Movie Script Keyness

Marie Rivers

2022-05-12

```r
library(tidyr) #text analysis in R
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```r
library(pdftools)
```

```
## Warning: package 'pdftools' was built under R version 4.1.2
```

```
## Using poppler version 22.02.0
```

```r
library(lubridate) #working with date data
```

```
##
## Attaching package: 'lubridate'
```

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

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.7     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
library(tidytext)
library(readr)
library(quanteda)
```

```
## Warning: package ’quanteda’ was built under R version 4.1.2
```

```
## Package version: 3.2.1
## Unicode version: 13.0
## ICU version: 69.1
```

```
## Parallel computing: 8 of 8 threads used.
```

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

```
library(readtext) #quanteda subpackage for reading pdf
library(quanteda.textstats)
library(quanteda.textplots)
```

```
## Warning: package ’quanteda.textplots’ was built under R version 4.1.2
```

```
library(ggplot2)
library(forcats)
library(stringr)
library(quanteda.textplots)
library(widyr)# pairwise correlations
library(igraph) #network plots
```

```
## Warning: package ’igraph’ was built under R version 4.1.2
```

```
##
## Attaching package: ’igraph’
```

```
## The following object is masked from ’package:quanteda.textplots’:
##
##     as.igraph
```

```
## The following objects are masked from ’package:dplyr’:
##
##     as_data_frame, groups, union
## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following objects are masked from 'package:lubridate':
##
##     %--%, union

## The following object is masked from 'package:tidyr':
##
##     crossing

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
library(ggraph)
library(here)
```

```
## here() starts at /Users/marierivers/Documents/UCSB_Environmental_Data_Science/EDS_231_Text_and_Senti
```

```
library(patchwork)
```

```
files <- list.files(path = here("data"),
                    pattern = "pdf$", full.names = TRUE)

scripts <- lapply(files, pdf_text)

scripts_pdf <- readtext(file = here("data", "*.pdf"),
                   docvarsfrom = "filenames",
                   docvarnames = c("title1", "title2", "title3"),
                   sep = NULL) # this isn't doing what I want it to do
#creating an initial corpus containing our data
scripts_corp <- corpus(x = scripts_pdf, text_field = "text" )
summary(scripts_corp) %>%
  knitr::kable(caption = "Summary of Scripts Corpus")
```

Table 1: Summary of Scripts Corpus

Text	Types	Tokens	Sentences	title1	title2	title3
an_inconvenient_truth.pdf	2245	10936	685	an	inconvenient	truth

Text	Types	Tokens	Sentences	title1	title2	title3
before_the_flood.pdf	2540	13634	863	before	the	flood
dont_look_up.pdf	4620	28016	2825	dont	look	up

```r
# Add some additional, context-specific stop words to stop word lexicon
more_stops <-c("randall", "kate", "dr", "president", "int", "oglethorpe", "jason", "brie", "orlean")
add_stops <- tibble(word = c(stop_words$word, more_stops))
stop_vec <- as_vector(add_stops)
```

xxx…look up code to remove numbers

Create different data objects that will be used for the subsequent analyses

```r
#convert to tidy format and apply my stop words
raw_text <- tidy(scripts_corp)

#Distribution of most frequent words across documents
raw_words <- raw_text %>%
  mutate(title = as.factor(title1)) %>%
  mutate(title = case_when(title == "dont" ~ "dont_look_up",
                           title == "an" ~ "an_inconvenient_truth",
                           title == "before" ~ "before_the_flood")) %>%
  unnest_tokens(word, text) %>%
  anti_join(add_stops, by = 'word') %>%
  count(title, word, sort = TRUE)
```

```r
#number of total words by document
total_words <- raw_words %>%
  group_by(title) %>%
  summarize(total = sum(n))

script_words <- left_join(raw_words, total_words)
```

```
## Joining, by = "title"
```

```r
par_tokens <- unnest_tokens(raw_text, output = paragraphs, input = text, token = "paragraphs")

par_tokens <- par_tokens %>%
 mutate(par_id = 1:n())

par_words <- unnest_tokens(par_tokens, output = word, input = paragraphs, token = "words") %>%
  mutate(title = case_when(title1 == "dont" ~ "dont_look_up",
                           title1 == "an" ~ "an_inconvenient_truth",
                           title1 == "before" ~ "before_the_flood"))
```

```r
tokens <- tokens(scripts_corp, remove_punct = TRUE)
toks1<- tokens_select(tokens, min_nchar = 3)
toks1 <- tokens_tolower(toks1)
toks1 <- tokens_remove(toks1, pattern = (stop_vec))
dfm <- dfm(toks1)

dfm$full_title <- c("an_inconvenient_truth", "before_the_flood", "dont_look_up")
docvars(dfm)
```

```
##   title1       title2 title3            full_title
## 1     an inconvenient  truth an_inconvenient_truth
## 2 before          the  flood      before_the_flood
## 3   dont         look     up          dont_look_up
```

```
par_words_inconvenient_truth <- par_words %>%
  filter(title == "an_inconvenient_truth")

par_words_before_the_flood <- par_words %>%
  filter(title == "before_the_flood")

par_words_dont_look_up <- par_words %>%
  filter(title == "dont_look_up")
```

```
word_cors_all <- par_words %>%
  add_count(par_id) %>%
  filter(n >= 50) %>%
  select(-n) %>%
  pairwise_cor(word, par_id, sort = TRUE)

word_cors_inconvenient_truth <- par_words_inconvenient_truth %>%
  add_count(par_id) %>%
  filter(n >= 50) %>%
  select(-n) %>%
  pairwise_cor(word, par_id, sort = TRUE)

word_cors_before_the_flood <- par_words_before_the_flood %>%
  add_count(par_id) %>%
  filter(n >= 50) %>%
  select(-n) %>%
  pairwise_cor(word, par_id, sort = TRUE)

word_cors_dont_look_up <- par_words_dont_look_up %>%
  add_count(par_id) %>%
  filter(n >= 50) %>%
  select(-n) %>%
  pairwise_cor(word, par_id, sort = TRUE)
```

```
dfm
```

```
## Document-feature matrix of: 3 documents, 5,042 features (56.61% sparse) and 4 docvars.
##                              features
## docs                           inconvenient truth transcript
##   an_inconvenient_truth.pdf            2     4          1
##   before_the_flood.pdf                 0     2          0
##   dont_look_up.pdf                     0     7          0
##                              features
## docs                           http://forumpolitics.com/blogs/2007/03/17/an-inconvient-truth-transcript
##   an_inconvenient_truth.pdf                                                                          1
##   before_the_flood.pdf                                                                               0
##   dont_look_up.pdf                                                                                   0
##                              features
## docs                           march 2007 introduction river gently flowing
##   an_inconvenient_truth.pdf     1    1            1     1      1       1
##   before_the_flood.pdf          1    0            0     1      0       0
##   dont_look_up.pdf              0    0            0     0      0       0
## [ reached max_nfeat ... 5,032 more features ]
```

```
#first the basic frequency stat
tstat_freq <- textstat_frequency(dfm, n = 5, groups = title1)
head(tstat_freq, 15) %>%
  knitr::kable(caption = "Subset of Top 5 Words")
```

Table 2: Subset of Top 5 Words

feature	frequency	rank	docfreq	group
ice	54	1	1	an
earth	32	2	1	an
time	32	2	1	an
warming	28	4	1	an
world	27	5	1	an
climate	69	1	1	before
page	63	2	1	before
change	52	3	1	before
people	48	4	1	before
world	38	5	1	before
comet	82	1	1	dont
time	78	2	1	dont
cont'd	76	3	1	dont
ext	70	4	1	dont
isherwell	59	5	1	dont

```
#let's zoom in on just one of our key terms
all_script_cors <- word_cors_all %>%
  filter(item1 == "climate") %>%
  mutate(n = 1:n())

all_script_cors_plot <- all_script_cors  %>%
  filter(n <= 30) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = correlation, edge_width = correlation), edge_colour = "steelblue3") +
  geom_node_point(size = 4) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```

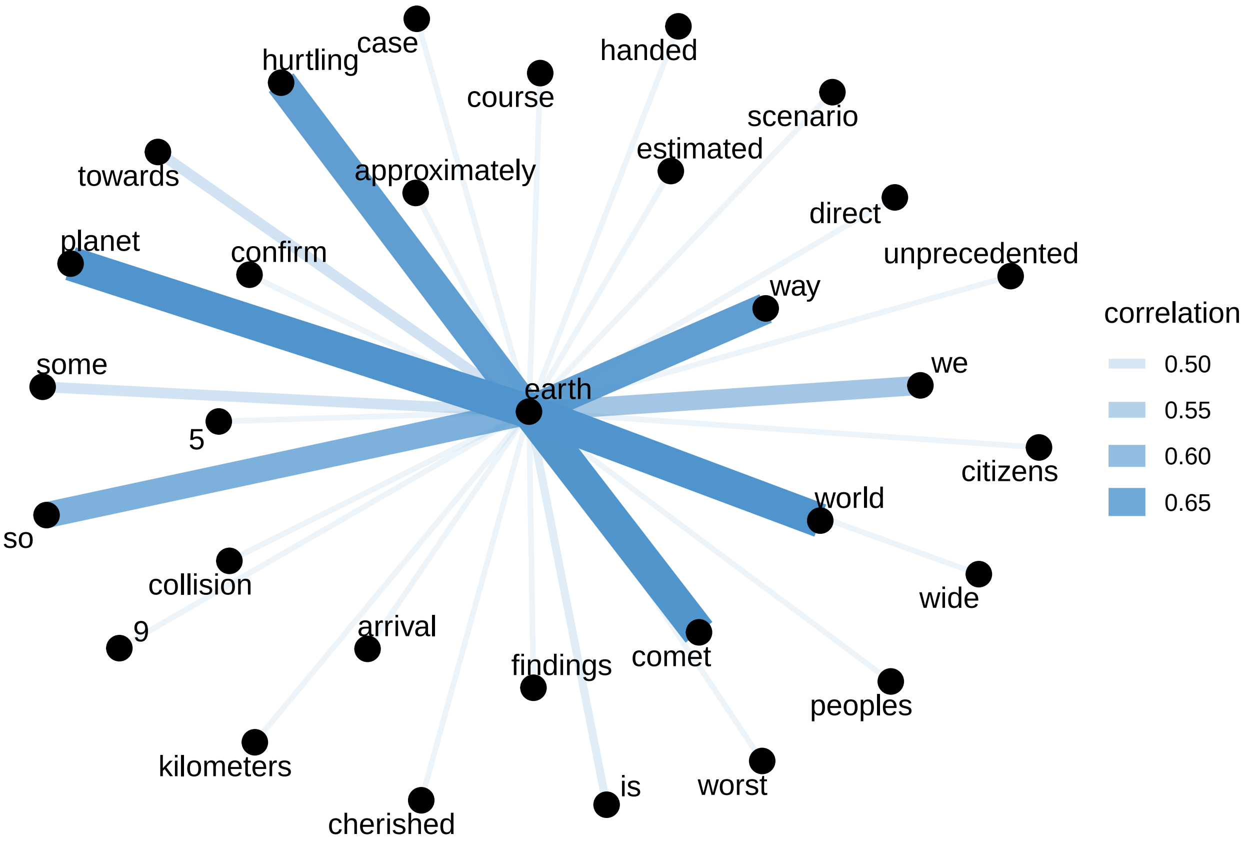
you can compare the use of "climate" in the 2 documentations to see how this word is used differently, but don't look up never mentions the word climate or climate change

```
#let's zoom in on just one of our key terms
inconvenient_truth_cors <- word_cors_inconvenient_truth %>%
  filter(item1 == "earth") %>%
  mutate(n = 1:n())
```

```r
inconvenient_truth_cors_plot <- inconvenient_truth_cors %>%
  filter(n <= 30) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = correlation, edge_width = correlation), edge_colour = "steelblue3") +
  geom_node_point(size = 4) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```

```r
#let's zoom in on just one of our key terms
before_the_flood_cors <- word_cors_before_the_flood %>%
  filter(item1 == "earth") %>%
  mutate(n = 1:n())

before_the_flood_cors_plot <- before_the_flood_cors %>%
  filter(n <= 30) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = correlation, edge_width = correlation), edge_colour = "steelblue3") +
  geom_node_point(size = 4) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```

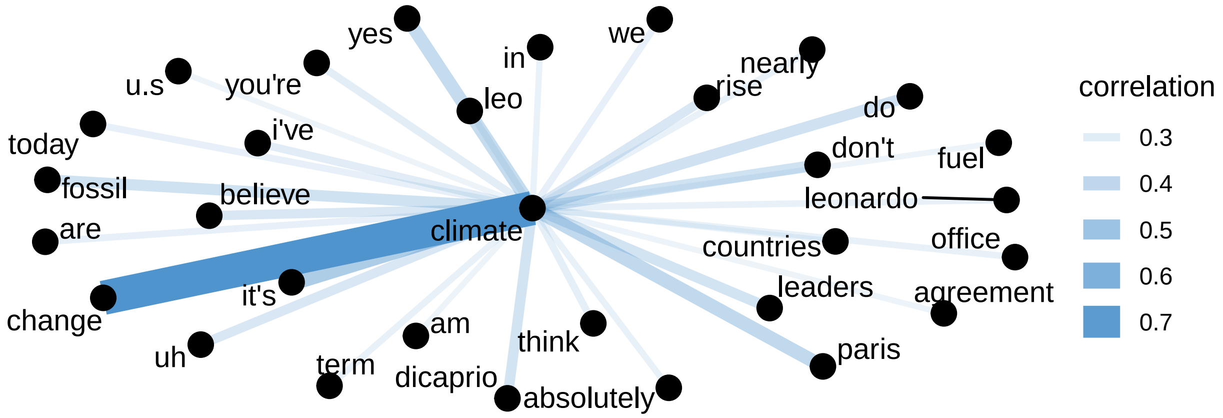
```r
# don't look up never mentions climate or climate change
#let's zoom in on just one of our key terms
dont_look_up_cors <-  word_cors_dont_look_up %>%
  filter(item1 == "earth") %>%
  mutate(n = 1:n())

dont_look_up_cors  %>%
  filter(n <= 30) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = correlation, edge_width = correlation), edge_colour = "steelblue3") +
  geom_node_point(size = 4) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```

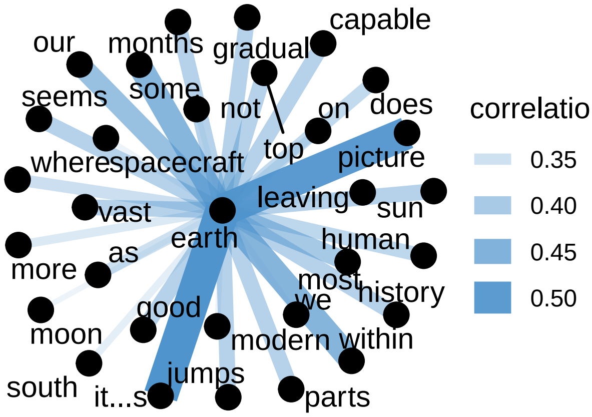


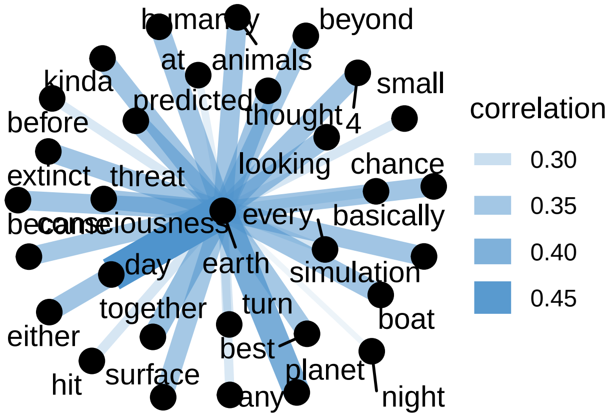
xxx…revise to have climate and earth separated

```
cors_plots <- all_script_cors_plot / (inconvenient_truth_cors_plot + before_the_flood_cors_plot)
cors_plots
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>
```

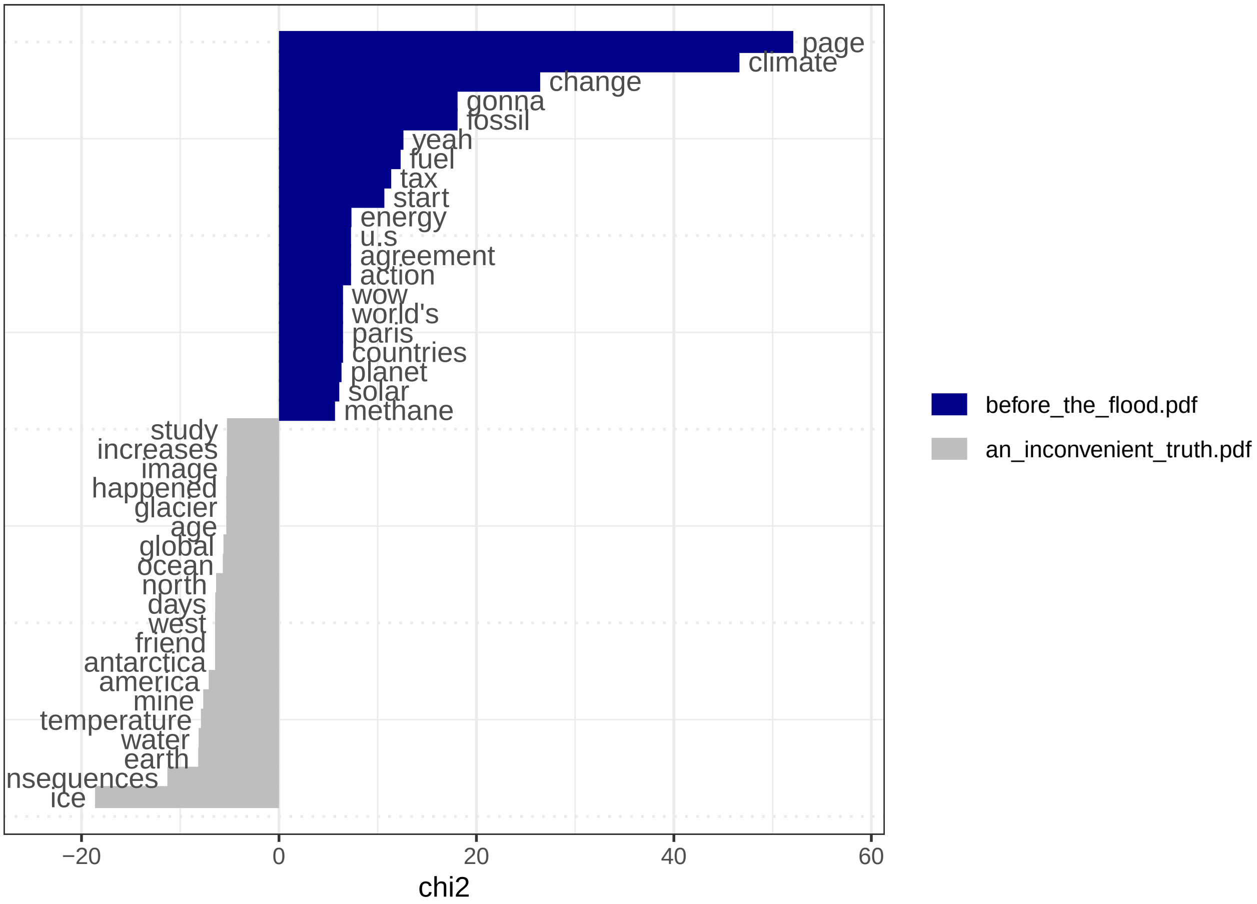




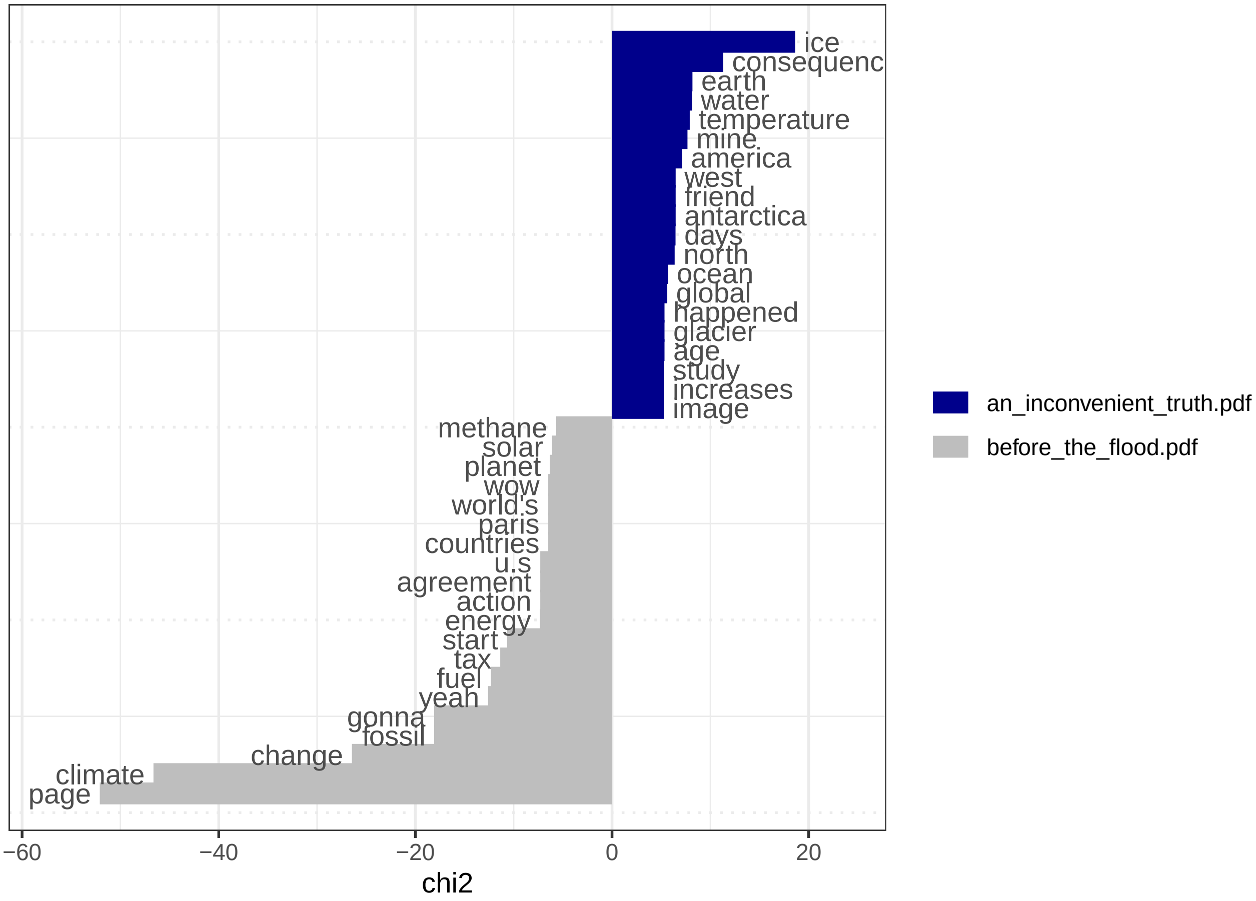


```r
keyness_function <- function(reference_report_title, target_report_title) {
  files <- list.files(path = here("data"),
                      pattern = "pdf$", full.names = TRUE)
  scripts <- lapply(files, pdf_text)
  scripts_pdf <- readtext(file = here("data", "*.pdf"),
                    docvarsfrom = "filenames",
                    docvarnames = c("title1", "title2", "title3"),
                    sep = "_")
  scripts_corp <- corpus(x = scripts_pdf, text_field = "text" )
  tokens <- tokens(scripts_corp, remove_punct = TRUE)
  toks1<- tokens_select(tokens, min_nchar = 3)
  toks1 <- tokens_tolower(toks1)
  toks1 <- tokens_remove(toks1, pattern = (stop_vec))
  dfm <- dfm(toks1)
  dfm$full_title <- c("an_inconvenient_truth", "before_the_flood", "dont_look_up")

  keyness_function_plot <- dfm %>%
    dfm_subset(full_title %in% c(reference_report_title, target_report_title)) %>%
    textstat_keyness(target = paste0(target_report_title, ".pdf")) %>%
    textplot_keyness()
  keyness_function_plot
}
```

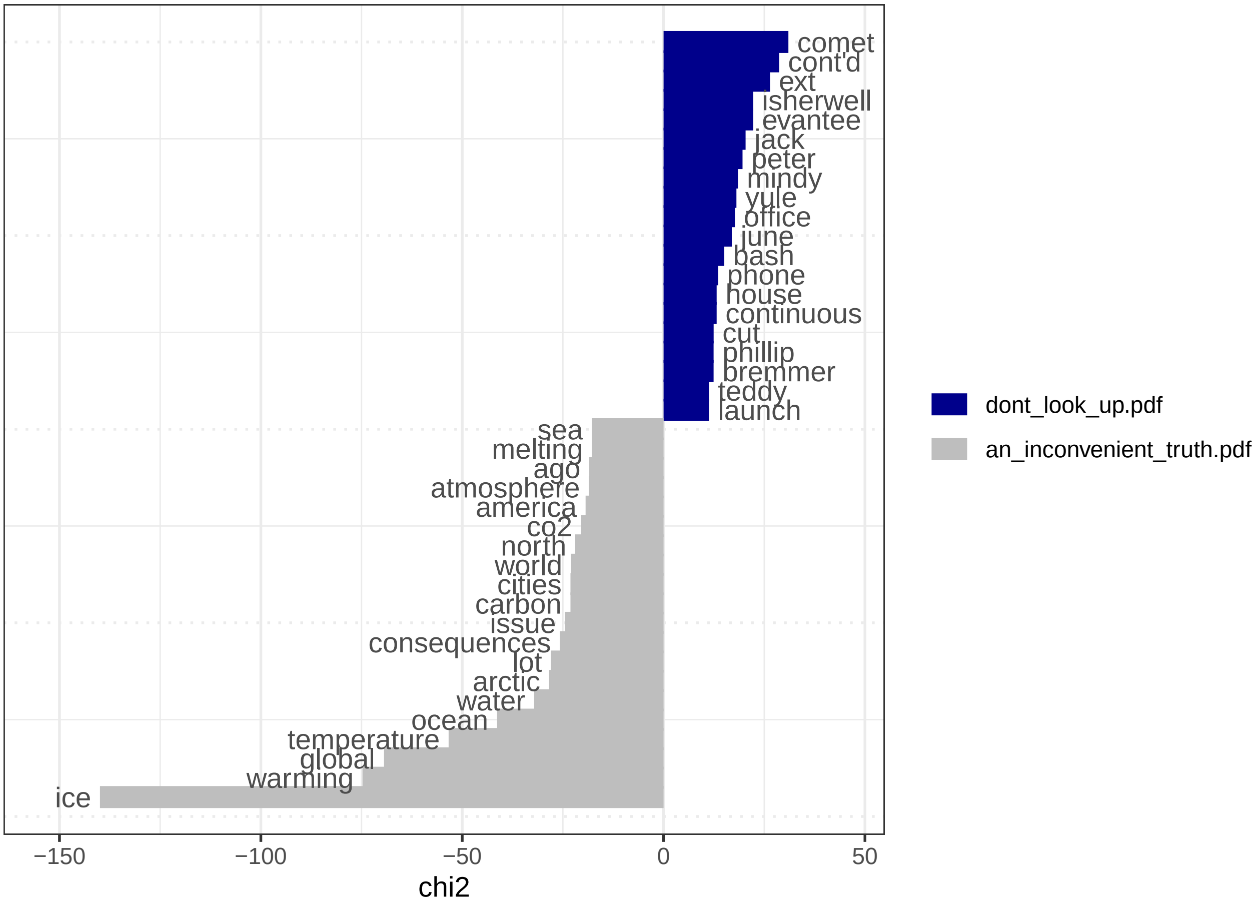
```r
# an_inconvenient_truth vs. before_the_flood
keyness_function(reference_report_title = "an_inconvenient_truth", target_report_title = "before_the_fl
```



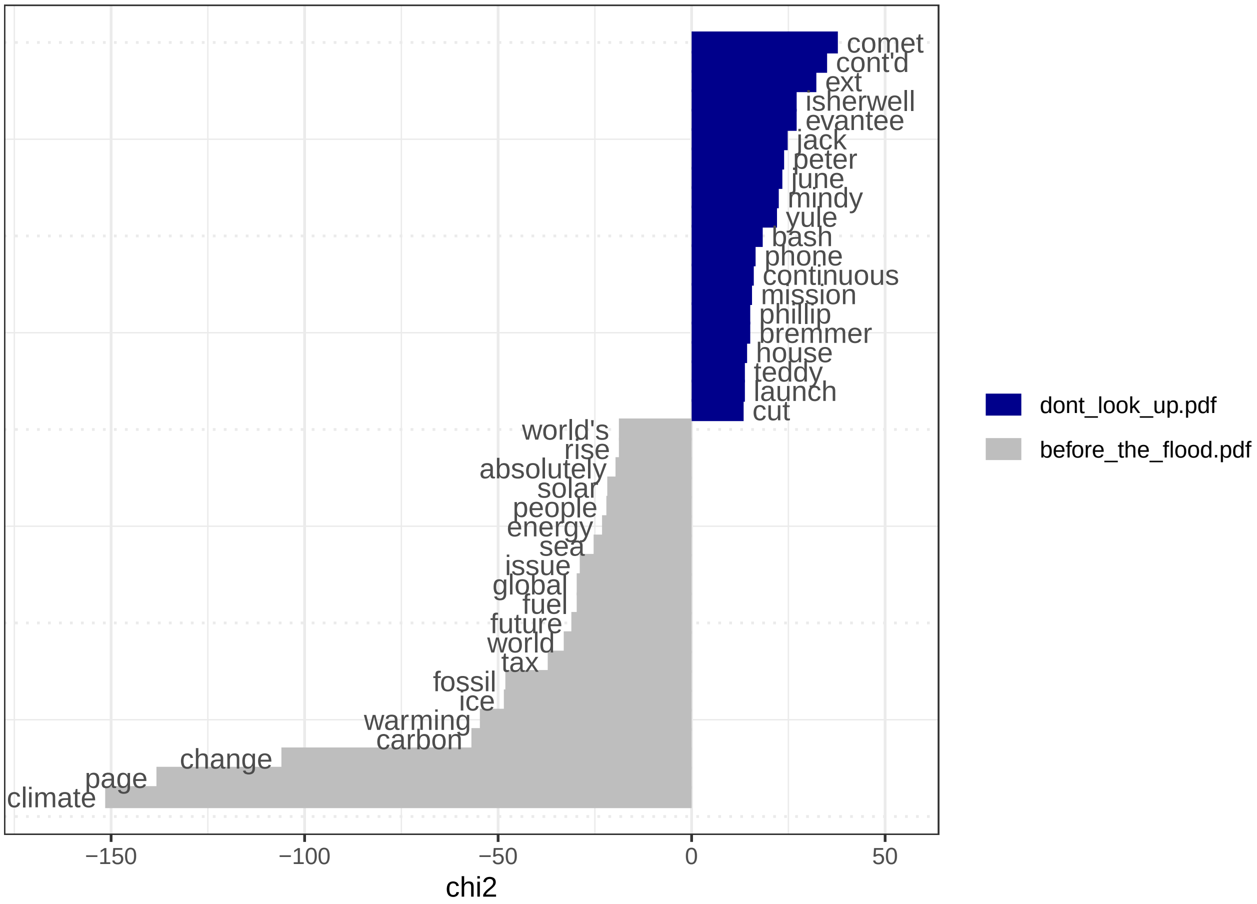
```
# before_the_flood vs. an_inconvenient_truth
keyness_function(reference_report_title = "before_the_flood", target_report_title = "an_inconvenient_tru
```



```r
# an_inconvenient_truth vs. don't look up
keyness_function(reference_report_title = "an_inconvenient_truth", target_report_title = "dont_look_up"
```



```
# an_inconvenient_truth vs. don't look up
keyness_function(reference_report_title = "before_the_flood", target_report_title = "dont_look_up")
```


comet
cont'd
ext
isherwell
evantee
jack
peter
june
mindy
yule
bash
phone
continuous
mission
phillip
bremmer
house
teddy
launch
cut
world's
rise
absolutely
solar
people
energy
sea
issue
global
fuel
future
world
tax
fossil
ice
warming
carbon
change
page
climate
−150
−100
−50
0
50
chi2
dont_look_up.pdf
before_the_flood.pdf