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
   theme(text = element_text(size=10),
        axis.text.x = element_text(angle=60, hjust=1)) +
```

```
  xlim(0.00,0.2)+
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free", ncol= 5) +
  scale_y_reordered()+
  ggtitle("Distribution of Top 5 Terms per Topic for Spanish Flu")
print(top_terms_spanish_graph)
```




```
## COVID
## Per Topic Term Distribution
covid_topics_gsub2 <- tidy(topicModel_covid_gsub2,matrix = "beta")
top_terms_covid_gsub2 <- covid_topics_gsub2 %>%
  group_by(topic) %>%
  slice_max(beta, n = 5) %>%
  ungroup() %>%
  arrange(topic, -beta)
top_terms_covid_graph_gsub2 <-top_terms_covid_gsub2 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  theme(text = element_text(size=7),
        axis.text.x = element_text(angle=60, hjust=1)) +
  xlim(0.00, 0.30) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free", ncol= 5) +
  scale_y_reordered()+
  ggtitle("Distribution of Top 5 Terms per Topic for COVID")
```

```
print(top_terms_covid_graph_gsub2)
```



Discussion

Topic Model of NYT Article Headlines from the First Month of The Spanish Flu Pandemic

Our topic model of published NYT article headlines was specified to extract 15 topics. Overall, it would appear that the majority of the topics extracted from this corpus seem to be related to WWI.

Topics 8 through 10 were comprised of different nations involved in WWI, as well as other war-related terms. Topic 11 was similar to topics 8 through 10, insofar as it appeared to be related to WWI and contained a nation involved in the war (i.e., "American"). Topic 2 was a bit ambiguous, but contained the words "war" and "president"—suggesting that this topic is also related to World War I (WWI). Topic 15 is yet another topic that appeared to be related to the war. It is specifically comprised of terms that could be used when describing the army. Topics 4 and 14 are both comprised of words that could either describe sports or war.

Other interpretable topics were also extracted from this corpus. Topic 1 is comprised of words specifically related to legislative processes in the United States, and topic 12 consists of terms that mostly seem related to judicial processes. Topics 5 through 7 all appear to be related to the US economy, markets, and/or goods.

There were a few other topics that varied in their composition with less interpretability. Topic 3 seemed to be related to New York City (NYC) and had the terms, "obituary" and "title". This topic could perhaps reflect obituaries of persons who resided in NYC. Topic 13 is not so easy to interpret, though it is gendered and related to women in some way. This topic could potentially be related to the woman's suffrage movement, which took place during this time period. It is also noteworthy that there are two terms for this particular topic that have the same probabilities ("wed" and "honor" with a β value of 0.017).

Topic Model of NYT Article Headlines from the First Month of The COVID-19 Pandemic

Our topic model of published NYT article headlines was specified to extract 15 topics from the COVID-19 corpus. Out of these 15 topics, 6 of the topics could not be interpreted. These topics were: 2, 4, 5, 6, 9 and 15.

The majority of remaining interpretable topics were related to the COVID-19 pandemic. Topic 1 seemed to describe the COVID-19 stimulus package introduced by the Trump administration. Topic 3 was comprised of terms that are clearly related to the COVID-19 pandemic (i.e., "case", "virus", and others). Topic 11 described social distancing, with the words "stay" and "home" in its top 5 terms per topic. Topic 12 was comprised of a mix of words related to the COVID-19 pandemic. It should also be noted that one additional cogent topic unrelated to the COVID-19 pandemic was extracted. This was topic 8, which consisted of terms related to the 2020 US presidential election.

There were a few other topics that were not entirely uninterpretable. Topic 7 could be specifically related to New York City, but also contained the words "art" and "school". Topic 10 seemed to possibly describe a briefing related to markets, although this is not entirely clear. Topic 13 might be indirectly related to the COVID-19 pandemic, with words such as "doctor" and "march" in its top 5 terms per topic. Topic 14 seemed to be specifically related to Donald Trump but also contained unrelated words.

Overall Discussion

Overall, the 15 topics extracted from the Spanish Flu corpus of headlines appeared to have more cogency than the topics that were extracted from the COVID-19 corpus of headlines. Specifically, topics in the Spanish Flu corpus appeared to be mostly related to WWI. Topics that were unrelated to WWI were nonetheless coherent, with additional topics about governmental processes, economy and markets, and a few others. The topics that were able to be understood from the COVID-19 corpus were scant in comparison to the topics from the Spanish Flu corpus. Generally, the topics that were extracted from the COVID-19 pandemic were related to the pandemic itself.

The topics that emerged from the Spanish Flu corpus suggests that the early stages of the Spanish Flu pandemic did not appear to receive much coverage in the New York Times during the month of March 1918. It seems that the first month of the Spanish Flu pandemic was eclipsed by the ongoing World War. The topics that emerged from the COVID-19 pandemic corpus suggest that, during the month of March 2020, the topics covered in the New York Times were largely dominated by the COVID-19 pandemic. However, we did observe coverage of the 2020 US presidential election in our topic model, with one topic clearly being related to the election.

The research question posed at the beginning of this paper was, "can we observe differences in the topics that emerge from published New York Times article headlines during the first respective month of the Spanish Flu and COVID-19 pandemic using Latent Dirichlet Allocation?" Our findings indicate that there are differences in the topics that emerged from published New York Times article headlines during the first months of these two pandemics. However, the differences we observed were reflected two different historical events—WWI and the COVID-19 pandemic. We were unable to observe any topics published about the Spanish Flu. Thus, we could not compare topics across both pandemics.

Limitations and Future Directions: If this project was to be repeated, there are a few things that could be done differently. First, future research could specify a larger reference period in the search parameters of the New York Times Archive API. For the Spanish Flu corpus, extending the query's time frame beyond the month of March 1918 may have resulted in headlines related to the Spanish Flu pandemic. In other words, it may have been to early on in the pandemic (with an ongoing war) to have been extensively covered. Only examining headlines from the first respective month of each pandemic may have been too narrow of a time frame to be able to make comparisons between the two pandemics.

Beyond modifying the search parameters of our query in the NYT Archive API, we could have used a different API altogether. If we were to use a different API to collect text data, such as the NYT Article

API, we may have been able to use additional endpoints—resulting in more specific corpora as inputs for our topic models. Using more specific corpora would potentially allow us to make better comparisons between the pandemics.

Another method to answering this research question could have involved entirely different media for topic modeling. One approach could be to use published books [1] about each pandemic as the two corpora for fitting topic models. If we were to use two books specifically about the history of each pandemic as corpora, we would then be able to directly compare topics across both pandemics.

References

In-text citations noted in brackets []

1. Baumer, B., Kaplan, D., & Horton, N. J. (2021). Modern data science with R. Crc Press. https://mdsr-book.github.io/mdsr2e/
2. Morens DM, Daszak P, Markel H, Taubenberger JK. 2020. Pandemic COVID-19 joins history's pandemic legion. https://doi.org/10.1128/mBio.00812-20.
3. Nytimes.com. (2022). https://developer.nytimes.com/docs/archive-product/1/overview
4. Reid, A,H, Taubenberger, J.K., & Fanning, T.G. (2001). The 1918 Spanish influenza: Integrating history and biology. Microbes and Infection, 3 (1) (2001), pp. 81-87
5. Silge, J., & Robinson, D. (2017). Text mining with R : a tidy approach. O'reilly. https://www.tidytextmining.com/
6. Wickham, H. (2019). ggplot2 Elegant Graphics for Data Analysis Second Edition GitHub. https://github.com/hadley/ggplot2-book