

# Markdown\_World12\_Analyzing\_data\_QuantedaOnly

Aymeric Collart

## 1. Prepare the environment

### 1.1 Load the libraries

```
library(quanteda)

## Package version: 4.3.1
## Unicode version: 14.0
## ICU version: 71.1

## Parallel computing: disabled

## See https://quanteda.io for tutorials and examples.

library(quanteda.textstats)
library(tidytext)
library(dplyr)

## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

library(stringr)
library(openxlsx)

#Sys.setlocale(category = "LC_ALL", locale = "cht")
```

### 1.2 Load the originally scraped data

```
load(file = "ArticleETTToday_CorpusCourse_CLEAN.Rdata")
```

## 2. Key Word In Context (KWIC)

### 2.1 Prepare the dataset for the analyses

```
Article_total2$docname <- paste0("text",
                                   1:nrow(Article_total2))

Article_tokens <- tokens(Article_total2$body)
```

### 2.2 Perform the KWIC segmentation

#### 2.2.1 KWIC segmentation

```
kwic_data <- kwic(Article_tokens,
                    pattern = "有",
                    window = 30)
```

#### 2.2.2 Annotate the KWIC dataset

```
kwic_data <- as.data.frame(kwic_data)

kwic_data <- right_join(kwic_data,
                         Article_total2,
                         by = "docname")

kwic_data <- na.omit(kwic_data)
```

#### 2.2.3 (Optional) Clean the context to keep only the phrase where the keyword is found

```
## Keep original information just in case
kwic_data$pre_original <- kwic_data$pre
kwic_data$post_original <- kwic_data$post

## Post context
symbol1 <- "\\".
kwic_data$post <- sub(paste0("(", symbol1, ").*"), "\\\1", kwic_data$post)

symbol2 <- "\\", "
kwic_data$post <- sub(paste0("(", symbol2, ").*"), "\\\1", kwic_data$post)

symbol3 <- "\\\? "
```

```

kwic_data$post <- sub(paste0("(", symbol3, ").*"), "\\\\1", kwic_data$post)

symbol4 <- "\\\\"!
kwic_data$post <- sub(paste0("(", symbol4, ").*"), "\\\\1", kwic_data$post)

## Pre context
kwic_data$pre <- sub(".*. ([^*. ]*)$", ". \\\\1", kwic_data$pre)
kwic_data$pre <- sub(".*, ([^*, ]*)$", ", \\\\1", kwic_data$pre)
kwic_data$pre <- sub(".* ? ([^*?]*)$", "? \\\\1", kwic_data$pre)
kwic_data$pre <- sub(".*! ([^*!]*)$", "! \\\\1", kwic_data$pre)

## Have a look at the data
knitr::kable(head(kwic_data))

```



doftime prekey postkey titlebody url yearmonth original\_article pre\_point original

doftime prekey postkey titlebody url yearmonth original\_article pre\_point original

text#19, 有玩有2020年當神人把《我的世界》改成《血源詛咒》。(圖 / 翻  
竟然家突發奇想將兩款F戲尬在一起，

2020 01 // 摄自 YouTube / Potomy)

01 把的月《我世01 的界》(Minecrafterday.  
01 世遇10:57界》上20231231/  
改《血成源htm

《血詛  
源咒》(Blood-  
詛 borne),

咒》竟

然還

原度

超度

高突

玩家發

狂想

敵將

碗兩

快款

點F

出戲

一起在

一起，兩

大名作

融合

在一

起的

F物

不僅

看來

煞有

其事，

而且

還原

度還

記者樓菀玲 / 綜合報導當《我的世界》(Minecraft) 遇上《血源詛咒》(Bloodborne)，竟然有玩家突發奇想將兩款遊戲尬在一起，兩大名作融合在一起的物不僅看起來煞有其事，而且還原度還很高，這一切都是要歸功於玩家 Potomy 和他的團隊，他們費時 2 年才終於經典魂系遊戲融入《我的世界》的像素世界。據了解，Potomy 最早在於 12 月 14 日透過推特 (X) 分享了這項計畫的新進度，目前這項計畫已進行了約 2 年，他們試著在《我的世界》重現《血源詛咒》的歐風城市雅南 (Yharnam)，還有一些原作中無法到達的區域。

2 Years of Recreating Bloodborne in Minecraft. ..We are getting there... pic.twitter.com/0hvG9qXpIq — Potomy (@\_potomy\_) December 14, 2023 Potomy 打算在《我的世界》重現原作元素，而且還打算融入自己的新想法，像是將原作遊戲中的敵人轉化為《我的世界》的生物敵人 (Mob)，甚至連攻擊動作等細節都被精心重現，使玩家能在《我的世界》中體驗到《血源詛咒》的醍醐味。這項計畫除了 Potomy 之外，還有不少神人高手參與其中，像是彼此分工角色建模、聲音設計、動畫和影片編輯等各方面工作，某種程度也算是一個制作遊戲的小型團隊，雖然目前尚未確定推出方式和時間，但已經有不少玩家表示相當期待，究竟要怎樣才能在《我的世界》當中，感受到來自《血源詛咒》的善意。| 當《我的世界》(Minecraft) 遇上《血源詛咒》(Bloodborne)，竟然 | 玩家突發奇想將兩款遊戲尬在一起，兩大名作融合在一起的物不僅看起來煞有其事，而且還原度還很高，這一切都是要歸功於 | |text6 | 55| 55|，但已經 | 有 | 不少玩家表示相當期待，| 有 | 2024 年 01 月 01 日 10:57 | 政治 | 神人把《我的世界》改成《血源詛咒》 | 還原度超高玩家狂敲碗：快點出 | 這項計畫除了 Potomy 之外，還有不少神人高手參與其中，像是彼此分工角色建模、聲音設計、動畫和影片編輯等各方面工作，某種程度也算是一個制作遊戲的小型團隊，雖然目前尚未確定推出方式和時間，但已經有不少玩家表示相當期待，究竟要怎樣才能在《我的世界》當中，感受到來自《血源詛咒》的善意。|<https://www.ettoday.net/news/20231231/2652951.htm> | 2024 | 01 | 01 | 神人把《我的世界》改成《血源詛咒》。(圖 / 翻攝自 YouTube / Potomy) 記者樓菀玲 / 綜合報導當《我的世界》(Minecraft) 遇上《血源詛咒》(Bloodborne)，竟然有玩家突發奇想將兩款遊戲尬在一起，兩大名作融合在一起的物不僅看起來煞有其事，而且還原度還很高，這一切都是要歸功於玩家 Potomy 和他的團隊，他們費時 2 年才終於經典魂系遊戲融入《我的世界》的像素世界。據了解，Potomy 最早在於 12 月 14 日透過推特 (X) 分享了這項計畫的新進度，目前這項計畫已進行了約 2 年，他們試著在《我的世界》重現《血源詛咒》的歐風城市雅南 (Yharnam)，還有一些原作中無法到達的區域。

2 Years of Recreating Bloodborne in Minecraft. ..We are getting there... pic.twitter.com/0hvG9qXpIq — Potomy (@\_potomy\_) December 14, 2023 Potomy 打算在《我的世界》重現原作元素，而且還打算融入自己的新想法，像是將原作遊戲中的敵人轉化為《我的世界》的生物敵人 (Mob)，甚至連攻擊動作等細節都被精心重現，使玩家能在《我的世界》中體驗到《血源詛咒》的醍醐味。這項計畫除了 Potomy 之外，還有不少神人高手參與其中，像是彼此分工角色建模、聲音設計、動畫和影片編輯等各方面工作，某種程度也算是一個制作遊戲的小型團隊，雖然目前尚未確定推出方式和時間，但已經有不少玩家表示相當期待，究竟要怎樣才能在《我的世界》當中，感受到來自《血源詛咒》的善意。| 動畫和影片編輯等各方面工作，某種程度也算是一個制作遊戲的小型團隊，雖然目前尚未確定推出方式和時間，但已經 | 不少玩家表示相當期待，究竟要怎樣才能在《我的世界》當中，感受到來自《血源詛咒》的善意。| |text12 | 18| 18|，常見補品 | 有 | 燒酒 |、| 母鴨、羊肉爐、藥 | 排骨等，| 有 | 2024 年 01 月 01 日 09:07 | 社會 | 跨年冬令進補爐火需留意 安裝住宅用火警報器避免悲劇 | 民跨年迎接 2024 年，常相約一起吃火鍋或是冬令進補，常見補品有燒酒 |、| 母鴨、羊肉爐、藥 | 排骨等，由於補料理方式需要慢火熬煮，因此煮需要時間，民常會趁著空檔去做家事、睡覺甚至外出，忘記房正在烹煮食物，當鍋煮到無湯汁時，便可能發生煮食不慎火案例。|<https://www.ettoday.net/news/20231231/2652951.htm> | 2024 | 01 | 01 | 爐火使用不慎，住家遭燒 |。(圖 / 記者 | | | 摄) 記者 | | | / 台北市報導民跨年迎接 2024 年，常相約一起吃火鍋或是冬令進補，常見補品有燒酒 |、| 母鴨、羊肉爐、藥 | 排骨等，由於補料理方式需要慢火熬煮，因此煮需要時間，民常會趁著空檔去做家事、睡覺甚至外出，忘記房正在烹煮食物，當鍋煮到無湯汁時，便可能發生煮食不慎火案例。消防局提醒，使用爐火應遵守「人離火熄」原則，若有事離開房，應先關閉爐火再離開房，同時避免鍋具部湯汁液面過高，導致沸騰時湯汁外溢，熄滅爐火而造成瓦斯外引起火事故。另外，也需留意家中長時間煮食物，忘記爐火使鍋煮到無湯，食物焦化而引起火。

住家房起火，所幸及時撲滅未釀成害。(圖 / 記者 | | | 摄) | 防止房火發生，消防局呼 |，房應安裝「住宅用火警報器」，偵測到濃或高時，將發出高分貝聲響提醒民；煮時瓦斯爐具附近勿放置易燃物品或危險物品。房可選用安全爐具，選擇有熄火安全裝置及度感知功能爐具，偵測爐火熄滅，或鍋爐達到一定度時會自動切斷瓦斯，有效預防爐火烹調不慎所造成之火案例。| 民跨年迎接 2024 年，常相約一起吃火鍋或是冬令進補，常見補品 | 燒酒 |、| 母鴨、羊肉爐、藥 | 排骨等，由於補料理方式需要慢火熬煮，因此煮需要時間，民常會趁著 | |text14 | 57| 57|，選擇 | 有 | 熄火安全裝置及度感知功能爐具，| 有 | 2024 年 01 月 01 日 09:07 | 社會 | 跨年冬令進補爐火需留意 安裝住宅用火警報器避免悲劇 | | 防止房火發生，消防局呼 |，房應安裝「住宅用火警報器」，偵測到濃或高時，將

發出高分貝聲響提醒民房；煮時瓦斯爐具附近勿放置易燃物品或危險物品。房可選用安全爐具，選擇有熄火安全裝置及感應功能爐具，偵測爐火熄滅，或鍋爐達到一定溫度時會自動切斷瓦斯，有效預防爐火烹調不慎所造成之火警案件。<https://www.ettoday.net/news/20231231/2652951.htm> | 2024 | 01 | 01 | 爐火使用不慎，住家遭燒。 (圖 / 記者[REDACTED]攝) 記者[REDACTED] / 台北市報導民房跨年迎接 2024 年，常相約一起吃火鍋或是冬令進補，常見補品有燒酒、母鴨、羊肉爐、藥排骨等，由於烹料理方式需要慢火熬煮，因此煮需要時間，民房常會趁著空檔去做家事、睡覺甚至外出，忘記房正在烹煮食物，當鍋煮到無湯汁時，便可能發生煮食不慎火警案例。消防局提醒，使用爐火應遵守「人離火熄」原則，若有事離開房，應先關閉爐火再離開房，同時避免鍋具底部湯汁液面過高，導致沸騰時湯汁外溢，熄滅爐火而造成瓦斯外引起火警事故。另外，也需留意家中長時間煮食物，忘記爐火使鍋煮到無湯，食物焦化而引起火警。

住家房起火，所幸及時撲滅未釀成災害。(圖 / 記者[REDACTED]攝) 防止房火發生，消防局呼房應安裝「住宅用火警報器」，偵測到濃煙或高溫時，將發出高分貝聲響提醒民房；煮時瓦斯爐具附近勿放置易燃物品或危險物品。房可選用安全爐具，選擇有熄火安全裝置及感應功能爐具，偵測爐火熄滅，或鍋爐達到一定溫度時會自動切斷瓦斯，有效預防爐火烹調不慎所造成之火警案件。 | 將發出高分貝聲響提醒民房；煮時瓦斯爐具附近勿放置易燃物品或危險物品。房可選用安全爐具，選擇 | 熄火安全裝置及感應功能爐具，偵測爐火熄滅，或鍋爐達到一定溫度時會自動切斷瓦斯，有效預防爐火烹調不慎所 | [text17 | 30] | 30 |，發現 29 歲林姓男子涉 | 有 | 重嫌， | 有 | 2024 年 01 月 01 日 10:29 | 社會 | 半工半讀買的機車被偷！23 歲女人生第一輛 警埋伏 10hrs 抓賊 | 警員依法受理後，陪同張女回到停放位置，確定機車遭竊後，立即調監視器，發現 29 歲林姓男子涉有重嫌，警方循線，發現林男將車騎走後，隨後置於中和民利街一帶，警方於現場埋伏；約中午 11 時許，林男出現靠近機車時，警員立即將其上逮捕，依竊罪移送新北地檢署偵辦，將機車發還給張女。<https://www.ettoday.net/news/20231231/2652951.htm> | 2024 | 01 | 01 | 張女首購機車遭竊，警方埋伏 10 小時逮獲竊車的林男。(圖 / 記者陸運[REDACTED]翻攝) 記者陸運[REDACTED] / 新北報導新北市中和區一名女子，第一次所購買的機車，停放在中山路三段路邊失竊，不僅是唯一代步工具，也是人生首輛機車，非常具有紀念價值。警方獲報，立即調監視器，發現女子失竊機車遭人停放於民利街路邊，經過將近 10 個小時埋伏，竊車的男子出現牽車時，警員一擁而上將其逮捕。據了解，23 歲張姓女子，就讀大學時半工半讀才購買該輛機車，從讀書時期陪伴到如今出社會，都作她的唯一代步工具。12 月中(21 日)深夜 2 時許，張女返回中和中山路三段一帶牽車時，赫然發現機車遭竊，立即前往警局報案，過程中，張女擔心機車無法找回，表示該車殘余價值不高，但有許多過往回憶。

警方發現林男將機車停放路旁，立即於一旁埋伏。(圖 / 記者陸運[REDACTED]翻攝) 警員依法受理後，陪同張女回到停放位置，確定機車遭竊後，立即調監視器，發現 29 歲林姓男子涉有重嫌，警方循線，發現林男將車騎走後，隨後置於中和民利街一帶，警方於現場埋伏；約中午 11 時許，林男出現靠近機車時，警員立即將其上逮捕，依竊罪移送新北地檢署偵辦，將機車發還給張女。

林男竊得張女機車後，將其停放於中和民利街一帶。(圖 / 記者陸運[REDACTED]翻攝) | 警員依法受理後，陪同張女回到停放位置，確定機車遭竊後，立即調監視器，發現 29 歲林姓男子涉 | 重嫌，警方循線，發現林男將車騎走後，隨後置於中和民利街一帶，警方於現場埋伏； | [text19 | 46] | 46 |，網友始知台灣改車界 | 有 | 這號人物存在。 | 有 | 2024 年 01 月 01 日 13:00 | 社會 | 揭密廖老大打龜號進化史！他台積電工師 兩岸改裝達人之[REDACTED]曝 | 這輛傳奇在車涯 0-400 公尺加速紀錄榜的「[REDACTED]榔號」幕後技術支持者林大安，也因廖老大和大陸網紅杰爺隔空較勁，網友始知台灣改車界有這號人物存在。事實上在兩岸汽車改裝圈林大安已是教父級人物。<https://www.ettoday.net/news/20231231/2652951.htm> | 2024 | 01 | 01 | 臺灣改裝車達人林大安接受專訪娓娓道來兩岸 04 加速賽諸多不为人知的細節。(圖 / 記者張君豪攝) 記者張君豪 / 台北報導 2022 年 4 月 22 日以打龜號創下 0-400 公尺加速 8.59 秒成績刷新當時中國車涯排行榜，但後來廖老大深陷飲料加盟店糾紛，打龜號轉手到中部[REDACTED]榔大亨，綽號「麥校長」麥姓業者手上，該車改名「[REDACTED]榔號」繼續連破紀錄！台海兩岸改裝車界[REDACTED]起 0-400 加速競賽大賽；2023 年 4 月[REDACTED]榔號再度打破中國車涯排行榜 0-400 公尺加速紀錄創下 8.43 秒成績。這輛傳奇在車涯 0-400 公尺加速紀錄榜的「[REDACTED]榔號」幕後技術支持者林大安，也因廖老大和大陸網紅杰爺隔空較勁，網友始知台灣改車界有這號人物存在。事實上在兩岸汽車改裝圈林大安已是教父級人物。林大安日前接受《ETtoday 新聞雲》訪問、娓娓道來他從愛好研究改裝車輛，到兩岸改裝界、車迷幾乎無人不知的心路歷程。林大安表示，自己對於汽車研究的熱愛，肇因家庭背景，家中本來就是貨車運輸業，從小有記憶開始，就是做在大貨車上面「每天看的都是車」。林大安是台灣台南人，他 3-4 歲時就因每天看著貨車、[REDACTED]覽車會問長輩很多關於汽車的問題，加上台南人講台語，很多關於汽車或汽修、零件詞語都是日本外來語、或者台語的簡稱。廖老大開直播親曝割愛「打龜號」原因，由中部 04 賽車圈資深高手接手持續測試刷新紀錄。(圖 / 翻攝自抖音) 在耳濡目染之下，林大安對汽車學[REDACTED]生興趣，「[REDACTED]實話，我的第一個汽車學老師就是我父親，幼稚園以前天天吃喝拉撒睡都在卡車上面」，但台灣人傳統觀念，長輩相當反對林大安做汽車相關工作，[REDACTED]在家人[REDACTED]迫下高工念了電子科，但念一學期就休學轉入協和工商汽修科。林大安進入汽車學領域就如同躍魚游入大海，從協和汽修科到後來考入台北技術學院汽修科系都是第一名成績，

他在大三幾乎修完大學 4 年課程，大四除了體育課外、埋首學校汽修間和外面車廠干起黑手、不是拆解汽車就是修車，「同學都覺得很奇怪，人家是來[大學我是來修車]」，他畢業後闖蕩江湖、曾開過網咖、咖啡簡餐店，甚至到台積電擔任工程師，但後來遭遇全球金融海嘯，台灣景氣寒冬，不少工程師被迫休無薪假或者待遇縮[。在前途未明的狀[，林大安毅然到大陸開設自己的改裝品牌，[把自己淵博的汽車專業知識運用到車輛改裝部品上、從原本進口汽車改裝零件，逐漸擴展成替客[規劃擬定完整的汽車動力提升計畫，[逐漸在對岸改裝車界打響名號，對岸改裝車界都知道「台灣來的林大安很會改快車」但命運總是弄人，他與對岸知名的改裝技師，也是操刀大陸網紅「杰爺」04 賽車的張佳琪，本來是事業上的合作夥伴，甚至杰爺 2022 年嗚聲「台灣[快車」時，自豪的改裝[迪 TTRS 賽車更是出自於當時在張佳琪改裝公司任職的林大安之手。廖老大「打龜號」2022 年創下車涯 TTRS 最快 0-400 加速記[。(圖 / 翻攝自抖音)對此林大安謙[澄清，「網傳是我幫阿杰改車，但我只幫他改裝引擎動力和水冷系統，行車電腦調教部分對岸有另外聘請技師，[不能[整合車是我改的」甚至張佳琪 2017 年在上海 04 加速賽以日[GTR 跑車所創下當時全中國最快的 8.63 秒成績，該部改裝 GTR 賽車的動力提升也出自林大安之手，「[了 GTR 能順利出賽、我約 1 周[睡，搞到肺炎去醫院吊點滴，打完再回來改車」。兩名改裝達人合作期間，[推出兩人合力開發的汽車動力改裝品牌，但在台陸爆發「[快車」[議前後，對岸網紅和張對林大安的中傷抹黑不斷、甚至原本要幫廖老大出賽車輛調整行車電腦的大陸編程廠商也遭張佳琪要求終止「助台」合作、這場本是廖老大和「杰爺」的台陸最快 04 改裝車競賽，[然上升[兩岸最知名技師改車工藝之[。[管林大安幫廖老大操刀設計的打龜號一戰成名，[從去年迄今陸續創下刷下車涯 TTRS 最快的 04 加速記[，而張佳琪操刀的杰爺賽車迄今[拿不出對岸成績，但從打龜號到後來的[GTR 榜號屢屢霸榜刷新紀[，這場兩岸改裝達人之[相信觀[心中已有答案。對此林大安不願多做評論，他[調每人想法不同，「張比較專注在改快車、但我不喜歡改車、也不會改車」做[一個資深汽車改裝達人，林大安專注在定期發布改車科普觀念，[致力在車用動力提升部品的商品零件開發，「我真的不會改車、能讓改車族群都能用上性能與安全、耐用度兼具的[品，[提升改車圈的觀念才是我比較想做的。】 【其他新聞】

賣 7 輛名車全撞爛出險千萬！這家車行真的有鬼 警揭詐保真相

獨 / 人體菸灰缸！竹聯淫煞[拍少女凌虐片 逼交 50 萬「退幫費」

爆乳辣媽網售毒咖啡被捕畫面曝 白短褲露大腿還不忘拎包包 | 號」幕後技術支持者林大安，也因廖老大和大陸網紅杰爺隔空較勁，網友始知台灣改車界 | 號人物存在。事實上在兩岸汽車改裝圈林大安已是教父級人物。|

## 2.2.4 Combined analysis: Frequency table of the first word following *you ‘to have’*

```
## Extract the first word
kwic_data$post_first_word <- word(kwic_data$post, 1)

## We need to tranform the tokenized data into a 'dfm' dataset
kwic_data_freq <- dfm(
  tokens(kwic_data$post_first_word,
  remove_punct = TRUE)
)

kwic_data_freq <- textstat_frequency(kwic_data_freq)

## Clean a little bit
kwic_data_freq <- kwic_data_freq[-grep("[[:digit:]]", kwic_data_freq$feature),]

## Recreate the rank
kwic_data_freq$rank <- 1:length(kwic_data_freq$rank)

knitr::kable(head(kwic_data_freq, 100))
```

	feature	frequency	rank	docfreq	group
1	民[方]	1291	1	1291	all
3	多	694	2	694	all
4	網友	656	3	656	all
5	很多	608	4	608	all
6	[方]	600	5	600	all
7	問題	588	6	588	all
8	什[方]	580	7	580	all
10	媒體	550	8	550	all
11	的	533	9	533	all
12	一名	503	10	503	all
14	任何	476	11	476	all
15	逃亡	445	12	445	all
16	可能	437	13	437	all
17	許多	399	14	399	all
18	機會	395	15	395	all
19	一個	338	16	338	all
21	信心	298	17	298	all
22	其他	298	18	298	all
23	在	271	19	271	all
24	更多	264	20	264	all
25	一些	262	21	262	all
26	勾	256	22	256	all
27	相當	249	23	249	all
28	不少	234	24	234	all
29	需要	228	25	228	all
30	重	223	26	223	all
31	這樣	222	27	222	all
32	跟	222	28	222	all
33	必要	220	29	220	all
35	時	207	30	207	all
36	不同	200	31	200	all
37	毒品	198	32	198	all
38	能力	196	33	196	all
39	多少	194	34	194	all
40	違	191	35	191	all
43	這[方]	181	36	181	all
44	非常	181	37	181	all
45	部分	177	38	177	all
46	[方]押	176	39	176	all
48	相關	172	40	172	all
49	一定	169	41	169	all
50	明顯	162	42	162	all
51	共識	156	43	156	all
53	串	153	44	153	all
54	超過	152	45	152	all
55	被	151	46	151	all
56	責任	149	47	149	all
57	重大	147	48	147	all
58	很大	147	49	147	all
59	違反	140	50	140	all
60	更	139	51	139	all
61	酒	135	52	135	all

	feature	frequency	rank	docfreq	group
62	看到	135	53	135	all
63	疑慮	134	54	134	all
64	意願	132	55	132	all
65	意見	130	56	130	all
66	對	128	57	128	all
67	興趣	126	58	126	all
68	會議	124	59	124	all
69	自己	124	60	124	all
70	發生	122	61	122	all
71	擦	120	62	120	all
72	大量	115	63	115	all
73	糾紛	111	64	111	all
74	幫助	111	65	111	all
75	一位	111	66	111	all
76	過	110	67	110	all
77	向	108	68	108	all
78	疏失	106	69	106	all
79	條件	106	70	106	all
80	債務	106	71	106	all
81	事實	104	72	104	all
82	不	103	73	103	all
83	兩個	102	74	102	all
84	高度	101	75	101	all
85	E狀	100	76	100	all
86	違法	99	77	99	all
87	據	96	78	96	all
88	多次	94	79	94	all
89	政治	93	80	93	all
90	性	93	81	93	all
92	兩	91	82	91	all
94	做	90	83	90	all
95	高達	89	84	89	all
96	男子	89	85	89	all
97	諸多	88	86	88	all
98	E常	86	87	86	all
99	瑕疵	86	88	86	all
100	大	86	89	86	all
101	過失	86	90	86	all
102	哪些	85	91	85	all
103	F	85	92	85	all
104	幾個	84	93	84	all
105	可能是	84	94	84	all
106	與	83	95	83	all
107	車輛	82	96	82	all
109	去	81	97	81	all
110	一輛	81	98	81	all
111	這種	80	99	80	all
112	藍	78	100	78	all

## 2.4 Save the data

### 2.4.1 Save as an Excel file

```
write.xlsx(kwic_data, "ArticleETTToday_KWIC_You.xlsx")
```

### 2.4.2 Save as an RData file

```
save(kwic_data, file = "ArticleETTToday_KWIC_You.Rdata")
```

## 3. Frequency tables

### 3.1 Create the overall frequency table

#### 3.1.1 Creation of the first table

```
## We need to transform the tokenized data into a 'dfm' dataset
Article_tokens_frequency <- dfm(
  tokens(Article_total2$body,
         remove_punct = TRUE))
Article_tokens_frequency <- textstat_frequency(Article_tokens_frequency)

table_AllWordsFreq_Top100 <- head(Article_tokens_frequency, 100)
table_AllWordsFreq_Top100
```

##	feature	frequency	rank	docfreq	group
## 1	的	357709	1	159298	all
## 2	在	116544	2	88581	all
## 3	日	84106	3	66187	all
## 4	後	71392	4	57501	all
## 5	人	69084	5	49761	all
## 6	時	67573	6	53558	all
## 7	有	65657	7	53298	all
## 8	男	63140	8	32390	all
## 9	與	61733	9	49197	all
## 10	是	59999	10	48402	all
## 11	也	59965	11	51054	all
## 12	及	49873	12	38094	all
## 13	表示	49840	13	47384	all
## 14	年	48807	14	33965	all
## 15	但	47947	15	43097	all
## 16	他	47910	16	34208	all
## 17	將	45675	17	39070	all
## 18	等	44905	18	36914	all
## 19	被	44599	19	36588	all
## 20	到	44505	20	38318	all
## 21	2	43707	21	34342	all

## 22	月姓	43394	22	32669	all
## 23	對	42361	23	28582	all
## 24	陳	41664	24	35874	all
## 25	民	41052	25	26393	all
## 26	要	40981	26	29991	all
## 27	F	39541	27	31461	all
## 28	F	38658	28	34703	all
## 29	台灣	38028	29	24641	all
## 30	案	36770	30	29132	all
## 31	1	36107	31	28584	all
## 32	F	35360	32	30056	all
## 33	警方	35205	33	26808	all
## 34	中	34824	34	30380	all
## 35	3	34319	35	28447	all
## 36	F	33596	36	29140	all
## 37	柯	33164	37	18553	all
## 38	文	32255	38	21397	all
## 39	以	32162	39	28049	all
## 40	F	31311	40	20937	all
## 41	上	31281	41	27774	all
## 42	不	30456	42	26271	all
## 43	民進黨	29648	43	20912	all
## 44	就	29633	44	25843	all
## 45	林	29426	45	19296	all
## 46	之	29209	46	18968	all
## 47	女	29041	47	16593	all
## 48	車	28849	48	18869	all
## 49	歲	27927	49	19862	all
## 50	讓	27838	50	24172	all
## 51	名	27598	51	22325	all
## 52	於	27440	52	23486	all
## 53	而	27073	53	25039	all
## 54	她	26964	54	17524	all
## 55	會	26866	55	22316	all
## 56	發生	26700	56	23129	all
## 57	了	25947	57	21844	all
## 58	前	25547	58	22155	all
## 59	大	25397	59	21497	all
## 60	國民黨	25126	60	17804	all
## 61	檢	24367	61	18605	all
## 62	發現	23872	62	20970	all
## 63	賴	23781	63	16125	all
## 64	立委	23697	64	17483	all
## 65	因	23330	65	21330	all
## 66	已	23243	66	21044	all
## 67	調查	23167	67	19220	all
## 68	都	23092	68	20346	all
## 69	指出	23045	69	22728	all
## 70	4	22781	70	19855	all
## 71	哲	22453	71	14297	all
## 72	自己	22438	72	19022	all
## 73	F有	21977	73	19189	all
## 74	和	21507	74	17793	all
## 75	5	21402	75	18870	all

```

## 76 跟 21213 76 17565 all
## 77 男子 20713 77 15449 all
## 78 我 20494 78 13075 all
## 79 依 20250 79 18416 all
## 80 多 20149 80 17965 all
## 81 這 20140 81 18193 all
## 82 黨 19999 82 13527 all
## 83 萬元 19934 83 14724 all
## 84 遭 19778 84 17403 all
## 85 分 19477 85 16186 all
## 86 小 19266 86 12596 all
## 87 今 19204 87 18762 all
## 88 德 19140 88 14233 all
## 89 該 19098 89 16180 all
## 90 李 18916 90 12213 all
## 91 人員 18698 91 14908 all
## 92 向 18625 92 17039 all
## 93 許 18586 93 15277 all
## 94 10 18579 94 16677 all
## 95 長 18579 94 15239 all
## 96 或 18303 96 14481 all
## 97 6 18281 97 15998 all
## 98 政府 18281 97 14640 all
## 99 清 17969 99 13651 all
## 100 相關 17961 100 16032 all

```

```
## Two problems occured: Digits, and Unknown words wrongly segmented
```

### 3.1.2 Clean it up a little bit

```

## Example with numbers
table_FreqWord <- Article_tokens_frequency[-grep("[[:digit:]]",
  Article_tokens_frequency$feature),]

## Redo the ranking
table_FreqWord$rank <- 1:length(table_FreqWord$rank)

```

### 3.1.3 Final table, addition of the percentage

```

table_FreqWord_Top100 <- head(table_FreqWord, 100)
table_FreqWord_Top100

```

```

##      feature frequency rank docfreq group
## 1      的 357709    1 159298 all
## 2      在 116544    2  88581 all
## 3      日  84106    3  66187 all
## 4      後  71392    4  57501 all
## 5      人  69084    5  49761 all
## 6      時  67573    6  53558 all

```

## 7	有	65657	7	53298	all
## 8	男	63140	8	32390	all
## 9	與	61733	9	49197	all
## 10	是	59999	10	48402	all
## 11	也	59965	11	51054	all
## 12	及	49873	12	38094	all
## 13	表示	49840	13	47384	all
## 14	年	48807	14	33965	all
## 15	但	47947	15	43097	all
## 16	他	47910	16	34208	all
## 17	將	45675	17	39070	all
## 18	等	44905	18	36914	all
## 19	被	44599	19	36588	all
## 20	到	44505	20	38318	all
## 22	月	43394	21	32669	all
## 23	姓	42361	22	28582	all
## 24	對	41664	23	35874	all
## 25	陳	41052	24	26393	all
## 26	民	40981	25	29991	all
## 27	要	39541	26	31461	all
## 28	F	38658	27	34703	all
## 29	台灣	38028	28	24641	all
## 30	案	36770	29	29132	all
## 32	F	35360	30	30056	all
## 33	警方	35205	31	26808	all
## 34	中	34824	32	30380	all
## 36	F	33596	33	29140	all
## 37	柯	33164	34	18553	all
## 38	文	32255	35	21397	all
## 39	以	32162	36	28049	all
## 40	F	31311	37	20937	all
## 41	上	31281	38	27774	all
## 42	不	30456	39	26271	all
## 43	民進黨	29648	40	20912	all
## 44	就	29633	41	25843	all
## 45	林	29426	42	19296	all
## 46	之	29209	43	18968	all
## 47	女	29041	44	16593	all
## 48	車	28849	45	18869	all
## 49	歲	27927	46	19862	all
## 50	讓	27838	47	24172	all
## 51	名	27598	48	22325	all
## 52	於	27440	49	23486	all
## 53	而	27073	50	25039	all
## 54	她	26964	51	17524	all
## 55	會	26866	52	22316	all
## 56	發生	26700	53	23129	all
## 57	了	25947	54	21844	all
## 58	前	25547	55	22155	all
## 59	大	25397	56	21497	all
## 60	國民黨	25126	57	17804	all
## 61	檢	24367	58	18605	all
## 62	發現	23872	59	20970	all
## 63	賴	23781	60	16125	all

```

## 64 立委 23697 61 17483 all
## 65 因 23330 62 21330 all
## 66 已 23243 63 21044 all
## 67 調查 23167 64 19220 all
## 68 都 23092 65 20346 all
## 69 指出 23045 66 22728 all
## 71 哲 22453 67 14297 all
## 72 自己 22438 68 19022 all
## 73 [F]有 21977 69 19189 all
## 74 和 21507 70 17793 all
## 76 跟 21213 71 17565 all
## 77 男子 20713 72 15449 all
## 78 我 20494 73 13075 all
## 79 依 20250 74 18416 all
## 80 多 20149 75 17965 all
## 81 這 20140 76 18193 all
## 82 黨 19999 77 13527 all
## 83 萬元 19934 78 14724 all
## 84 遭分 19778 79 17403 all
## 85 小 19477 80 16186 all
## 86 今 19266 81 12596 all
## 87 德 19204 82 18762 all
## 88 該 19140 83 14233 all
## 89 李 19098 84 16180 all
## 90 人員 18916 85 12213 all
## 91 向 18698 86 14908 all
## 92 許 18625 87 17039 all
## 93 長 18586 88 15277 all
## 95 或 18579 89 15239 all
## 96 政府 18303 90 14481 all
## 98 清 18281 91 14640 all
## 99 相關 17969 92 13651 all
## 100 國 17961 93 16032 all
## 101 處警 17914 94 13623 all
## 102 警 17898 95 15036 all
## 103 總統 17804 96 15484 all
## 104 [F] 17774 97 13169 all
## 105 張 17680 98 12860 all
## 107 進行 17594 99 11644 all
## 108 進行 17570 100 15739 all

```

```

table_FreqWord_Top100$percentage <-
  round(table_FreqWord_Top100$frequency/sum(table_FreqWord$frequency)*100, 5)
table_FreqWord_Top100

```

	feature	frequency	rank	docfreq	group	percentage
## 1	的	357709	1	159298	all	2.37874
## 2	在	116544	2	88581	all	0.77501
## 3	日	84106	3	66187	all	0.55930
## 4	後	71392	4	57501	all	0.47475
## 5	人	69084	5	49761	all	0.45940
## 6	時	67573	6	53558	all	0.44936
## 7	有	65657	7	53298	all	0.43661

## 8	男	63140	8	32390	all	0.41988
## 9	與	61733	9	49197	all	0.41052
## 10	是	59999	10	48402	all	0.39899
## 11	也	59965	11	51054	all	0.39876
## 12	及	49873	12	38094	all	0.33165
## 13	表示	49840	13	47384	all	0.33143
## 14	年	48807	14	33965	all	0.32456
## 15	但	47947	15	43097	all	0.31884
## 16	他	47910	16	34208	all	0.31860
## 17	將	45675	17	39070	all	0.30374
## 18	等	44905	18	36914	all	0.29862
## 19	被	44599	19	36588	all	0.29658
## 20	到	44505	20	38318	all	0.29596
## 22	月	43394	21	32669	all	0.28857
## 23	姓	42361	22	28582	all	0.28170
## 24	對	41664	23	35874	all	0.27706
## 25	陳	41052	24	26393	all	0.27299
## 26	民	40981	25	29991	all	0.27252
## 27	要	39541	26	31461	all	0.26295
## 28	F	38658	27	34703	all	0.25707
## 29	台灣	38028	28	24641	all	0.25288
## 30	案	36770	29	29132	all	0.24452
## 32	F	35360	30	30056	all	0.23514
## 33	警方	35205	31	26808	all	0.23411
## 34	中	34824	32	30380	all	0.23158
## 36	F	33596	33	29140	all	0.22341
## 37	柯	33164	34	18553	all	0.22054
## 38	文	32255	35	21397	all	0.21449
## 39	以	32162	36	28049	all	0.21388
## 40	F	31311	37	20937	all	0.20822
## 41	上	31281	38	27774	all	0.20802
## 42	不	30456	39	26271	all	0.20253
## 43	民進黨	29648	40	20912	all	0.19716
## 44	就	29633	41	25843	all	0.19706
## 45	林	29426	42	19296	all	0.19568
## 46	之	29209	43	18968	all	0.19424
## 47	女	29041	44	16593	all	0.19312
## 48	車	28849	45	18869	all	0.19184
## 49	歲	27927	46	19862	all	0.18571
## 50	讓	27838	47	24172	all	0.18512
## 51	名	27598	48	22325	all	0.18352
## 52	於	27440	49	23486	all	0.18247
## 53	而	27073	50	25039	all	0.18003
## 54	她	26964	51	17524	all	0.17931
## 55	會	26866	52	22316	all	0.17866
## 56	發生	26700	53	23129	all	0.17755
## 57	了	25947	54	21844	all	0.17255
## 58	前	25547	55	22155	all	0.16989
## 59	大	25397	56	21497	all	0.16889
## 60	國民黨	25126	57	17804	all	0.16709
## 61	檢	24367	58	18605	all	0.16204
## 62	發現	23872	59	20970	all	0.15875
## 63	賴	23781	60	16125	all	0.15814
## 64	立委	23697	61	17483	all	0.15758

```

## 65 因 23330 62 21330 all 0.15514
## 66 已 23243 63 21044 all 0.15456
## 67 調查 23167 64 19220 all 0.15406
## 68 都 23092 65 20346 all 0.15356
## 69 指出 23045 66 22728 all 0.15325
## 71 哲 22453 67 14297 all 0.14931
## 72 自己 22438 68 19022 all 0.14921
## 73 [F]有 21977 69 19189 all 0.14615
## 74 和 21507 70 17793 all 0.14302
## 76 跟 21213 71 17565 all 0.14107
## 77 男子 20713 72 15449 all 0.13774
## 78 我 20494 73 13075 all 0.13628
## 79 依 20250 74 18416 all 0.13466
## 80 多 20149 75 17965 all 0.13399
## 81 這 20140 76 18193 all 0.13393
## 82 黨 19999 77 13527 all 0.13299
## 83 萬元 19934 78 14724 all 0.13256
## 84 遭 19778 79 17403 all 0.13152
## 85 分 19477 80 16186 all 0.12952
## 86 小 19266 81 12596 all 0.12812
## 87 今 19204 82 18762 all 0.12771
## 88 德 19140 83 14233 all 0.12728
## 89 該 19098 84 16180 all 0.12700
## 90 李 18916 85 12213 all 0.12579
## 91 人員 18698 86 14908 all 0.12434
## 92 向 18625 87 17039 all 0.12386
## 93 許 18586 88 15277 all 0.12360
## 95 長 18579 89 15239 all 0.12355
## 96 或 18303 90 14481 all 0.12171
## 98 政府 18281 91 14640 all 0.12157
## 99 清 17969 92 13651 all 0.11949
## 100 相關 17961 93 16032 all 0.11944
## 101 國 17914 94 13623 all 0.11913
## 102 處 17898 95 15036 all 0.11902
## 103 警 17804 96 15484 all 0.11840
## 104 總統 17774 97 13169 all 0.11820
## 105 [F] 17680 98 12860 all 0.11757
## 107 張 17594 99 11644 all 0.11700
## 108 進行 17570 100 15739 all 0.11684

```

## 3.2 Save the data

### 3.2.1 Save as an Excel file

```
write.xlsx(table_FreqWord_Top100, "ArticleETToday_Top100nouns.xlsx")
```

### 3.2.2 Save as an RData file

```
save(table_FreqWord_Top100, file = "ArticleETToday_Top100nouns.Rdata")
```