New York Times 기사 분석을 통한 COVID19 현상의 글로벌 흐름 분석

Author

고귀환 2018314374 인공지능 융합전공
김시인 2017310474 아동청소년학과, 데이터사이언스 융합전공
박준혁 2018311199 글로벌바이오메디컬공학과, 인공지능 융합전공
최재영 2016312411 문헌정보학과
한승희 2018314848 인공지능 융합전공

Abstract

1월에 중국 우한을 시작으로 불과 몇 달 만에 전세계적 역병(pandemic)으로 번진 역사적 역병인 covid-19는 생활습관부터 경제까지 흐름을 완전히 뒤바꿔 버리는 역사책에 남을 역사적 사건이 되어버렸다. 우리는 이런 COVID-19가 대중의 의식에 미치는 영향을 알고 싶었고, 대중들은 어떻게 이러한 역사적 사건을 받아들이는지 알고 싶었다. 그렇기에 해당 연구는 COVID-19가 발생하고 전세계적으로 퍼지는 과정에서 언론사의 COVID-19과 관련된 표현 변화를 연구하였다. 미국의 대표적인 언론인 뉴욕 타임즈가 이 사건을 어떻게 바라보는지에 대한 관점 변화를 중심으로 조사하고, 대중의 관심이 어디로 옮겨가는지 알아보려고 하였다. 연구자료 범위는 뉴욕 타임즈의 기사로, 기간은 1월부터 5월까지로 한정하였다. 전처리 후 단어의 개수를 세어 월별로 분류하는 방식으로 연구하였다. 연구 결과는 월별/단어별 빈도수이며 word cloud와 bar plot으로 시각화 하였다. 이와 같은 연구를 통해서 뉴욕 타임즈의 코로나에 대한 견해가 1,2월에는 중국 및 동아시아에서 벌어지는 유행병에서 3월에는 전세계적 감염, 4,5월에는 전세계적 역병(pandemic)으로 변했다는 것을 알 수 있었다.

Keywords: 뉴욕 타임즈, 데이터 분석, COVID-19, pandemic

1. Proposed scheme

1.1. 주제

New York Times 기사 분석을 통한 COVID19 현상의 글로벌 흐름 분석

1.2. 연구 배경

Coronavirus는 2020년 1월부터 지금까지 전 세계적으로 사람들에게 커다란 영향력을 준 이슈였다. Coronavirus로 인해 경제, 문화, 교육 등 사람들의 생활이 많이 변화하였다. 사회적 거리두기로 인해 비대면 수업과 회의가 이루어지고, outdoor 활동 대신 indoor 활동을 즐기는 사람들이 많이 증가하였다. 언론에서도 Coronavirus는 중요한 이슈이다. 언론에서 Coronavirus가 어떻게 다루어 지는지 분석을 통해 2020년 초기부터 현재까지의 Coronavirus에 대한 전체적인 흐름을 파악하고 싶어 연구를 진행하게 되었다. Coronavirus가 우리나라에 한정되는 문제가 아닌 전세계적인 문제라는 점을 고려해서 미국의 대표 언론사 중 하나인 New York Times를 연구 대상으로 선택하였다.

1.3. 연구 계획

Coronavirus 관련 New York Times 기사 데이터를 수집하여 분석할 계획이다. 한국에 첫 확진자가 발생한 이후 2020년 1월 21일부터 2020년 5월 19일까지 New York Times 기사를 웹 스크래핑할 것이다. 기사 단어들을 전처리한 후, 단어들의 빈도를 계산하여 정량적 분석을 할 예정이다. 마지막으로, word cloud와 bar plot 두 가지 시각화 표현 기법을 통해 Coronavirus에 대한 New York Times 기사 단어 흐름을 표현하고 언론에서의 단어 흐름 변화를 분석해보고자 한다.

1.4. 연구 방법

1) 분석 데이터
New York Times 홈페이지에서 “coronavirus” 또는 “COVID19”라고 검색하였을 때 나온 관련 기사들이 분석 대상 데이터이다. 수집 기사 날짜는 2020년 01월 21일부터 2020년 05월 19일까지이다. 기사 개수는 coronavirus 검색 결과 1134개, 그리고 COVID19 검색 결과 796개로 총 분석 기사 개수는 1930개이다. 1930개의 New York Times 기사 제목과 본문을 웹 스크래핑한다.

2) 분석 절차
먼저, 단어의 출현 빈도는 정량적으로 접근이 가능하므로, 수량화 하여 단어의 빈도를 계산하고 1차적 정량적 분석을 수행한다. 분석에 앞서, New York Times 기사 웹 스크래핑과 불필요한 단어들을 전처리하는 작업을 진행한다. 다음으로, 분석 결과를 한 눈에 파악하기 쉽도록 word cloud와 bar plot 두 가지 시각화 방법을 활용하여 2차적 시각화 표현과 COVID19 현상의 글로벌 흐름 분석을 수행한다.

3) 분석 기법
구체적으로, R programming을 활용해 텍스트 마이닝 기반으로 분석을 진행한다. 먼저 웹 스크래핑을 통한 New York Times 기사 데이터를 수집한다. 그리고 tm과 textstem을 통해 분석에 불필요한 단어들에 전처리 작업을 한다. 다음으로 rJava와 KoNLP를 통한 형태소 분석과 단어 빈도 계산을 통해 1차적 정량적 분석을 수행한다. 마지막으로 word cloud와 bar plot을 통한 기사 단어 빈도 기반 2차적 시각화 분석을 진행한다.

1.5. 연구 기대결과
New York Times 기사 분석을 통해, 1월부터 5월까지 월별로 언론에서의 Coronavirus 관련 단어의 흐름 변화 파악과 분석을 기대한다. 구체적으로, 1월에는 아시아권 국가에서의 Coronavirus에 대한 표현이 많이 나타날 것이라고 예상한다. 이후 영국, 이탈리아와 같은 유럽 국가들과 미국 등 다양한 지역에서의 Coronavirus 표현이 언론에서 사용되었을 것이라고 예상한다. 이와 더불어, Coronavirus 관련 using mask, social distance와 같은 사람들의 보건과 건강, 생활 방침과 관련된 단어가 많이 등장할 것이라고 예상한다.

Coronavirus는 사실 2020년 상반기의 가장 큰 사회적, 국제적 이슈였다. Coronavirus에 대한 New York Times 기사 분석을 통해, Coronavirus가 세계적으로 어떻게 전파되고 진행이 되었는지에 대한 분석과 흐름 이해를 기대한다. 그 결과, 최종적으로 Coronavirus에 대한 사회적, 국제적 전반적인 흐름 변화를 파악할 수 있을 것으로 기대한다.

2. Previous work

2019년 12월 중증급성호흡기증후군 코로나바이러스 2(SARS-CoV-2)로 지정된 한 바이러스가 중국 중부 우한에서 온 환자들에게 원인 불명의 폐렴으로 처음 확인됐다. 코로나바이러스 2019년 1차 보고 이후 전 세계적인 감염병으로 번져 대규모 감염이 발생했다. 현재 ‘코로나 바이러스’는 개인의 사소한 생활습관부터 국가 경제, 외교상황까지 영향을 미치지 않는 곳이 없을 정도로 모든 것을 바꾸어 놓았으며, 전 세계적으로 가장 관심이 많은 주제이다. 그 영향은 학술계에도 큰 영향을 미쳤고, 코로나에 관련된 데이터를 분석하는 연구가 활발히 진행되었다. 그래서 우리 팀 프로젝트의 진행에 도움이 되고자, 어떤 선행 연구들이 있는지 먼저 분석하였고, 웹 크롤링으로 데이터를 모으기 전에 어떤 코로나에 관련된 어떤 기사들이 있는지 조사하였다.

2.1. Cross-Country Comparison of Case Fatality Rates of COVID-19/SARS-COV-2

Objectives
Case fatality rates (CFR) and recovery rates are important readouts during epidemics and pandemics. In this article, an international analysis was performed on the ongoing coronavirus disease 2019 (COVID-19) pandemic.

Methods
Data were retrieved from accurate databases according to the user’s guide of data sources for patient registries, CFR and recovery rates were calculated for each country. A comparison of CFR between countries with total cases ≥ 1,000 was observed for 12th and 23rd March.

Results

Italy's CFR was the highest of all countries studied for both time points (12th March, 6.22% versus 23rd March, 9.26%). The data showed that even though Italy was the only European country reported on 12rd March, Spain and France had the highest CFR of 6.16 and 4.21%, respectively, on 23rd March, which was strikingly higher than the overall CFR of 3.61%.

Conclusion
Obtaining detailed and accurate medical history from COVID-19 patients, and analyzing CFR alongside the recovery rate, may enable the identification of the highest risk areas so that efficient medical care may be provided. This may lead to the development of point-of-care tools to help clinicians in stratifying patients based on possible requirements in the level of care, to increase the probabilities of survival from COVID-19 disease.

Keywords: coronavirus, COVID-19, case fatality rates

전염병을 분석하는데 중요한 정보인 치사율(CFR)과 회복률(RR)을 기준으로 COVID-19에 대한 국제적인 분석을 한 논문이다. 한 국가만 분석한 게 아니라 중국, 이탈리아를 포함해 총 25개국을 인구수, GDP등 많은 요인을 고려하여 정확한 수치를 나타냈다는 점에서 우리 팀 프로젝트에 유의미하게 사용될 수 있을 것이다.

2.2. The Novel Coronavirus (COVID-2019) Outbreak: Amplification of Public Health Consequences by Media Exposure

Abstract
The 2019 novel coronavirus (COVID-2019) has led to a serious outbreak of often severe respiratory disease, which originated in China and has quickly become a global pandemic, with far-reaching consequences that are unprecedented in the modern era. As public health officials seek to contain the virus and mitigate the deleterious effects on worldwide population health, a related threat has emerged: global media exposure to the crisis. We review research suggesting that repeated media exposure to community crisis can lead to increased anxiety, heightened stress responses that can lead to downstream effects on health, and misplaced health-protective and help-seeking behaviors that can overburden health care facilities and tax available resources. We draw from work on previous public health crises (i.e., Ebola and H1N1 outbreaks) and other collective trauma (e.g., terrorist attacks) where media coverage of events had unintended consequences for those at relatively low risk for direct exposure, leading to potentially severe public health repercussions. We conclude with recommendations for individuals, researchers, and public health officials with respect to receiving and providing effective communications during a public health crisis.

Keywords: infectious disease, media, coronavirus

코로나 사태와 비슷한 전례(에볼라 등)로부터 알아내는 언론 노출이 군중 심리에 미치는 영향을 다룬 논문이다. 반복적인 미디어 노출이 지역사회 위기에 대한 불안감 증가, 건강에 악영향을 주는 스트레스 증가, 의료 시설과 보유 자원에 부담을 크게 줄 수 있는 잘못된 자기방어 및 도움으로 이루어 질 수 있다는 연구에 기반한다. 팀 프로젝트에서 단순히 많이 사용된 단어를 분석하는 텍스트분석으로만 끝나는 것이 아니라 그에 따라 변화하는 군중의 심리까지 예측해 볼 수 있게 하는 좋은 자료이다.

2.3. Estimates of the severity of coronavirus disease 2019: a model-based analysis

Background
In the face of rapidly changing data, a range of case fatality ratio estimates for coronavirus disease 2019 (COVID-19) have been produced that differ substantially in magnitude. We aimed to provide robust estimates, accounting for censoring and ascertainment biases.

Methods
We collected individual-case data for patients who died from COVID-19 in Hubei, mainland China (reported by national and provincial health commissions to Feb 8, 2020), and for cases outside of mainland China (from government or ministry of health websites and media reports for 37 countries, as well as Hong Kong and Macau, until Feb 25, 2020). These individual-case data were used to estimate the time between onset of symptoms and outcome (death or discharge from hospital). We next obtained age-stratified estimates of the case fatality ratio by relating the aggregate distribution of cases to the observed cumulative deaths in China, assuming a constant attack rate by age and adjusting for demography and age-based and location-based under-

ascertainment. We also estimated the case fatality ratio from individual line-list data on 1334 cases identified outside of mainland China. Using data on the prevalence of PCR-confirmed cases in international residents repatriated from China, we obtained age-stratified estimates of the infection fatality ratio. Furthermore, data on age-stratified severity in a subset of 3665 cases from China were used to estimate the proportion of infected individuals who are likely to require hospitalisation.

Findings
Using data on 24 deaths that occurred in mainland China and 165 recoveries outside of China, we estimated the mean duration from onset of symptoms to death to be 17·8 days (95% credible interval [CrI] 16·9–19·2) and to hospital discharge to be 24·7 days (22·9–28·1). In all laboratory confirmed and clinically diagnosed cases from mainland China (n=70 117), we estimated a crude case fatality ratio (adjusted for censoring) of 3·67% (95% CrI 3·56–3·80). However, after further adjusting for demography and under-ascertainment, we obtained a best estimate of the case fatality ratio in China of 1·38% (1·23–1·53), with substantially higher ratios in older age groups (0·32% [0·27–0·38] in those aged <60 years vs 6·4% [5·7–7·2] in those aged ≥60 years), up to 13·4% (11·2–15·9) in those aged 80 years or older. Estimates of case fatality ratio from international cases stratified by age were consistent with those from China (parametric estimate 1·4% [0·4–3·5] in those aged <60 years [n=360] and 4·5% [1·8–11·1] in those aged ≥60 years [n=151]). Our estimated overall infection fatality ratio for China was 0·66% (0·39–1·33), with an increasing profile with age. Similarly, estimates of the proportion of infected individuals likely to be hospitalised increased with age up to a maximum of 18·4% (11·0–37·6) in those aged 80 years or older.

Interpretation
These early estimates give an indication of the fatality ratio across the spectrum of COVID-19 disease and show a strong age gradient in risk of death.

Funding
UK Medical Research Council.

중국 본토 사망자와 본토 외의 사망자를 대상으로 개별 사례 데이터를 수집해서 증상의 시작과 결과 사이의 시간(사망 또는 퇴원)을 추정하고 인구통계학적, 연령 기반 및 위치 기반 미실현에 따라 조정하여 중국에서 관찰된 누적 사망률과 사례의 총 분포를 연관시킴으로써 사례 사망률의 연령 제한 추정치를 계산하여 가시적으로 보여주었다. 결과를 모델, 그래프 등으로 표현해 가시적으로 표현했다는 점에서 본 프로젝트에 중요한 참고자료가 될 것이다.

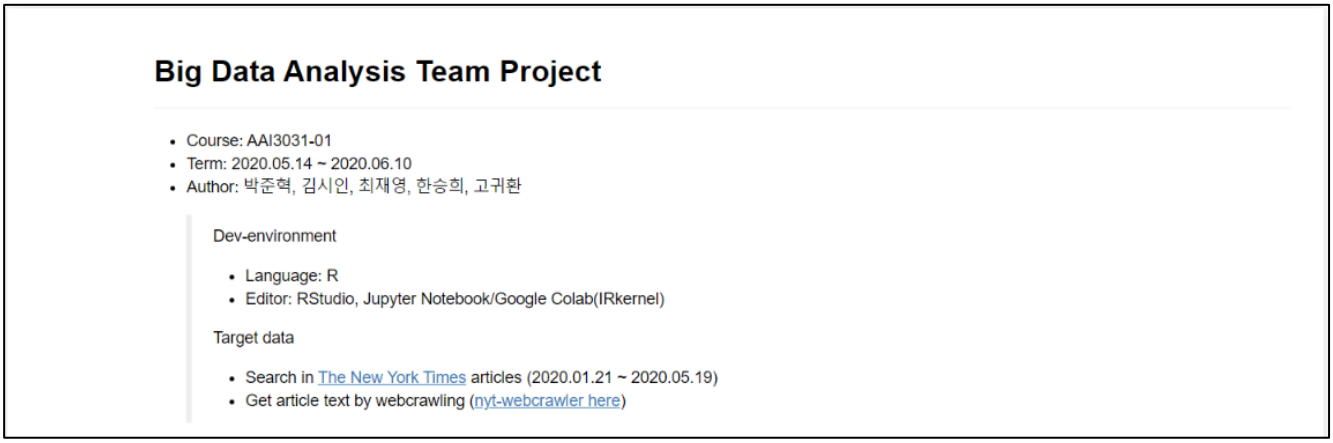
2.4. References

Cross-Country Comparison of Case Fatality Rates of COVID-19/SARS-COV-2:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7104689

The Novel Coronavirus (COVID-2019) Outbreak: Amplification of Public Health Consequences by Media Exposure:
https://psycnet.apa.org/fulltext/2020-20168-001.html

Estimates of the severity of coronavirus disease 2019: a model-based analysis:
https://www.sciencedirect.com/science/article/pii/S1473309920302437#!

3. Performance evaluation



Big Data Analysis Team Project

- Course: AAI3031-01
- Term: 2020.05.14 ~ 2020.06.10
- Author: 박준혁, 김시인, 최재영, 한승희, 고귀환

Dev-environment

- Language: R
- Editor: RStudio, Jupyter Notebook/Google Colab(IRkernel)

Target data

- Search in The New York Times articles (2020.01.21 ~ 2020.05.19)
- Get article text by webcrawling (nyt-webcrawler here)

프로젝트는 2020.05.14에 시작해서 2020.06.10에 끝마쳤다. 분석 코드는 RStudio, Jupyter Notebook/Google Colab(IRkernel) 환경에서 R language로 작성하였다. 협업 환경을 위해 Colab을 일부 이용하였으나 한계가 있어 주로 RStudio로 개별 작성하고, Google drive에 공유하는 방식으로 진행되었다. 분석 대상이 되는 data는 "The New York Times"에서 검색한 기사의 원문을 web crawling하여 수집하였다. 프로젝트 초기에는 아직 R로 crawling하는 방법을 배우지 않아 python으로 nyt-webcrawler를 만들어 사용하였다.
전반적인 코드의 흐름은 아래의 순서로 진행된다.

0. Settings: 코드 실행에 필요한 패키지 설치 및 불러오기
1. Create data file: raw data에서 필요한 data만을 가져와 분석에 쓰일 data file 생성하기
2. Preprocessing: 본문에서 단어를 추출하기 위한 전처리
3. Get word sets: 전처리한 data에서 단어 추출하여 word set 생성
4. Analysis: word set으로부터 의미 있는 결과 확인하기

그리고 전체적인 스크립트는 다음과 같이 작성되었다.

1. 코드에서 single hash-tagged line은 uncomment하여 실행할 수 있다.
2. Pseudo code는 알파벳 대문자 동사로 시작하는 코멘트로 스크립트 내 코드 위에 작성되었다.
3. 결과 확인을 위한 코드를 중간중간 삽입하였으며 데이터가 많아 일부만 출력하기로 한다.

3.0 Settings

0. Settings

In [1]:
```
## You can uncomment and run single hash-tagged lines
## Pseudo code start with UPPERCASE verb
## Install R packages to use if not already installed
## command: install.packages('packagename')

# install.packages('tm') # Install dependencies together
## dependency package required for loading 'tm' package: 'NLP'...

# install.packages('textstem') # Install dependencies together
## dependency packages required for loading 'textstem' package:'sylly' 'koRpus' 'koRpus.lang.en'

# install.packages('rJava')
## required for loading 'KoNLP' package

# install.packages('KoNLP')
## IF NOT INSTALLED : trouble shooting here (https://github.com/haven-jeon/KoNLP)

# install.packages('memoise')

# install.packages('RColorBrewer')

# install.packages('wordcloud2')


## ATTACH packages
library(plyr);library(dplyr);
library(NLP);library(tm);
library(sylly);library(koRpus);library(koRpus.lang.en);library(textstem);
library(rJava);library(KoNLP);
library(memoise);
library(RColorBrewer);library(wordcloud2);
```

```
Attaching package: 'dplyr'


The following objects are masked from 'package:plyr':

    arrange, count, desc, failwith, id, mutate, rename, summarise,
    summarize


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


For information on available language packages for 'koRpus', run

  available.koRpus.lang()

and see ?install.koRpus.lang()


Checking user defined dictionary!
```

install.packages('패키지명')로 필요한 패키지를 설치할 수 있으며 이미 설치되어 있으면 굳이 설치할 필요가 없다. 설치할 패키지는 'tm', 'textstem', 'rJava', 'KoNLP', 'memoise', 'RColorBrewer', 'wordcloud2'이고, 나머지는 dependency package로 함께 설치된다. library(패키지명)으로 필요한 패키지를 불러온다. 아래 결과는 패키지가 사용하는 object이니 사용자 변수와 중복되지 않도록 주의하자.

3.1 Creat data file

1. Creat combined data file for analysis

(0) load raw data files

In [2]:
```
## READ csv data files, THEN GET dataframes 'corona', 'covid'
corona <- read.csv('articles_search_coronavirus.csv')
covid <- read.csv('articles_search_covid-19.csv')

## ADD 'condition' values of each dataframe
corona$condition <- 'corona'
covid$condition <- 'covid'
```

In [3]:
```
## SHOW data info
'corona';str(corona)
'covid';str(covid)
```

```
'corona'

'data.frame':   1134 obs. of  6 variables:
 $ title    : Factor w/ 1134 levels " 'A Heart-Wrenching Thing' : Hospital Bans on Visits Devastate Families",..: 885 463 901 565 1123 7
11 571 393 566 1128 ...
 $ date     : int  20200121 20200121 20200121 20200121 20200121 20200121 20200121 20200121 20200121 20200122 ...
 $ text     : Factor w/ 1134 levels " [Want to get New York Today by email? Here' s the sign-up.]It' s Thursday. Weather: Mostly sunny,
with a high "| __truncated__,..: 82 74 1114 15 20 693 9 349 24 199 ...
 $ section  : Factor w/ 60 levels "Africa","Americas",..: 22 22 5 9 9 5 9 15 9 5 ...
 $ link     : Factor w/ 1134 levels "https://www.nytimes.com/2020/01/21/briefing/impeachment-china-virus-france.html?searchResultPositio
n=10",..: 1134 6 7 2 4 8 3 5 1 18 ...
 $ condition: chr  "corona" "corona" "corona" "corona" ...

'covid'

'data.frame':   796 obs. of  6 variables:
 $ title    : Factor w/ 781 levels " 'A Storm Is Coming' : Fears of an Inmate Epidemic as the Virus Spreads in the Jails",..: 603 285 47
4 537 414 700 443 196 187 766 ...
 $ date     : int  20200121 20200121 20200128 20200129 20200130 20200130 20200131 20200203 20200204 20200205 ...
 $ text     : Factor w/ 796 levels " [Want to get New York Today by email? Here' s the sign-up.]It' s Monday. Because of the coronavirus
outbreak, "| __truncated__,..: 64 56 3 337 588 638 144 625 265 622 ...
 $ section  : Factor w/ 60 levels "Africa","Americas",..: 21 21 33 21 36 21 57 57 57 21 ...
 $ link     : Factor w/ 796 levels "https://www.nytimes.com/2020/01/21/health/cdc-coronavirus.html?searchResultPosition=3",..: 796 1 3 2
4 5 6 7 8 9 ...
 $ condition: chr  "covid" "covid" "covid" "covid" ...
```

nyt-crawler로 얻은 csv file을 각각 read.csv() 함수를 통해 'corona', 'covid' dataframe으로 읽어 들이고, str() 함수를 통해 데이터 객체를 확인한다.

(1) join the datasets, 'corona' and 'covid'

In [4]:
```
## MERGE 'corona', 'covid' dataframes, THEN GET new raw dataframe 'rawdf'
rawdf <- rbind(corona, covid)

## COMPUTE N/As in the dataframe
colSums(is.na(rawdf)) # if there is no N/A, it is well combined.
```

title: 0 **date:** 0 **text:** 0 **section:** 0 **link:** 0 **condition:** 0

두 dataframe을 rbind() 함수로 행방향 결합을 시키고 'rawdf'로 입력한다. 그리고 'rawdf'에 NA가 있는지 계산하여 잘 결합되었는지 알아본다.

(2) split the 'date' into 'month' and 'day'

In [5]:
```
## GET 'month', 'day' data FROM data 'date'
rawdf$month <- rawdf$date %>% substr(5, 6)
rawdf$day <- rawdf$date %>% substr(7, 8)
```

1-(0)의 In[3]에 대한 출력에서 $date 정보가 연, 월, 일이 결합된 형태임을 확인하여 나중을 위해 월($month), 일($day)을 뽑아내어 'rawdf'의 새로운 열로 추가한다.

(3) select needed data

In [6]:
```
## SELECT data FROM raw dataframe 'rawdf', THEN GET new dataframe 'df'
## needed data:  month, day(just in case), title, text, date (just in case)
df <- rawdf %>% dplyr::select(month, day, title, text, date)

## SET data labels
colnames(df) <- c('Month', 'Day', 'Title', 'Text', 'Date')
```

dplyr 패키지의 select를 이용하여 'rawdf'에서 필요한 정보 $month, $day, $title, $text, $date를 골라 새로운dataframe 'df'를 생성한다.
'df'에서 각 행은 $Month, $Day, $Title, $Text, $Date로 지정한다.

(4) combine data, 'Title' and 'Text'

In [7]:
```
## MERGE 'Title', 'Text' data, THEN SET data type AS character, THEN GET data 'Article'
df$Article <- paste(df$Title, df$Text) %>% as.character()
```

기사의 제목($Title)과 내용($Text)를 하나로 결합하여 charcter type의 ($Article)을 새로운 열로 추가한다.

(5) save dataset into csv file

In [8]:
```
## WRITE the dataframe 'df' AS csv file
write.csv(df, 'data.csv', row.names = FALSE)
```

현재까지의 결과 'df'를 'data.csv' 파일로 저장한다.

3.2 Preprocessing

2. Preprocess using 'tm', 'textstem' packages

(0) load data file, change data types

In [9]:
```
## READ csv data files, THEN GET dataframes 'data'
data <- read.csv('data.csv')

## SET data type of data in dataframe AS character
data$Article <- data$Article %>% as.character()
data$Title <- data$Title %>% as.character()
data$Text <- data$Text %>% as.character()

## SET variable of data in dataframe
article <- data$Article
```

In [10]:
```
## SHOW data info and an example of data
str(data)
data %>% tail(1) %>% t()
```

data.csv 파일을 read.csv() 함수를 통해 'data' dataframe으로 불러들인다. dataframe에 담긴 data의 type을character로 변경하고, preprocess할 $Article을 article로 지정한다.
In[10]에서 불러들인 data 예시는 아래에서 확인할 수 있다.

```
'data.frame':   1930 obs. of  6 variables:
 $ Month  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Day    : int  21 21 21 21 21 21 21 21 21 22 ...
 $ Title  : chr  "The Coronavirus: What Scientists Have Learned So Far" "First Patient With Wuhan Coronavirus Is Identified in the U.S."
"The Test a Deadly Coronavirus Outbreak Poses to China' s Leadership" "Impeachment Trial, Davos, Coronavirus: Your Tuesday Briefing" ...
 $ Text   : chr  "A novel respiratory virus that originated in Wuhan, China, last December has spread to six continents. Hundreds"| __tr
uncated__ "A man in Washington State is infected with the Wuhan coronavirus, the first confirmed case in the United States"| __truncated
__ "WUHAN, China — Facing growing pressure to contain a deadly viral outbreak that has spread halfway around the wo"| __truncated__ "(Wa
nt to get this briefing by email? Here' s the sign-up.)After a ceremonial opening last week for the third pr"| __truncated__ ...
 $ Date   : int  20200121 20200121 20200121 20200121 20200121 20200121 20200121 20200121 20200121 20200122 ...
 $ Article: chr  "The Coronavirus: What Scientists Have Learned So Far A novel respiratory virus that originated in Wuhan, China,"| __tr
uncated__ "First Patient With Wuhan Coronavirus Is Identified in the U.S. A man in Washington State is infected with the W"| __truncated
__ "The Test a Deadly Coronavirus Outbreak Poses to China' s Leadership WUHAN, China — Facing growing pressure to c"| __truncated__ "Imp
eachment Trial, Davos, Coronavirus: Your Tuesday Briefing (Want to get this briefing by email? Here' s the s"| __truncated__ ...
```

str(data) data
객체정보를 표시한다.

A matrix: 6 × 1 of type chr

	1930
Month	5
Day	19
Title	These N.Y.C. Neighborhoods Have the Highest Rates of Virus Deaths
Text	New data released Monday sheds light on one of the biggest questions about the toll the coronavirus has taken on New York: Where are people dying?The data, which shows death rates in each of the city's ZIP codes, underscores the deep disparities already unearthed by the outbreak. While the majority of the deaths across the city have been older residents, race and income have proven to be the largest factors in determining who lives and who dies.Neighborhoods with high concentrations of black and Latino people, as well as low-income residents, suffered the highest death rates, while some wealthier areas — primarily in Manhattan — saw almost no deaths, according to the new data, which was published by the New York City Health Department.The findings reinforced earlier reports showing that black and Latino New Yorkers were dying at twice the rate of white residents when the data is adjusted for age. Across the United States, the virus has infected and killed black people at disproportionately high rates."We may all be in the same storm, but we're not all in the same boat," said Inez Barron, a city councilwoman whose Brooklyn district includes the ZIP code with the highest death rate in the city.The data, which was current as of Monday, includes only deaths of people who had tested positive for Covid-19, the disease caused by the novel coronavirus. Probable cases of the virus among those who had not been tested account for 1 in 4 deaths.The death rate in the 11239 ZIP code — a community of about 13,000 people — is the city's highest, and almost 40 percent higher than in the area with the next highest rate. It is home to many older and African-American residents and includes Starrett City, a sprawling low- and middle-income housing complex on a peninsula jutting into Jamaica Bay.Although the area has the city's highest concentration of people over age 65, it was unclear why its death rate is so high. The total number of confirmed deaths there was 76.Ms. Barron, the city councilwoman, said the people she represents have long been underserved by the city and live in conditions that make it difficult to control the spread of the disease."We might have instances of multigenerational families in Starrett City, and one person who is sick doesn't have the luxury of going out to Long Island or going to their vacation home," she said.While the vast majority of the city's deaths have been people 65 and older, the overwhelming difference between the neighborhoods that suffered most and least has been race and income, not age.Of the 10 ZIP codes with the highest death rates, eight have populations that are predominantly black or Hispanic and include every borough except for Manhattan.Most of the neighborhoods with the lowest death rates are in Manhattan, and each has a six-figure median household income. The group also includes some of the richest ZIP codes in the city, the same areas that emptied out when the virus hit New York. All but one is majority white.The neighborhoods in the bottom quarter for death rates have double the income of the group in the top quarter. On average, the most affected areas are also more populous.The Bronx has the highest rate for coronavirus cases, hospitalizations and deaths. And in each measure, Manhattan has been the least affected.The three whitest ZIP codes in the Bronx — around Pelham-Throgs Neck and the Northeast Bronx — show among the lowest death rates in the borough.The same trends with race and income can be seen in Manhattan. A ZIP code stretching over Central Harlem and Morningside Heights had the borough's highest death rate; the neighborhoods are 90 percent black and Hispanic and one of the poorest areas in Manhattan.As stark as the disparities appear to be, it is possible that the reality is much worse.The city's ZIP code data includes only cases in which a person tested positive for Covid-19, meaning the same poor neighborhoods where testing has been lagging may have had even more deaths from the virus than the current statistics reflect. The city has been counting "probable" deaths from the virus but has not released those numbers by ZIP code.Across the city, a median of 6 percent of residents have been tested for the virus.In the ZIP codes with the highest rates of death, a median of 38 percent of the tests came back positive; in the areas with the lowest rates, about 25 percent came back positive — suggesting that if more tests were done, the death rates in the hardest-hit areas could be even higher.The rate of deaths at public housing projects mirrors that of the city overall, suggesting that fears the pandemic might disproportionately affect residents in buildings operated by the New York City Housing Authority have not borne out, according to a Health Department analysis.As of last week, 943 residents of city housing projects who had tested positive for Covid-19 had died. In its analysis of death rates in NYCHA buildings, which house about 400,000 New Yorkers, the Health Department also included deaths of people presumed to have had the virus, which added another 298 cases.In all, 7,818 public housing residents have tested positive for the disease.
Date	20200519
Article	These N.Y.C. Neighborhoods Have the Highest Rates of Virus Deaths New data released Monday sheds light on one of the biggest questions about the toll the coronavirus has taken on New York: Where are people dying?The data, which shows death rates in each of the city's ZIP codes, underscores the deep disparities already unearthed by the outbreak. While the majority of the deaths across the city have been older residents, race and income have proven to be the largest factors in determining who lives and who dies.Neighborhoods with high concentrations of black and Latino people, as well as low-income residents, suffered the highest death rates, while some wealthier areas — primarily in Manhattan — saw almost no deaths, according to the new data, which was published by the New York City Health Department.The findings reinforced earlier reports showing that black and Latino New Yorkers were dying at twice the rate of white residents when the data is adjusted for age. Across the United States, the virus has infected and killed black people at disproportionately high rates."We may all be in the same storm, but we're not all in the same boat," said Inez Barron, a city councilwoman whose Brooklyn district includes the ZIP code with the highest death rate in the city.The data, which was current as of Monday, includes only deaths of people who had tested positive for Covid-19, the disease caused by the novel coronavirus. Probable cases of the virus among those who had not been tested account for 1 in 4 deaths.The death rate in the 11239 ZIP code — a community of about 13,000 people — is the city's highest, and almost 40 percent higher than in the area with the next highest rate. It is home to many older and African-American residents and includes Starrett City, a sprawling low- and middle-income housing complex on a peninsula jutting into Jamaica Bay.Although the area has the city's highest concentration of people over age 65, it was unclear why its death rate is so high. The total number of confirmed deaths there was 76.Ms. Barron, the city councilwoman, said the people she represents have long been underserved by the city and live in conditions that make it difficult to control the spread of the disease."We might have instances of multigenerational families in Starrett City, and one person who is sick doesn't have the luxury of going out to Long Island or going to their vacation home," she said.While the vast majority of the city's deaths have been people 65 and older, the overwhelming difference between the neighborhoods that suffered most and least has been race and income, not age.Of the 10 ZIP codes with the highest death rates, eight have populations that are predominantly black or Hispanic and include every borough except for Manhattan.Most of the neighborhoods with the lowest death rates are in Manhattan, and each has a six-figure median household income. The group also includes some of the richest ZIP codes in the city, the same areas that emptied out when the virus hit New York. All but one is majority white.The neighborhoods in the bottom quarter for death rates have double the income of the group in the top quarter. On average, the most affected areas are also more populous.The Bronx has the highest rate for coronavirus cases, hospitalizations and deaths. And in each measure, Manhattan has been the least affected.The three whitest ZIP codes in the Bronx — around Pelham-Throgs Neck and the Northeast Bronx — show among the lowest death rates in the borough.The same trends with race and income can be seen in Manhattan. A ZIP code stretching over Central Harlem and Morningside Heights had the borough's highest death rate; the neighborhoods are 90 percent black and Hispanic and one of the poorest areas in Manhattan.As stark as the disparities appear to be, it is possible that the reality is much worse.The city's ZIP code data includes only cases in which a person tested positive for Covid-19, meaning the same poor neighborhoods where testing has been lagging may have had even more deaths from the virus than the current statistics reflect. The city has been counting "probable" deaths from the virus but has not released those numbers by ZIP code.Across the city, a median of 6 percent of residents have been tested for the virus.In the ZIP codes with the highest rates of death, a median of 38 percent of the tests came back positive; in the areas with the lowest rates, about 25 percent came back positive — suggesting that if more tests were done, the death rates in the hardest-hit areas could be even higher.The rate of deaths at public housing projects mirrors that of the city overall, suggesting that fears the pandemic might disproportionately affect residents in buildings operated by the New York City Housing Authority have not borne out, according to a Health Department analysis.As of last week, 943 residents of city housing projects who had tested positive for Covid-19 had died. In its analysis of death rates in NYCHA buildings, which house about 400,000 New Yorkers, the Health Department also included deaths of people presumed to have had the virus, which added another 298 cases.In all, 7,818 public housing residents have tested positive for the disease.

data %>% tail(1) %>% t()
출력결과가 길어지므로 data의 맨 마지막(tail(1) 함수) 열 만을 transpose(t() 함수)하여 나타내었다.

아래부터는 불러온 data의 전처리가 본격적으로 시작되는 과정이다. 전반적으로 다음 순서로 진행된다.

1. 모든 문자 소문자화 : lower
2. 의미 없는 단어, 문자, 문장부호 등 제거
    (1) 문장부호(punctuation), 수(digit) 제거 : del_punct
    (2) 불용어(stopwords) 제거 : del_stopword
    (3) 단일 문자(one-letter) 제거 : del_oneletter
    (4) 공백(space) 제거 : del_space
3. 표제어 추출(lemmatization) : lem

따라서 다음과 같이 순차적으로 변수명이 바뀌며, 앞선 과정이 누적되어 적용된 결과를 확인할 수 있다.

article → lower → del_punct → del_stopword → del_oneletter → del_sapce → lem

(1) make all words lowercase

In [11]:
```
## SET uppercase letters to lowercase
lower <- sapply(article, tolower, USE.NAMES = FALSE)
```

In [12]:
```
## SHOW an example before lower vs. after lower
cbind(article, lower) %>% head(1) %>% substr(1,1000)
```

A matrix: 1 × 2 of type chr

article	lower
The Coronavirus: What Scientists Have Learned So Far A novel respiratory virus that originated in Wuhan, China, last December has spread to six continents. Hundreds of thousands have been infected, at least 20,000 people have died and the spread of the coronavirus was called a pandemic by the World Health Organization in March. Much remains unknown about the virus, including how many people may have very mild or asymptomatic infections, and whether they can transmit the virus. The precise dimensions of the outbreak are hard to know. Here's what scientists have learned so far about the virus and the outbreak.Coronaviruses are named for the spikes that protrude from their surfaces, resembling a crown or the sun's corona. They can infect both animals and people, and can cause illnesses of the respiratory tract. At least four types of coronaviruses cause very mild infections every year, like the common cold. Most people get infected with one or more of these viruses at some point in their	the coronavirus: what scientists have learned so far a novel respiratory virus that originated in wuhan, china, last december has spread to six continents. hundreds of thousands have been infected, at least 20,000 people have died and the spread of the coronavirus was called a pandemic by the world health organization in march. much remains unknown about the virus, including how many people may have very mild or asymptomatic infections, and whether they can transmit the virus. the precise dimensions of the outbreak are hard to know. here's what scientists have learned so far about the virus and the outbreak.coronaviruses are named for the spikes that protrude from their surfaces, resembling a crown or the sun's corona. they can infect both animals and people, and can cause illnesses of the respiratory tract. at least four types of coronaviruses cause very mild infections every year, like the common cold. most people get infected with one or more of these viruses at some point in their

article → lower

좌(article) 우(lower)를 비교해보면 모든 대문자가 소문자로 변경됨을 확인할 수 있다.

(2) remove meaningless words in article

In [13]:
```
## REPLACE punctuations (, -.) and numbers WITH space
del_punct <- sapply(lower,
                    function(x) as.character(gsub('[[:punct:][:digit:]]', ' ', x)),
                    USE.NAMES = FALSE)

## DEFINE mystopwords to be deleted
## stopwords() in 'tm' package
mystopwords <- c(tm::stopwords('en'),
            'will', 'can', 'may',
            'also', 'still', 'yet',
            'much', 'get', 'say',
            'one', 'two', 'go')
## can add other stopwords in the 'mystopwords' vector

## DELETE 'mystopwords'
## removeWord() in 'tm' package
del_stopword <- sapply(del_punct,
                       function(x) tm::removeWords(x, mystopwords),
                       USE.NAMES = FALSE)

## DELETE one-letter words (like a, s ...)
del_oneletter <- sapply(del_stopword,
                        function(x) as.character(gsub('\\b[a-z]{1}\\b', '', x)),
                        USE.NAMES = FALSE)

## DELETE 2 or more spaces
del_space <- sapply(del_oneletter,
                    function(x) as.character(gsub('\\s+', ' ', x)),
                    USE.NAMES = FALSE)
```

lower → del_punct

[:punct:]는 모든 문장부호(punctuations), [:digit:]은 모든 수를 뜻하고, 이를 공백(‘ ’)으로 치환한다. 여기에서 생긴 공백은 뒤에서 제거할 예정이다.

del_punct → del_stopword

stopword는 문장내에 자주 등장하지만 분석하는데 큰 의미가 없는 단어를 뜻한다. stopwords는 tm 패키지에서 불러올 수 있다. 여기에 몇 가지 단어를 추가하여 mystopword를 만들어 이에 해당하는 단어를 removeWords() 함수를 적용해 del_punct에서 제거해준다.

del_stopword → del_oneletter

\\b는 경계를 나타내고, [a-z]{1}는 알파벳 소문자 a부터 z까지 중 하나를 의미한다. 앞 뒤 공백인 단일문자를 공백(‘ ’)으로 치환한다. 여기에서 생긴 공백은 다음과정에서 제거된다.

del_oneletter → del_sapce

\\s+ 는 공백(whitespace)이 하나 2개 이상(+)을 의미하고, 이를 단일 공백(‘ ‘)으로 바꾸어 단어 사이 공백을 하나로 바꾼다.

In [14]:
```
## COMPUTE the number of characters in each step
cbind(lower, del_punct,del_stopword, del_oneletter, del_space) %>% nchar() %>% colSums()

## SHOW an example of each step (cumulatively applied)
cbind(lower, del_punct,del_stopword, del_oneletter, del_space) %>% head(1) %>% substr(1,700)
```

lower: 16132840 **del_punct:** 16132840 **del_stopword:** 12719952 **del_oneletter:** 12673394 **del_space:** 10971987

A matrix: 1 × 5 of type chr

lower	del_punct	del_stopword	del_oneletter	del_space
the coronavirus: what scientists have learned so far a novel respiratory virus that originated in wuhan, china, last december has spread to six continents. hundreds of thousands have been infected, at least 20,000 people have died and the spread of the coronavirus was called a pandemic by the world health organization in march. much remains unknown about the virus, including how many people may have very mild or asymptomatic infections, and whether they can transmit the virus. the precise dimensions of the outbreak are hard to know. here's what scientists have learned so far about the virus and the outbreak.coronaviruses are named for the spikes that protrude from their surfaces, resembling	the coronavirus what scientists have learned so far a novel respiratory virus that originated in wuhan china last december has spread to six continents hundreds of thousands have been infected at least people have died and the spread of the coronavirus was called a pandemic by the world health organization in march much remains unknown about the virus including how many people may have very mild or asymptomatic infections and whether they can transmit the virus the precise dimensions of the outbreak are hard to know here s what scientists have learned so far about the virus and the outbreak coronaviruses are named for the spikes that protrude from their surfaces resembling	coronavirus scientists learned far novel respiratory virus originated wuhan china last december spread six continents hundreds thousands infected least people died spread coronavirus called pandemic world health organization march remains unknown virus including many people mild asymptomatic infections whether transmit virus precise dimensions outbreak hard know s scientists learned far virus outbreak coronaviruses named spikes protrude surfaces resembling crown sun s corona infect animals people cause illnesses respiratory tract least four types coronaviruses cause mild infections every year like comm	coronavirus scientists learned far novel respiratory virus originated wuhan china last december spread six continents hundreds thousands infected least people died spread coronavirus called pandemic world health organization march remains unknown virus including many people mild asymptomatic infections whether transmit virus precise dimensions outbreak hard know scientists learned far virus outbreak coronaviruses named spikes protrude surfaces resembling crown sun corona infect animals people cause illnesses respiratory tract least four types coronaviruses cause mild infections every year like common	coronavirus scientists learned far novel respiratory virus originated wuhan china last december spread six continents hundreds thousands infected least people died spread coronavirus called pandemic world health organization march remains unknown virus including many people mild asymptomatic infections whether transmit virus precise dimensions outbreak hard know scientists learned far virus outbreak coronaviruses named spikes protrude surfaces resembling crown sun corona infect animals people cause illnesses respiratory tract least four types coronaviruses cause mild infections every year like common cold people infected viruses point lives another coronavirus circulated china caused danger

nchar() 함수로 계산한 문자 수가 점점 줄어듦을 확인할 수 있다.
(16,132,840) → (16,132,840) → (12,719,952) → (12,673,394) → (10,971,987)
그 아래는 순차적으로 전처리 되어가는 article 일부를 보여주고 있다.

(3) lemmatize and change data type

In [15]:
```
## LEMMANTIZE articles
## lemmatize_strings in 'textstem' package
lem <- sapply(del_space,
              textstem::lemmatize_strings,
              USE.NAMES = FALSE)

## SET data AS dataframe
words <- lem %>% as.data.frame()

## SET data label
colnames(words) <- 'Word'

## SET data type of data in dataframe AS character
words$Word <- words$Word %>% as.character()
```

In [16]:
```
## COMPUTE the number of characters in each step
cbind(del_space, lem) %>% nchar() %>% colSums()

## SHOW an example before lem vs. after lem
cbind(del_space, lem) %>% head(1) %>% substr(1,1000)
```

del_space: 10971987 **lem:** 10272403

del_sapce → lem : textstem 패키지의 lemmentize_strings 함수를 적용하여 표제어를 추출한다.

(10,971,987) → (10,272,403)으로 문자의 수가 줄었다. 즉, 하나의 표제어로 통합되면서 다른 형태의 동일 의미의 단어 길이가 줄었다.

A matrix: 1 × 2 of type chr

del_space	lem
coronavirus scientists learned far novel respiratory virus originated wuhan china last december spread six continents hundreds thousands infected least people died spread coronavirus called pandemic world health organization march remains unknown virus including many people mild asymptomatic infections whether transmit virus precise dimensions outbreak hard know scientists learned far virus outbreak coronaviruses named spikes protrude surfaces resembling crown sun corona infect animals people cause illnesses respiratory tract least four types coronaviruses cause mild infections every year like common cold people infected viruses point lives another coronavirus circulated china caused dangerous condition known severe acute respiratory syndrome sars virus contained sickened people killed middle east respiratory syndrome mers first reported saudi arabia caused coronavirus new virus named sars cov disease causes called covid hard accurately assess lethality new virus appears less often fa	coronavirus scientist learn far novel respiratory virus originate wuhan china last december spread six continent hundred thousand infect less people die spread coronavirus call pandemic world health organization march remain unknown virus include many people mild asymptomatic infection whether transmit virus precise dimension outbreak hard know scientist learn far virus outbreak coronavirus name spike protrude surface resemble crown sun corona infect animal people cause illness respiratory tract less four type coronavirus cause mild infection every year like common cold people infect virus point life another coronavirus circulate china cause dangerous condition know severe acute respiratory syndrome sars virus contain sicken people kill middle east respiratory syndrome mers first report saudi arabia cause coronavirus new virus name sars cov disease cause call covid hard accurately assess lethality new virus appear little often fatal coronavirus cause sars mers significantly seasonal fl

좌(del_space) 우(lem)를 비교해보면, 첫 줄의 learned → learn, originated → originate 와 같이 기본형의 단어로 바뀌었다.

(4) append 'words' to dataset (and save 'words' separately if needed)

In [17]:
```
## MERGE 'data', 'words' dataframes, THEN GET new dataframe ''
dataset <- cbind(data, words)

## WRITE the dataframes AS csv files (you can skip)
# write.csv(words, 'words.csv', row.names = FALSE)
# write.csv(dataset, 'dataset.csv', row.names = FALSE)
```

다음 단계에서 쓰일 dataset을 묶어서 저장해 둔다.

3.3 Get word sets

3. Get unique wordset by month

(0) load dataset (can skip)

In [18]:
```
## READ csv data file, THEN GET same dataframes as above'dataset'
# dataset <- read.csv('dataset.csv')

## SET data AS character
# dataset$Word <- dataset$Word %>% as.character()
```

(1) get unique wordset

In [19]:
```
## ADD a new column to the dataset
dataset$unique_words <- NA

## FOR each row in rows of dataset
## SPLIT 'Word' data according to spaces THEN GET only unique words
## ADD the unique words TO dataset
## ENDFOR
for (i in 1:nrow(dataset)) {
  row_words <- dataset$Word[i]
  listed_unique_words <- lapply(strsplit(row_words, ' '), unique)
  unique_words <- unlist(listed_unique_words) %>% paste(collapse = ' ')
  dataset$unique_words[i] <- unique_words
}
```

In [20]:
```
## COMPUTE the number of characters to compare Word with unique_words
dataset %>% dplyr::select(Word,unique_words) %>% nchar()
```

Word: 10282037 **unique_words:** 5881087

처음에 두 개의 raw data를 가져오면서 중복된 article을 수집했을 가능성이 있다. 따라서 중복 방지를 위해 dataset에 빈 열의 unique_words를 추가하고, for loop를 돌면서, unique한 단어를 unique_words에 각각 입력한다. dataset의 행 단위로 탐색하면 중복을 방지할 수 있다.

(2) combine data by month (Jan ~ May)

In [21]:
```
## GROUP dataset BY Month
## THEN MERGE unique_words
## THEN GET dataframe'dataset_monthly'
dataset_monthly <- dataset %>%
    dplyr::group_by(Month) %>%
    dplyr::summarise(wordset = paste(unique_words, collapse = ' '))
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

dataset을 월별 data로 나누고 unique_words를 wordset으로 하여 단어별 개수로 정리한 다음 dataset_monthly에 입력한다. dplyr 패키지의 summarise()는 단어를 집계하는 함수로 쓰이고, group_by()는 입력받은 dataset에서 $Month를 기준으로 그룹화하는 함수로 쓰인다.

In [22]:
```
## SHOW dataset info
dataset_monthly %>% str()
## COMPUTE
dataset_monthly$wordset %>% t() %>% nchar() #each month has this amount of words
```

```
tibble [5 × 2] (S3: tbl_df/tbl/data.frame)
 $ Month  : int [1:5] 1 2 3 4 5
 $ wordset: chr [1:5] "coronavirus scientist learn far novel respiratory virus originate wuhan china last december spread six continen"| __truncated__ "new york city eye first suspect case coronavirus health official announce saturday patient bellevue hospital ce"| __truncated__ "work home coronavirus affect workplace furlough sick leave experience measure business try prevent employee exp"| __truncated__ "approve first coronavirus antibody test food drug administration thursday new use unite state currently availab"| __truncated__ ...
```

A matrix: 1 × 5 of type int

311498	1016027	1607159	1771955	1166745

dataset_monthly의 정보를 확인한다. nchar()로 계산한 월별 dataset의 단어 길이 수는 1월부터 5월까지 각각 311,498(1월), 1,016,027(2월), 1,607,159(3월), 1,771,955(4월), 1,166,745(5월)이다. 1월은 21일부터 모인 기사라 단어 길이 수가 적다.

(3) divide data sets by month and save as txt file

In [23]:
```
## WRITE the dataframes AS csv files (you can skip)
# write.csv(dataset_monthly, 'monthly_words.csv', row.names = F)
## it takes long time since a lot of words are in the dataset

## SPLIT the dataset INTO 5 dataframes (Jan~May)
## corresponding each row (for easier loading)
Jan <- dataset_monthly$wordset[1]
Feb <- dataset_monthly$wordset[2]
Mar <- dataset_monthly$wordset[3]
Apr <- dataset_monthly$wordset[4]
May <- dataset_monthly$wordset[5]

## WRITE txt files
write(Jan, 'Jan.txt')
write(Feb, 'Feb.txt')
write(Mar, 'Mar.txt')
write(Apr, 'Apr.txt')
write(May, 'May.txt')
```

dataset_monthly의 wordset을 각각 Jan, Feb, Mar, Apr, May로 나누어 선언하고, txt파일로 저장해둔다.

3.4 Analysis

4. Analysis

(0) load dataset (can skip)

In [24]:
```
## READ txt data file, THEN GET same dataset as above (Jan~May)
# Jan <- readLines('Jan.txt') %>% substr(1,35000)
# Feb <- readLines('Feb.txt') %>% substr(1,35000)
# Mar <- readLines('Mar.txt') %>% substr(1,35000)
# Apr <- readLines('Apr.txt') %>% substr(1,35000)
# May <- readLines('May.txt') %>% substr(1,35000)
```

(1) extract nouns

In [25]:
```
## EXTRACT noun from datasets (Jan~May), THEN GET (noun1~noun5)
noun1 <- sapply(Jan, extractNoun, USE.NAMES = F)
noun2 <- sapply(Feb, extractNoun, USE.NAMES = F)
noun3 <- sapply(Mar, extractNoun, USE.NAMES = F)
noun4 <- sapply(Apr, extractNoun, USE.NAMES = F)
noun5 <- sapply(May, extractNoun, USE.NAMES = F)
## extractNoun in 'KoNLP' package
```

Jan~May를 KoNLP 패키지의 extractNoun함수를 적용하여 단어를 추출해내고, 이를 noun1~noun5로 지정한다.

(2) remove unnecessary words

In [26]:
```
## FOR noun in (noun1~noun5)
## REPLACE unnecessary words in noun WITH N/As
## REMOVE N/As in noun
## ENDFOR

## N01
noun1 <- noun1 %>% gsub('people',NA,.)
noun1 <- noun1 %>% gsub('say',NA,.)
noun1 <- noun1 %>% gsub('one',NA,.)
noun1 <- noun1 %>% gsub('also',NA,.)
noun1 <- noun1 %>% gsub('will',NA,.)
noun1 <- noun1 %>% gsub('day',NA,.)
noun1 <- noun1 %>% gsub('make',NA,.)
noun1 <- noun1 %>% gsub('take',NA,.)
noun1 <- noun1 %>% gsub('two',NA,.)
noun1 <- noun1 %>% gsub('like',NA,.)
noun1 <- noun1 %>% gsub('case',NA,.)
noun1 <- noun1 %>% gsub('go',NA,.)
noun1 <- noun1 %>% gsub('can',NA,.)

## N02
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no
noun2 <- no

## N03
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no
noun3 <- no

## N04
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no
noun4 <- no

## N05
noun5 <- noun5 %>% gsub('people',NA,.)
noun5 <- noun5 %>% gsub('say',NA,.)
noun5 <- noun5 %>% gsub('one',NA,.)
noun5 <- noun5 %>% gsub('also',NA,.)
noun5 <- noun5 %>% gsub('will',NA,.)
noun5 <- noun5 %>% gsub('day',NA,.)
noun5 <- noun5 %>% gsub('make',NA,.)
noun5 <- noun5 %>% gsub('take',NA,.)
noun5 <- noun5 %>% gsub('two',NA,.)
noun5 <- noun5 %>% gsub('like',NA,.)
noun5 <- noun5 %>% gsub('case',NA,.)
noun5 <- noun5 %>% gsub('go',NA,.)
noun5 <- noun5 %>% gsub('can',NA,.)
```

gsub() 함수로 전처리에서 처리되지 않은 단어를 재처리한다.
noun1~noun5 동일한 과정을 거친다.

```
noun1 <- na.omit(noun1)
noun2 <- na.omit(noun2)
noun3 <- na.omit(noun3)
noun4 <- na.omit(noun4)
noun5 <- na.omit(noun5)
```

gsub()으로 인해 NA로 대체된 것을 na.omit()으로 제거한다.

(3) count nouns

In [27]:
```
## GET tables to count the nouns
count_jan <- table(noun1) %>% sort(decreasing = T)
count_feb <- table(noun2) %>% sort(decreasing = T)
count_mar <- table(noun3) %>% sort(decreasing = T)
count_apr <- table(noun4) %>% sort(decreasing = T)
count_may <- table(noun5) %>% sort(decreasing = T)
```

각 noun을 table로 변경한(table() 함수) 다음 내림차순(decreasing=T)으로 정렬(sort() 함수)한다.
각 table은 count_jan, count_feb, count_mar, count_apr, count_may이다.

In [28]:
```
## Word count for the entire period, Top 100
count_table <- cbind(count_jan, count_feb, count_mar, count_apr, count_may)
count_table %>% head()
count_all <- rowSums(count_table) %>% as.table()
rowSums(count_table) %>% head(100)
```

```
Warning message in cbind(count_jan, count_feb, count_mar, count_apr, count_may):
 "number of rows of result is not a multiple of vector length (arg 1)"
```

A matrix: 6 × 5 of type int

	count_jan	count_feb	count_mar	count_apr	count_may
coronavirus	11	12	14	14	12
early	11	12	13	13	11
test	11	12	12	12	11
virus	11	11	12	12	11
begin	11	11	12	12	10
dr	11	11	12	11	10

coronavirus: 63 **early:** 60 **test:** 58 **virus:** 57 **begin:** 56 **dr:** 55 **new:** 54 **week:** 53 **call:** 51 **covid:** 51 **director:** 49 **even:** 49 **find:** 48 **include:** 47 **infection:** 47 **know:** 47 **look:** 47 **many:** 47 **medical:** 47 **time:** 47 **use:** 47 **work:** 46 **world:** 46 **around:** 45 **become:** 44 **cause:** 43 **change:** 43 **come:** 43 **country:** 43 **disease:** 43 **home:** 43 **pandemic:** 43 **part:** 43 **provide:** 43 **school:** 42 **university:** 41 **way:** 41 **care:** 39 **china:** 39 **control:** 39 **don:** 39 **end:** 39 **every:** 39 **get:** 38 **health:** 38 **infect:** 38 **just:** 38 **may:** 38 **medicine:** 38 **now:** 38 **numb:** 38 **outbreak:** 38 **public:** 38 **re:** 38 **really:** 36 **see:** 36 **spread:** 36 **start:** 35 **think:** 35 **three:** 35 **without:** 35 **year:** 35 **another:** 34 **april:** 34 **develop:** 34 **effect:** 34 **emergency:** 34 **enough:** 34 **expert:** 34 **first:** 34 **follow:** 33 **give:** 33 **grow:** 33 **have:** 33 **help:** 33 **history:** 33 **ill:** 33 **life:** 33 **live:** 33 **long:** 33 **month:** 33 **need:** 33 **news:** 32 **official:** 31 **other:** 31 **patient:** 31 **realize:** 31 **researcher:** 31 **result:** 31 **return:** 30 **run:** 30 **science:** 30 **scientist:** 30 **six:** 30 **state:** 30 **study:** 30 **track:** 30 **treatment:** 30 **add:** 29 **africa:** 29

count_jan ~ count_may를 cbind()로 열방향으로 결합한 count_table을 생성한다. 위 결과는 count_table의 head() 일부와 각 열을 합산한 count_all에서 top 100에 해당하는 단어와 그 개수를 나열한 것이다.

(4) wordcloud of each month

In [29]:
```
## SET variable 'palette' for argument 'colors'
## brewer.pal() in 'RColorBrewer' package
palette <- brewer.pal(8, 'Set2')

## GET monthly wordclouds
## wordcloud2() in 'wordcloud2' package

## 1(Jan)
wordcloud2(data = count_jan, size=0.7, color = palette)

## 2(Feb)
wordcloud2(data = count_feb, size=0.7, color = palette)

## 3(Mar)
wordcloud2(data = count_mar, size=0.7, color = palette)

## 4(Apr)
wordcloud2(data = count_apr, size=0.7, color = palette)

## 5(May)
wordcloud2(data = count_may, size=0.7, color = palette)

## ALL(Jun~May)
wordcloud2(data = count_all, size=0.7, color = palette)
```

RColorBrewer 패키지에서 brewer.par() 함수로 word cloud에 쓰일 색상 팔레트 'palette'를 만든다. wordcloud2 패키지의 wordcloud2() 함수로 count table을 word cloud로 만든다. 결과는 4. Conclusion에서 확인할 수 있다.

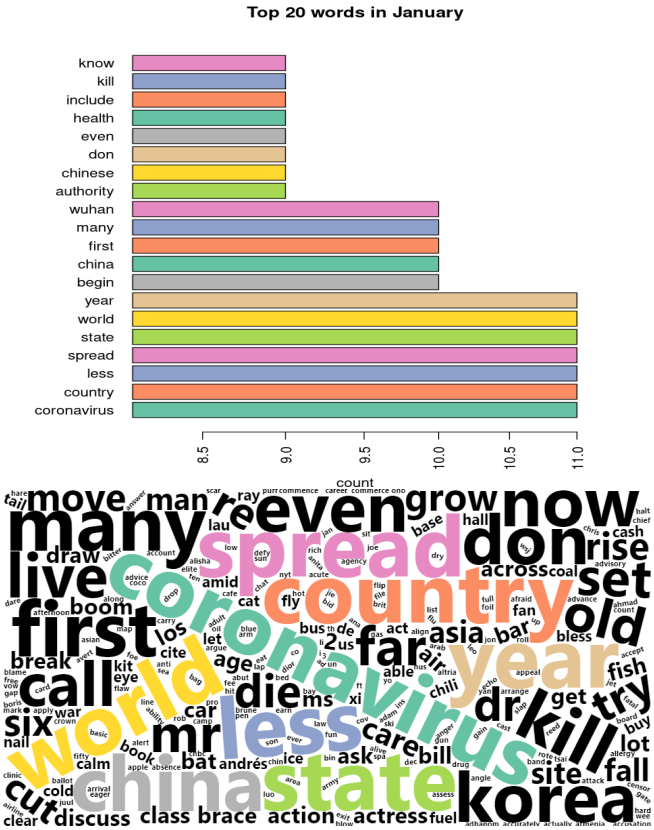
(5) top 20 words in month

In [30]:
```
## GET monthly barplot, top 20 words
par(mar=c(4,7,4,4))
barplot(count_jan[1:20], col=palette, las=2, log="x", horiz=T, xlab = "count", main = "Top 20 words in January")
barplot(count_feb[1:20], col=palette, las=2, log="x", horiz=T, xlab = "count", main = "Top 20 words in February")
barplot(count_mar[1:20], col=palette, las=2, log="x", horiz=T, xlab = "count", main = "Top 20 words in March")
barplot(count_apr[1:20], col=palette, las=2, log="x", horiz=T, xlab = "count", main = "Top 20 words in April")
barplot(count_may[1:20], col=palette, las=2, log="x", horiz=T, xlab = "count", main = "Top 20 words in May")
barplot(count_all[1:20], col=palette, las=2, log="x", horiz=T, xlab = "count", main = "Top 20 words")
```

각 월의 top 20 단어를 bar plot으로 확인해본다. 결과는 4. Conclusion에서 확인할 수 있다.

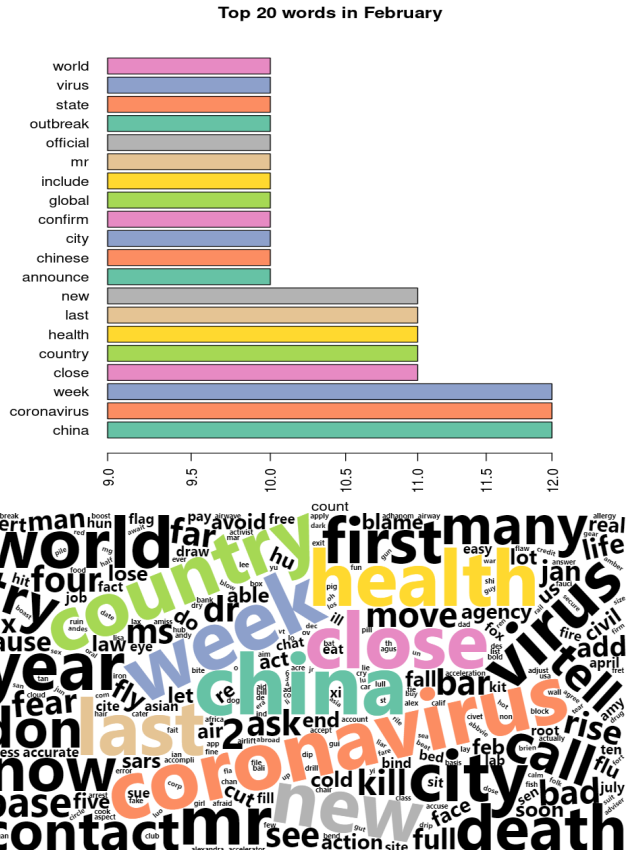
4. Conclusion

1월의 word cloud 및 bar plot 해석



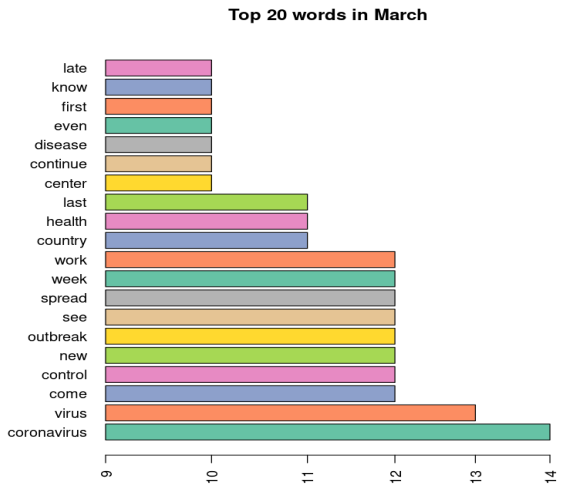
막 중국 우한에서 발병이 시작하기 시작한 때이다. 아직 세계적 pandemic이 될 거라는 예측은 없이, 가볍게 COVID-19를 다루고 있다. 중국' 과 '발병'(outbreak)이 주로 등장하고 'begin' 등 발병의 시작 등을 주로 나타내는 단어들이 등장하고 있다. 그와 함께 'infect', 'disease' 등 전염병 관련 단어들이 등장한다. 1/23일 우한 도시가 봉쇄되었다. 그와 관련된 주 단어 'city', quthority', 'government' 등도 보이기 시작한다.

2월의 word cloud 및 bar plot 해석



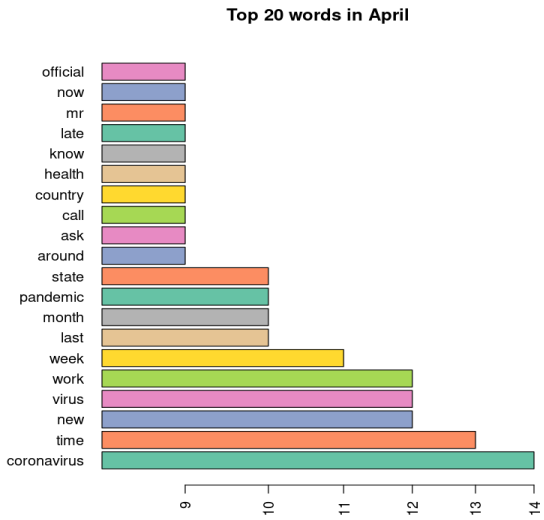
여전히 'china'는 주된 관심사이다. 중국 내에서 COVID-19가 급격히 유행하고 감염자가 급격히 늘어나는 시기가 1월말-2월초중순이다. 정부가 우한을 봉쇄하였기 때문에, 여전히 그와 관련된 'wuhan', 'government', 'control', 'city' 단어들이 보이고, 코로나 감염자가 전세계적으로 하나 둘 등장하고 있어서 'quarantine'(자가격리) 이라는 용어도 등장하기 시작한다. 'spread' 란 단어는 전세계적으로 퍼지고 있는 상황을 알려주고 있다.

3월의 word cloud 및 bar plot 해석

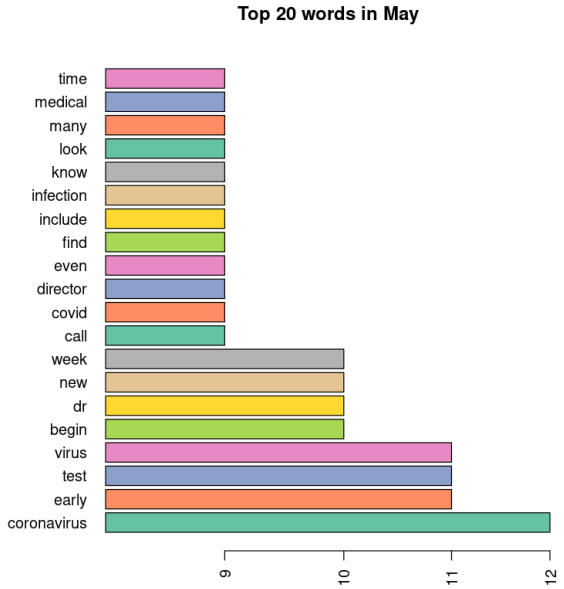


본격적으로 전세계에서 대 유행하기 시작한다는 것이 보인다. 높아진 'world', 'country', 'public', 'state', 'spread' 단어들의 빈도수와 빈도수가 확연히 줄어든 'china', 'chinese' 단어에서 엿볼 수 있다. 건강에 대한 관심이 늘어나면서 'health', 'help' 등이 등장하고, 도시, 국가 봉쇄가 연달아 일어나면서 'close', 'government' 등이 등장하고 있다. 장기적 상황이 예상되면서 단순 권고 조치인 'quarantine'(자가격리)였던 지난달과 달리, 외출을 자제시키는 'home' 이란 단어가 등장하기 시작하였고 같은 맥락의 'long'이라는 단어도 등장하였다. 외출 자제는 경제상황과 직결되기 때문에 그 상황이 아주 심각함을 반영한다.

4월의 word cloud 및 bar plot 해석



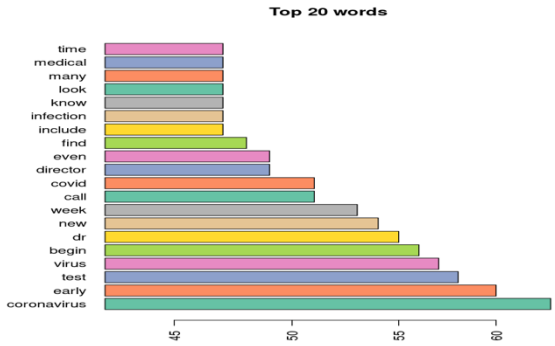
여전히 'china','chinese'에 관한 단어는 등장하지 않고 전세계적 감염에 대한 인식을 보여주듯 'world'란 단어가 주로 드러난다. 특히, 전세계적으로 상황이 악화되었다는 것을 'pandemic'을 통해서 알 수 있다. 4월에는 미국에서도 covid-19 유행이 악화되었기 때문에, 미국 자국 국민들에게 있어서 더 영향이 컸다는 것을 알 수 있다. 'health', 'test', 'care'등이 자주 등장하였음을 통해서 그 사실을 유추해 볼 수 있다.



5월의 word cloud 및 bar plot 해석

4월에 pandemic 이란 단어가 등장하면서 계속해서 pandemic이란 단어가 자주 언급되고 있다. 그와 함께 'long' , 'health', 'home' 등의 COVID 와 관련된 단어가 계속해서 등장하면서 코로나 19에 대한 현재 견해는 '세계적 역병' 과 '경계 상태' 임을 알 수 있다. 단어 빈도수는 4월의 것과 비슷하다.

5 개월간의 통합 word cloud 및 bar plot



최종 결론

주목해야할 흐름은 1 월~2 월까지는 COVID-19 를 단순히 중국 및 동아시아에서 유행하는 이전에 종종 있었던 유행병과 같은 생각을 가지고 있었다는 것이다. 중국을 중심으로 주요 보도가 이루어지고 있고, 아직은 세계적 감염이라는 인식을 보이고 있지 않다. 하지만, 3 월달부터는 COVID-19 를 세계적 감염으로 바라보기 시작한다. 빈도수가 높아지는 'world' 단어와 빈도가 확연히 줄어든 'china'라는 단어에서 알 수 있다. 특히 4 월달부터는 세계적 감염의 심각성을 나타내는 'pandemic'이라는 단어의 빈도수가 특히 높아지고 그 흐름이 5 월까지 이어지는 것으로 보아 현재 뉴욕 타임스 언론사의 코로나 19 에 대한 견해는 1-2 월과는 달리 '전세계적 대 역병 및 초긴장 상태' 라는 것을 알 수 있다.