Recitation 12: Text analytics

Seamus Wagner

2021-11-23

Text analytics

Text as data and text analytics are coming into quantitative political science work rapidly. There are two main avenues for improvement that I see. Because text as data is such a broad topic, this is heavilty influenced by how I see the application and what *should* be done as we move forward. The two main avenues for cutting edge work is in measurement (pre-processing) and how to adapt/create models using different types of text (whether this is through different languages, different media, or different topics). Those two are incredibly broad themselves, but by measurement or pre-processing, the prevailing literature revolves around a number of decisions that are more or less uniformly applied or (maybe?) theoretically derived. These types of decisions are what are necessary to go from raw data to some series of numbers usually a document term matrix/term document matrix. They are interchanable and you will see both DTM/TDM. Moving from raw data to a DTM is where we will spend most of this lab. The second topic is adapting existing models to new data scenarios. The two main types of learning models we use are supervised and unsupurvised.

Supervised models

There are others, but for the introduction, these two are the important frameworks of modeling. Supervised models are used for classification, where you feed a model a bunch of text (or images, or some other data) with identifiers present to train on. Then, we feed that model unidentified text (or images, or other data) and the model will classify what we give it using the trained identifiers. These are popularly used for document classification, speech recognition, and sentiment recognition. The goal of supervised models is to increase certainty and accuracy of classification. When using these models, you can actually tell how well these perform (assuming you know what the underlying thing you are training it for).

Unsupervised Models

For unsupervised models, you are trying to understand some latent structure. In practice, many of the models we use are likely a combination of supervised and unsupervised. For instance, topic modeling is likely the most commonly used unsupervised approach in social sciences. Topic models treat your text as a bag of words that is unstructured. It then finds “topics" present in the data but varyingly so. A good link is here to show what they do visually. Topic modeling (LDA) specifically, is one of the most common models used. It is also not perfect, as no mdoels are. BUt, it is important to note with many of these models, there *will* be misclassifications and some readers will not like that. It is important to think about how accurate they can be and building into your research design some level of human validity check is wise. See Table 3 here for how even the *best* model only captures $~66%$ of human coded set of Tweets. One of the bigger hurdles for social scientists using this approach is a removal from statistical significance and causality. You can validate against other models, which is typically what is done, or use smaller subsets and human expert coding. There is room for more validation techniques. Though, remember that human coders are unreliable, and so you need intercoder reliability checks, this gets incredibly expensive, nearly prohibitively so for non-native languages. See here for a good introduction to this concept. There is and likely will always be room for improvement in how humans interact with computers to emulate human decision-making.

Pre-processing

Pre-processing is one of(*the* imo) the most important decisions that all of you will have to make if you use text for quantitative analyses, and that is to turn those words into numbers somehow. The world of ComSci where many of these techniques evolved out of hold human judgment as the gold standard. Think about what sentiment analysis does, it codes a sentences/document into positive/negative sentiments. There are many many types of models, some of which use more than positive/negative as the only binary, see here for NRC sentiment analysis developed by Saif Mohammad. This is implemented in the syuzhet R package as well. Some of the issues with these models are English language centrality, human subjectivity, and the roots in product review. Those three are not the only issues, but some that I find troubling. First, these types of models, where a human or some number of humans, determine a list of words that invoke some latent sentiment (itself subjective) that is coded into the computer in English. That means for other languages we need reliable translation for the latent meaning of words beyond literal translations. Automated translations are getting better, but are still not great for *all* languages. Second, related to the first, as language shapes how people see the worlds, humans determine what is subjectively positive, negative, or any other socially constructed latent topic we may be interested in. Third, these were all built for determining whether people will buy your shit in the future and are mostly concerned with how people review products and services, which is a different set of linguistic rules as political discourse, newspaper writing, Twitter posts, and party manifestos. Twitter for instance, is very lax with syntax and structure compared to party manifestos or academic articles or newspaper articles. The structure and choice of words shifts with these contests, making things like sentiment analysis difficult to justify in many social science circumstances. It is a good place to start maybe, but usually not sufficiently convincing to me at least, this is not necessarily a lab of what will get you published but how to think about approaching text data.

The above is maybe a bit of a tangent to actual pre-processing, which I will get into here. Matt Denny has a really good PA article here on this with Arthur Spirling that I think you should read if you plan to do any unsupervised learning models. The bulk of the decisions you will make will be whether to include punctuation, numbers, all lowercase words, just the root words, what/if any stopwords to remove, how many n-grams to include, and whether to include infrequent words. What to include and not is not only down to theory. Particularly in the unsupervised world, where theory may not be as well-identified. Further, drawing on supervised norms may not be ideal when we shift to unsupervised. When we think about what the goal of a supervised model versus an unsupervised model is, we move from classification to unveiling latent structures.

Practical examples

Moving forward, here is some R code for how to pre-process and packages that you may find helpful. That being said, R is not good at this compared to others. R is very good, quanteda specifically, at generating DTMs though. Matt Denny highlights the following two as faster mallet and coreNLP.

```r
rm(list=ls())
library(dplyr)
library(data.table)
library(ggplot2)
library(stringr)
library(quanteda)
library(quanteda.textmodels)
library(quanteda.textstats)
library(quanteda.textplots)
library(readtext)
library(newsmap)
library(seededlda)
library(spacyr)
```

There are a number of other useful packages, firstly being stringr. A cheatsheet for stringr is found here. stringr is part of the tidyverse and has four main functions. The first is character manipulation, the second is dealing with whitespace, the third is case sensative and locale operations, and the fourth is pattern matching functions (mostly used for regular expressions). The grepl() family is the base R equivalent I use both, mostly since I learned certain operations in one or the other. I do try to work within tidyverse as I have mentioned previously, just so packages and syntax is consistent.

Regular expressions are something that I still need to look up every time I use them. Regular expressions are a way to represent patterns in strings. stringr's page on regular expressions is found here. Essentially, in R you must add an additional ** before a regular expressions because of some reason that the specific character cannot be represented in a character string in R. I will walk through a bit of an example here and hopefully you will understand the intuition behind regular expressions and will look into them more if you need them. These are not all regular expressions, but some useful ones and I only use str_match(_all) in this section, stringr has more applications than just this.

```
# We will start with a basic series of strings
letters <- c("abc", "abcd", "abcdf", "nmo")
# Then use the str_match function to find all instances of ab as a pattern
str_match(letters, "ab")
```

```
##      [,1]
## [1,] "ab"
## [2,] "ab"
## [3,] "ab"
## [4,] NA
```

```
# We can also define the pattern and then use it in the function
# This pattern looks complex, but will hopefully make sense in a minute
pattern <- '.*(\\d{3}).*(\\d{3}).*(\\d{4})'

# Here is a fake list of phone numbers I came up with
numbers <- c(
  "814-571 4848",
  "(614) 978 8765",
  "Work: 458-874-8145",
  "8149337991",
  "I do not own a phone"
)
# Now let's use str_match to find these phone numbers, can anyone see what the regex is doing?
str_match(numbers, pattern)
```

```
##      [,1]                 [,2]  [,3]  [,4]
## [1,] "814-571 4848"       "814" "571" "4848"
## [2,] "(614) 978 8765"     "614" "978" "8765"
## [3,] "Work: 458-874-8145" "458" "874" "8145"
## [4,] "8149337991"         "814" "933" "7991"
## [5,] NA                   NA    NA    NA
```

Let’s break it down.

```
# \d is a number in regex, we need \\d in R.

# {3} is how many characters(numbers) to select (match in our case)

# If we did {1} it would only pull the first number in each match
str_match(numbers, "\\d")
```

```
##      [,1]
## [1,] "8"
## [2,] "6"
## [3,] "4"
## [4,] "8"
## [5,] NA
```

```
# We can also add the + to match one or more
str_match(numbers, "\\d+")
```

```
##      [,1]
## [1,] "814"
## [2,] "614"
## [3,] "458"
## [4,] "8149337991"
## [5,] NA
```

```
# This can be read asmatch one or more digits from object numbers
```

```
# In regex (to my knowledge) all characters used as a regular expression have a complementary operator
str_match(numbers, "\\D+")
```

```
##      [,1]
## [1,] "-"
## [2,] "("
## [3,] "Work: "
## [4,] NA
## [5,] "I do not own a phone"
```

```
# We can also use str_match_all if we want to the functino to continue after it finds the first match
str_match_all(numbers, "\\d+")
```

```
## [[1]]
##      [,1]
## [1,] "814"
## [2,] "571"
## [3,] "4848"
##
## [[2]]
##      [,1]
## [1,] "614"
## [2,] "978"
## [3,] "8765"
##
## [[3]]
##      [,1]
## [1,] "458"
## [2,] "874"
## [3,] "8145"
##
## [[4]]
##      [,1]
## [1,] "8149337991"
##
## [[5]]
##      [,1]
```

```r
# w is another useful one, where it returns alpha-numeric characters
str_match_all(numbers, "\\w+")
```

```
## [[1]]
##      [,1]
## [1,] "814"
## [2,] "571"
## [3,] "4848"
##
## [[2]]
##      [,1]
## [1,] "614"
## [2,] "978"
## [3,] "8765"
##
## [[3]]
##      [,1]
## [1,] "Work"
## [2,] "458"
## [3,] "874"
## [4,] "8145"
##
## [[4]]
##      [,1]
## [1,] "8149337991"
##
## [[5]]
##      [,1]
## [1,] "I"
## [2,] "do"
## [3,] "not"
## [4,] "own"
## [5,] "a"
## [6,] "phone"
```

```
# The opposite is helpful for what all you are exlcuding (typically you will filter out whitespace and p
str_match(numbers, "\\W+")
```

```
##      [,1]
## [1,] "-"
## [2,] "("
## [3,] ": "
## [4,] NA
## [5,] " "
```

```
# If you have interest in whitespace, s and S are helpful.
str_match(numbers, "\\s+")
```

```
##      [,1]
## [1,] " "
## [2,] " "
## [3,] " "
## [4,] NA
## [5,] " "
```

```
str_match(numbers, "\\S+")
```

```
##      [,1]
## [1,] "814-571"
## [2,] "(614)"
## [3,] "Work:"
## [4,] "8149337991"
## [5,] "I"
```

```
# If we do not know what we have in our character vectors, we use the wildcard expression .

sports <- c("swimming", "Rugby", "football", "american Football", "ice hockey")
# This returns the first thing that appears
str_match(sports, ".")
```

```
##      [,1]
## [1,] "s"
## [2,] "R"
## [3,] "f"
## [4,] "a"
## [5,] "i"
```

```
# Returns all of it
str_match(sports, ".+")
```

```
##      [,1]
## [1,] "swimming"
## [2,] "Rugby"
## [3,] "football"
## [4,] "american Football"
## [5,] "ice hockey"
```

```
# Sometimes we want to match things based on a ngram so we can use this code
str_match(sports, "am.+")
```

```
##      [,1]
## [1,] NA
## [2,] NA
## [3,] NA
## [4,] "american Football"
## [5,] NA
```

```
#Regular expressions turn really useful when we have multiple conditions

# what if we want anything that contains any set but not necessarily in that order
str_match(sports, "oo.+")
```

```
##      [,1]
## [1,] NA
## [2,] NA
## [3,] "ootball"
## [4,] "ootball"
## [5,] NA
```

```
str_match(sports, "[oo].+")
```

```
##      [,1]
## [1,] NA
## [2,] NA
## [3,] "ootball"
## [4,] "ootball"
## [5,] "ockey"
```

```
str_match(sports, "[ruoam].+")
```

```
##      [,1]
## [1,] "mming"
## [2,] "ugby"
## [3,] "ootball"
## [4,] "american Football"
## [5,] "ockey"
```

```
# You can also use the conditional bar | to separate each of them, but you need () not [] for multiple
str_match(sports, "(r|u|o|a|m).*")
```

```
##      [,1]                [,2]
## [1,] "mming"             "m"
## [2,] "ugby"              "u"
## [3,] "ootball"           "o"
## [4,] "american Football" "a"
## [5,] "ockey"             "o"
```

Now let's read the first one again

```
# return 0 or more wildcard characters, then give me three digits
# then 0 or more wildcard characters, then give me three digits
# then 0 or more wildcard characters, then give me 4 numbers
# The digits being wrapped in a () means we capture those, and then ignore stuff outside of it.

# Quick note that this only works for an exmaple like this, real phone number regular expressions shoul

pattern <- '.*(\\d{3}).*(\\d{3}).*(\\d{4})'

numbers <- c(
  "814-571 4848",
  "(614) 978 8765",
  "Work: 458-874-8145",
  "8149337991",
  "I do not own a phone"
)
str_match(numbers, pattern)
```

```
##      [,1]                 [,2]  [,3]  [,4]
## [1,] "814-571 4848"       "814" "571" "4848"
## [2,] "(614) 978 8765"     "614" "978" "8765"
## [3,] "Work: 458-874-8145" "458" "874" "8145"
## [4,] "8149337991"         "814" "933" "7991"
## [5,] NA                   NA    NA    NA
```

Other useful text as data stuff to know

Locale is selected by ISO2c codes. Base R also supports this functionality. This is important for non-English words and alphabets where letter order may be different if you are doing a sorting function for instance. stringr allows to set it within a function, since the global environment sort() and order() use English.

```
# Let's make a string and change it around a bit

x <- "I am using this text as practice."

str_to_upper(x)
```

```
## [1] "I AM USING THIS TEXT AS PRACTICE."
```

```
str_to_title(x)
```

```
## [1] "I Am Using This Text As Practice."
```

```
str_to_lower(x) # This one is most often used in preprocessing.
```

```
## [1] "i am using this text as practice."
```

```
str_to_lower(x, "tr") #Turkish has two lowercase i's so it is a useful example.
```

```
## [1] "i am using this text as practice."
```

Quick applied example

I will use a bit of code from an earlier lab to get some text data from BBC news articles.

```
library(rvest)
final_df <- data.table()
get_data <- function(article_link) {
article_page <- read_html(article_link) #URL defined from page
article_content <- article_page %>% html_nodes("#main-heading") %>% #Heading of each page
html_text() #Text in that heading
article_content$author <- article_page %>% html_nodes("strong") %>% #Author of news article
html_text() #Author name
article_content$date <- article_page %>% html_nodes("time") %>% #Date of publication
html_text() #Date
article_content$content <- article_page %>% html_nodes("p") %>% #Content of article
html_text() %>% paste(collapse = ",") #Content of article
return(article_content)
}

for (page_result in c(1,2)){ #The 2 in this case is manually entered as the nubmer of pages the search
link <- paste0("https://www.bbc.co.uk/search?q=tanzania+corruption&page=", #The page= is where the c(1,
page_result, "")
page <- read_html(link)
title <- page %>% html_nodes(".headline") %>% #Title of article
html_text() #Text of title
article_links <- page %>% html_nodes(".e1f5wbog0") %>% #Links in the article to go into the page for co
html_attr("href") %>% paste('', ., sep = "")
article_data <- sapply(article_links, FUN = get_data, USE.NAMES = F) #Do this for all links (pages)
final_df <- rbind(final_df, rbindlist(article_data, fill=TRUE)) #Combining all the data
print(paste("Page:", page_result))
}
```

```
## [1] "Page: 1"
## [1] "Page: 2"
```

Now we are only really interested in one column here, the content column. So we can take a number of approaches. The first is a corpus, which combines the text data with document level data and you can manipulate them after. Here is a quick example,

```
#?corpus
text <- final_df$content
corp_corruption <- corpus(text)
print(corp_corruption)
```

```
## Corpus consisting of 24 documents.
## text1 :
```

```
## "By Mark DoyleBBC International Development Correspondent ,Th..."
##
## text2 :
## ",Mike Corey is off the coast of Tanzania, exploring a coral ..."
##
## text3 :
## ",Last updated on 12 December 201912 December 2019.From the s..."
##
## text4 :
## ",Last updated on 12 December 201912 December 2019.From the s..."
##
## text5 :
## ",Last updated on 12 December 201912 December 2019.From the s..."
##
## text6 :
## "The President of Tanzania, Benjamin Mkapa, has defended the ..."
##
## [ reached max_ndoc ... 18 more documents ]
```

```
summary(corp_corruption)
```

```
## Corpus consisting of 24 documents, showing 24 documents:
##
##    Text Types Tokens Sentences
##   text1   349    763         4
##   text2    84    143         1
##   text3   348    575         4
##   text4   348    575         4
##   text5   348    575         4
##   text6    95    151         2
##   text7   124    192         3
##   text8   124    192         3
##   text9   124    192         3
##  text10   290    509         5
##  text11   310    583         5
##  text12   544   1107         8
##  text13   312    552         4
##  text14   341    654         4
##  text15   327    579         5
##  text16    83    124         5
##  text17    18     20         1
##  text18    30     33         1
##  text19    74    106         1
##  text20    46     59         1
##  text21   219    401         2
##  text22    34     41         1
##  text23    61    106         3
##  text24    26     29         1
```

Notice the names are text1, … is not ideal. One real-world issue is that these data are not identified correctly by authors (NA for all but 2), there are not titles or dates for a number of others.

A second option is to tekenize it. Tokens break apart each token of a text, characters, punctuation, numbers, etc. You then get a list of each item of the corpus.

```r
toks_corruption <- tokens(corp_corruption)
print(toks_corruption)
```

```
## Tokens consisting of 24 documents.
## text1 :
##  [1] "By"            "Mark"          "DoyleBBC"      "International"
##  [5] "Development"   "Correspondent" ","             "The"          
##  [9] "British"       "arms"          "and"           "aircraft"     
## [ ... and 751 more ]
## 
## text2 :
##  [1] ","         "Mike"      "Corey"     "is"        "off"       "the"      
##  [7] "coast"     "of"        "Tanzania"  ","         "exploring" "a"        
## [ ... and 131 more ]
## 
## text3 :
##  [1] ","        "Last"     "updated"  "on"       "12"       "December"
##  [7] "201912"   "December" "2019"     "."        "From"     "the"     
## [ ... and 563 more ]
## 
## text4 :
##  [1] ","        "Last"     "updated"  "on"       "12"       "December"
##  [7] "201912"   "December" "2019"     "."        "From"     "the"     
## [ ... and 563 more ]
## 
## text5 :
##  [1] ","        "Last"     "updated"  "on"       "12"       "December"
##  [7] "201912"   "December" "2019"     "."        "From"     "the"     
## [ ... and 563 more ]
## 
## text6 :
##  [1] "The"       "President" "of"        "Tanzania"  ","         "Benjamin" 
##  [7] "Mkapa"     ","         "has"       "defended"  "the"       "work"     
## [ ... and 139 more ]
## 
## [ reached max_ndoc ... 18 more documents ]
```

To show a keyword in context approach, which shows you a few words around specific keywords, you can use the kwic command. It will take keywords (typically a root(s) of some sort) and show you what is around the words. You can also specify how many words with window(). This is useful ad a face validity check for yourself to see what your scraping or data collection process gave you. You can also specify exact phrases, using phrase() instead of pattern(). Notice that these are not super helpful beyond identifying some themes for you to look into.

```r
toks <- final_df$content
kwic(toks, pattern =  "corrup*")
```

```
## Keyword-in-context with 65 matches.
##    [text1, 27]          UK parliamentary inquiry into a | corruption |
##    [text1, 97]                found guilty of any wider | corruption |
##   [text1, 131]          campaigners who say bribery and | corruption |
##   [text1, 566]                  payments, Timeline: BAE | corruption |
##   [text3, 38]                      guilty on one of 20 | corruption |
##   [text3, 97]    country’s Prevention and Combating of | Corruption |
##   [text4, 38]                      guilty on one of 20 | corruption |
##   [text4, 97]    country’s Prevention and Combating of | Corruption |
##   [text5, 38]                      guilty on one of 20 | corruption |
##   [text5, 97]    country’s Prevention and Combating of | Corruption |
##   [text6, 18]        of his administration in tackling | corruption |
##   [text6, 90]                    he had no evidence of | corruption |
##  [text6, 119]               made a pledge to eradicate | corruption |
##   [text7, 40]    country’s Prevention and Combating of | Corruption |
##   [text8, 40]    country’s Prevention and Combating of | Corruption |
##   [text9, 40]    country’s Prevention and Combating of | Corruption |
##  [text10, 20]                payroll in a crackdown on | corruption |
## [text10, 248]       International on its perception of | corruption |
##  [text11, 16]                  the first casualty of a | corruption |
##  [text11, 75]                     statement said., The | corruption |
##  [text11, 99]                 with a promise to tackle | corruption |
## [text11, 371]            takes action over the alleged | corruption |
## [text11, 377]                          ., PM caught in | corruption |
## [text11, 387]              , Can elections help tackle | corruption |
##  [text12, 44]                    " I will vote against | corruption |
##  [text12, 51]                         .. I can’t elect |  corrupt   |
## [text12, 113] parliament where he consistently exposed | corruption |
## [text12, 140]          in on people’s feelings towards | corruption |
## [text12, 176]              ruling party ] cannot fight | corruption |
## [text12, 183]                     it is the product of | corruption |
## [text12, 193]                   elected, we will fight | corruption |
## [text12, 214]                   for its poor record on | corruption |
## [text12, 226]                  its fair share of major | corruption |
## [text12, 388]                   five years, has fought | corruption |
## [text12, 423]                  been taken to court for | corruption |
## [text12, 462]       Daniel Yona are currently battling | corruption |
## [text12, 616]      party is committed towards fighting | corruption |
## [text12, 624]                 who are accused of grand | corruption |
## [text12, 637]         has demonstrated that he tackles | corruption |
## [text12, 742]             , Ansibert Ngurumo, believes | corruption |
##  [text13, 13] ministers amid allegations of government | corruption |
## [text13, 296]       government has struggled to tackle | corruption |
## [text13, 331]  Tanzania after expressing concern about | corruption |
## [text13, 349]                         , MPs rap BAE in | corruption |
## [text13, 357]       Can elections help Tanzania tackle | corruption |
##  [text14, 69]             a high-profile casualty in a | corruption |
## [text14, 426]             took action over the alleged | corruption |
## [text14, 440]                 with a promise to tackle | corruption |
## [text14, 448]                          ., PM caught in | corruption |
## [text14, 458]              , Can elections help tackle | corruption |
##  [text15, 96]                 with a promise to tackle | corruption |
## [text15, 311]             took action over the alleged | corruption |
## [text15, 383]       Can elections help Tanzania tackle | corruption |
##   [text18, 6]       , How politicial and institutional | corruption |
##  [text19, 23]                   with a pledge to fight | corruption |
##  [text19, 41]        countries that efforts to prevent | corruption |
##  [text19, 81]     pressing for stronger action against | corruption |
##   [text20, 8]     News Arabic investigates the biggest | corruption |
##   [text21, 46]         and employees of parastatals for | corruption |
##  [text21, 179]               by the government to fight | corruption |
##    [text22, 8]     News Arabic investigates the biggest | corruption |
##   [text23, 18]             illegal rhino horn trade and | corruption |
##   [text23, 72]            of poachers and traffickers - | corruption |
##   [text23, 95]             illegal rhino horn trade and | corruption |
##   [text24, 18]             illegal rhino horn trade and | corruption |
##
##  case surrounding an air-traffic-control deal
##  charges., The controversy
##  are a brake on economic
##  probes, High Court halts
##  charges., Malinzi was
##  Bureau., Although Malinzi
##  charges., Malinzi was
##  Bureau., Although Malinzi
##  charges., Malinzi was
##  Bureau., Although Malinzi
##  ., Speaking in a
##  among top government or ruling
##  when he came to power
##  Bureau., He was
##  Bureau., He was
##  Bureau., He was
##  ., Payments to the
##  index., Many countries
##  scandal that has rocked the
##  allegations have led to donors
##  in government, but critics
##  ., PM caught in
##  row, Tanzania profile,
##  ?, News in Swahili
##  ... I can't
##  politicians," he says
##  scandals, and he is
##  , and has promised to
##  because it is the product
##  ... If elected
##  from the top."
##  ., But it has
##  scandals in the past few
##  zealously. We have amended
##  , what else do you
##  allegations in court.,
##  . Those who are accused
##  are in court, and
##  from the top,"
##  is worse now than under
##  ., He has been
##  which has adversely hampered economic
##  and the slow pace of
##  inquiry, Can elections help
##  ?, Tanzania country profile
##  scandal that has rocked Tanzania's
##  ., Mr Kikwete took
##  in government., PM
##  row, Tanzania profile,
##  ?, News in Swahili
##  in government, but critics
##  , Reuters news agency reported
##  ?, News in Swahili
##  impacts on regimes around the
##  in his country.,
##  would be strengthened as part
##  since Mr Mkapa came to
##  probe in the history of
##  and dishonesty., Opening
##  , the international community had
##  probe in the history of
##  in Zululand’s courts and its
##  is apparently fuelling the crisis
##  in Zululand’s courts and its
##  in Zululand’s courts and its
```

```r
kwic_corrup <- kwic(toks, pattern = c("corrup*", "accou*"), window = 5)
kwic_corrup
```

```
## Keyword-in-context with 79 matches.
##    [text1, 27]          UK parliamentary inquiry into a |   corruption   |
##    [text1, 64]             admitted to not keeping full |   accounting   |
##    [text1, 97]                found guilty of any wider |   corruption   |
##   [text1, 131]          campaigners who say bribery and |   corruption   |
##   [text1, 566]                  payments, Timeline: BAE |   corruption   |
##    [text3, 38]                      guilty on one of 20 |   corruption   |
##    [text3, 97]    country’s Prevention and Combating of |   Corruption   |
##    [text4, 38]                      guilty on one of 20 |   corruption   |
##    [text4, 97]    country’s Prevention and Combating of |   Corruption   |
##    [text5, 38]                      guilty on one of 20 |   corruption   |
##    [text5, 97]    country’s Prevention and Combating of |   Corruption   |
##    [text6, 18]        of his administration in tackling |   corruption   |
##    [text6, 90]                    he had no evidence of |   corruption   |
##   [text6, 119]               made a pledge to eradicate |   corruption   |
##    [text7, 40]    country’s Prevention and Combating of |   Corruption   |
##    [text8, 40]    country’s Prevention and Combating of |   Corruption   |
##    [text9, 40]    country’s Prevention and Combating of |   Corruption   |
##   [text10, 20]                payroll in a crackdown on |   corruption   |
##  [text10, 176]               government who are honest, |  accountable   |
##  [text10, 248]       International on its perception of |   corruption   |
##   [text11, 16]                  the first casualty of a |   corruption   |
##   [text11, 75]                     statement said., The |   corruption   |
##   [text11, 99]                 with a promise to tackle |   corruption   |
##  [text11, 206]                been taken from an escrow |    account     |
##  [text11, 244]                            the... escrow |    account     |
##  [text11, 274]                      saying., The escrow |    account     |
##  [text11, 371]            takes action over the alleged |   corruption   |
##  [text11, 377]                          ., PM caught in |   corruption   |
##  [text11, 387]              , Can elections help tackle |   corruption   |
##   [text12, 44]                    " I will vote against |   corruption   |
##   [text12, 51]                         .. I can't elect |    corrupt     |
##  [text12, 113] parliament where he consistently exposed |   corruption   |
##  [text12, 140]          in on people's feelings towards |   corruption   |
##  [text12, 176]              ruling party ] cannot fight |   corruption   |
##  [text12, 183]                     it is the product of |   corruption   |
##  [text12, 193]                   elected, we will fight |   corruption   |
##  [text12, 214]                   for its poor record on |   corruption   |
##  [text12, 226]                  its fair share of major |   corruption   |
##  [text12, 315]              revolution in the system of | accountability |
##  [text12, 388]                   five years, has fought |   corruption   |
##  [text12, 423]                  been taken to court for |   corruption   |
##  [text12, 462]       Daniel Yona are currently battling |   corruption   |
##  [text12, 616]      party is committed towards fighting |   corruption   |
##  [text12, 624]                 who are accused of grand |   corruption   |
##  [text12, 637]         has demonstrated that he tackles |   corruption   |
##  [text12, 742]             , Ansibert Ngurumo, believes |   corruption   |
##  [text12, 805]                money in an offshore bank |    account     |
##   [text13, 13] ministers amid allegations of government |   corruption   |
##   [text13, 42]        The inspector of the government's |    accounts    |
##  [text13, 119]            , President Kikwete said that | accountability |
##  [text13, 296]       government has struggled to tackle |   corruption   |
##  [text13, 331]  Tanzania after expressing concern about |   corruption   |
##  [text13, 349]                         , MPs rap BAE in |   corruption   |
##  [text13, 357]       Can elections help Tanzania tackle |   corruption   |
##   [text14, 69]             a high-profile casualty in a |   corruption   |
##  [text14, 270]             deposited in a personal bank |    account     |
##  [text14, 373]                been taken from an escrow |    account     |
##  [text14, 426]             took action over the alleged |   corruption   |
##  [text14, 440]                 with a promise to tackle |   corruption   |
##  [text14, 448]                          ., PM caught in |   corruption   |
##  [text14, 458]              , Can elections help tackle |   corruption   |
##   [text15, 96]                 with a promise to tackle |   corruption   |
##  [text15, 138]            session to discuss the public |    accounts    |
##  [text15, 196]                been taken from an escrow |    account     |
##  [text15, 311]             took action over the alleged |   corruption   |
##  [text15, 338]                   $ 120m from the escrow |    account     |
##  [text15, 383]       Can elections help Tanzania tackle |   corruption   |
##    [text18, 6]       , How politicial and institutional |   corruption   |
##   [text19, 23]                   with a pledge to fight |   corruption   |
##   [text19, 41]        countries that efforts to prevent |   corruption   |
##   [text19, 81]     pressing for stronger action against |   corruption   |
##    [text20, 8]     News Arabic investigates the biggest |   corruption   |
##   [text21, 46]         and employees of parastatals for |   corruption   |
##  [text21, 179]               by the government to fight |   corruption   |
##    [text22, 8]     News Arabic investigates the biggest |   corruption   |
##   [text23, 18]             illegal rhino horn trade and |   corruption   |
##   [text23, 72]            of poachers and traffickers - |   corruption   |
##   [text23, 95]             illegal rhino horn trade and |   corruption   |
##   [text24, 18]             illegal rhino horn trade and |   corruption   |
##
##  case surrounding an air-traffic-control deal
##  records of £ 8m (
##  charges., The controversy
##  are a brake on economic
##  probes, High Court halts
##  charges., Malinzi was
##  Bureau., Although Malinzi
##  charges., Malinzi was
##  Bureau., Although Malinzi
##  charges., Malinzi was
##  Bureau., Although Malinzi
##  ., Speaking in a
##  among top government or ruling
##  when he came to power
##  Bureau., He was
##  Bureau., He was
##  Bureau., He was
##  ., Payments to the
##  and hardworking. This is
##  index., Many countries
##  scandal that has rocked the
##  allegations have led to donors
##  in government, but critics
##  , paid to an energy
##  issue had not been understood
##  was held jointly by state
##  ., PM caught in
##  row, Tanzania profile,
##  ?, News in Swahili
##  ... I can't
##  politicians," he says
##  scandals, and he is
##  , and has promised to
##  because it is the product
##  ... If elected
##  from the top."
##  ., But it has
##  scandals in the past few
##  and leadership," says
##  zealously. We have amended
##  , what else do you
##  allegations in court.,
##  . Those who are accused
##  are in court, and
##  from the top,"
##  is worse now than under
##  ., But now,
##  ., He has been
##  noted the rampant misuse of
##  would be taken seriously and
##  which has adversely hampered economic
##  and the slow pace of
##  inquiry, Can elections help
##  ?, Tanzania country profile
##  scandal that has rocked Tanzania's
##  in her name,"
##  , paid to an energy
##  ., Mr Kikwete took
##  in government., PM
##  row, Tanzania profile,
##  ?, News in Swahili
##  in government, but critics
##  committee’s call for Mr Pinda’s
##  , paid to an energy
##  , Reuters news agency reported
##  held jointly by state power
##  ?, News in Swahili
##  impacts on regimes around the
##  in his country.,
##  would be strengthened as part
##  since Mr Mkapa came to
##  probe in the history of
##  and dishonesty., Opening
##  , the international community had
##  probe in the history of
##  in Zululand’s courts and its
##  is apparently fuelling the crisis
##  in Zululand’s courts and its
##  in Zululand’s courts and its
```

Another use of tokens is to create n-grams. n-grams are a useful way to create common word pairings from 2-5 sometimes but 2-3 often. This means two and three word pairs. People often combine these when the type of topic you are interested in contains many of of these. Think of scraping something around the midterms in newspaper article, you probably want n3-grams for things like Critical Race Theory, because when they are combined as three, they are distinct from when they are separate.

```
toks <- tokens(final_df$content)
toks_ngram <- tokens_ngrams(toks, n = 2:4)
head(toks_ngram[[1]], 30)
```

```
##  [1] "By_Mark"                      "Mark_DoyleBBC"
##  [3] "DoyleBBC_International"    "International_Development"
##  [5] "Development_Correspondent" "Correspondent_,"
##  [7] ",_The"                        "The_British"
##  [9] "British_arms"                 "arms_and"
## [11] "and_aircraft"                 "aircraft_firm"
## [13] "firm_BAE"                     "BAE_Systems"
## [15] "Systems_has"                  "has_been"
## [17] "been_severely"                "severely_criticised"
## [19] "criticised_by"                "by_a"
## [21] "a_UK"                         "UK_parliamentary"
## [23] "parliamentary_inquiry"        "inquiry_into"
## [25] "into_a"                       "a_corruption"
## [27] "corruption_case"              "case_surrounding"
## [29] "surrounding_an"               "an_air-traffic-control"
```

```
# the word() function from string is potentially a more helpful route. An example of why regular expres
word(string = corp_corruption, start = 1, end = 20, sep = fixed(" "))
```

```
##  [1] "By Mark DoyleBBC International Development Correspondent ,The British arms and aircraft firm B
##  [2] ",Mike Corey is off the coast of Tanzania, exploring a coral reef that scientists say could be a
##  [3] ",Last updated on 12 December 201912 December 2019.From the section African,Former Tanzanian Foo
##  [4] ",Last updated on 12 December 201912 December 2019.From the section African,Former Tanzanian Foo
##  [5] ",Last updated on 12 December 201912 December 2019.From the section African,Former Tanzanian Foo
##  [6] "The President of Tanzania, Benjamin Mkapa, has defended the work of his administration in tackl
##  [7] ",Last updated on 29 June 201729 June 2017.From the section Football,The president of the Tanzan
##  [8] ",Last updated on 29 June 201729 June 2017.From the section Football,The president of the Tanzan
##  [9] ",Last updated on 29 June 201729 June 2017.From the section Football,The president of the Tanzan
## [10] "Tanzania has removed more than 10,000 \"ghost workers\" from its public sector payroll in a cra
## [11] "Tanzania's Attorney General Frederick Werema has resigned, making him the first casualty of a c
## [12] "By Emmanuel MugaBBC News, Dar es Salaam,As he gets set to vote for the first time in his life r
## [13] "Tanzania's President Jakaya Kikwete has sacked six ministers amid allegations of government cor
## [14] "Tanzania's President Jakaya Kikwete has fired a senior government minister accused of wrongly
## [15] "Tanzania's prime minister is under pressure to resign over alleged fraudulent payments worth $
## [16] ",Prayers and herbal remedies were the solution Tanzanian President John Magufuli prescribed for
## [17] ",The Children in Crossfire director takes Stephen Travers to see conditions in the African cour
## [18] ",How politicial and institutional corruption impacts on regimes around the world,\n         "
## [19] ",\n\n\t,\n\t,The Tanzanian President, Benjamin Mkapa, has opened a conference of donor countri
## [20] ",BBC News Arabic investigates the biggest corruption probe in the history of Lebanon's energy s
## [21] ",\n\n\t,\n\t,\n\nDodoma: The chairman of the [ruling] Chama cha Mapinduzi [CCM] party, Presiden
## [22] ",BBC News Arabic investigates the biggest corruption probe in the history of Lebanon's energy s
## [23] ",BBC Africa Correspondent Alastair Leithead investigates the alleged links between the illegal
## [24] ",BBC Africa correspondent Alastair Leithead investigates the alleged links between the illegal
```