PubMed COVID-19 Dataset Processing

MIT COVID 19 Hackathon
Qiaoru Zhang

````
---
title: |
  | PubMed COVID-19 Dataset Processing
  | MIT COVID 19 Hackathon
header-includes: \usepackage{float}
output:
  word_document: default
  pdf_document:
    extra_dependencies: float
subtitle: |
  | Qiaoru Zhang
fontsize: 12pt
---

## Required library papckages
```{r}
library(tidyverse)
library(readxl)
library(writexl)
```

##  Transform the txt file into a character string
```{r}
COVdataset <- readLines("~/Desktop/abstract-COVID19sym-set.txt")
length <- length(COVdataset)
breaks <- which(! nzchar(COVdataset))
nbreaks <- length(breaks)
if (breaks[nbreaks] < length) {
  breaks <- c(breaks, length + 1L)
  nbreaks <- nbreaks + 1L
}
if (nbreaks > 0L) {
  COVdataset <- mapply(function(a,b) paste(COVdataset[a:b], collapse = " "),
                       c(1L, 1L + breaks[-nbreaks]),
                       breaks - 1L)
}
COVdataset[1]
```
````

Change the transformed file into dataframe file.

```{r}
abstract <- data.frame(COVdataset,stringsAsFactors = FALSE)
```

Count the character number for each rows

```{r}
abstract$noChar <- nchar(abstract$COVdataset)
```

Filter out the rows which have long paragraphs

```{r}
dfabstract<-as_tibble(abstract)
df<-dfabstract %>% filter(noChar >600)
```

Remove the counted number

```{r}
Abstract_PubMed<-within(df, rm(noChar))
```

Information and white space filtrate：

```{r}
COVID19_symptoms<-gsub("Author
information.*","",COVID_19_symptoms$COVdataset)
COVID19_symptoms <- COVID19_symptoms[!(COVID19_symptoms$COVdataset ==
""), ]
COVID19_symptoms<-data.frame(COVID19_symptoms)
```

Create the data file

```{r}
write_xlsx(COVID19_symptoms,"~/Desktop/MIT/COVID19_symptoms.xlsx")
```

Split a large dataframe into seprate dataframe groups by 100 rows per group.

```{r}
chunk <- 100
n <- nrow(COVID19_symptoms)
r <- rep(1:ceiling(n/chunk),each=chunk)[1:n]
COVID19_symptoms_split <- split(COVID19_symptoms,r)
r <- ggplot2::cut_width(1:n,chunk,boundary=0)
```

Create the splited data file

```{r}
write_xlsx(COVID19_symptoms_split,"~/Desktop/COVID19_symptoms_split.xlsx")
```

Save them into diferent files

```{r}
for (Abstract in unique(COVID19_symptoms_split$Abstract)) {
  write.csv(data["Abstract" == Abstract,], file = paste0("newfile_", Abstract, ".csv"))
}
```