Keywords-in-Context and Dictionaries

Tom Paskhalis

29 November, 2018

```
library("readr")
library("dplyr")
library("stringr")
library("lubridate")
library("stopwords")
library("quanteda")
```

Reading in the file

We use the Congressional Record date from the current (115th Congress). There are 4 datasets, with one for each year and each chamber. All are stored in the `data` folder and comressed with gzip to save space.

```
# House of Representatives
us_house_2017 <- readr::read_csv("../data/us-house-2017.csv.gz")
us_house_2018 <- readr::read_csv("../data/us-house-2018.csv.gz")
# Senate
us_senate_2017 <- readr::read_csv("../data/us-senate-2017.csv.gz")
us_senate_2018 <- readr::read_csv("../data/us-senate-2018.csv.gz")
```

While the Senate speeches comes with many additional covariates, House data is more limited in this regard. However, we can extract date from the API field `granuleId`, using this, rather convoluted, regular expression.

```
pattern <- "20\\d\\d-(0[1-9]|1[012])-(0[1-9]|[12][0-9]|3[01])"
us_house_2017 <- us_house_2017 %>%
    dplyr::mutate(date = lubridate::ymd(stringr::str_extract(granuleId, pattern)))
us_house_2018 <- us_house_2018 %>%
    dplyr::mutate(date = lubridate::ymd(stringr::str_extract(granuleId, pattern)))
```

Create Corpus

First, we combine the datasets for two chambers (House of Representatives and Senate) and two years (2017 and 2018) into one. Then, we can create a corpus by simply passing the resultant data frame to `corpus()` function in `quanteda`. The variable `text` is automatically recognised as the one containing the speeches and all the other variables are treated as `docvars`.

```
congress115 <- dplyr::select(us_house_2017, date, chamber, speaker, text) %>%
  dplyr::bind_rows(dplyr::select(us_house_2018, date, chamber, speaker, text)) %>%
  dplyr::bind_rows(dplyr::select(us_senate_2017, date, chamber, speaker, text)) %>%
  dplyr::bind_rows(dplyr::select(us_senate_2018, date, chamber, speaker, text))

head(congress115, 10)
```

```
## # A tibble: 10 x 4
##    date       chamber speaker           text
##    <date>     <chr>   <chr>             <chr>
##  1 2017-01-03 H       The CLERK         The Representatives-elect and th~
##  2 2017-01-03 H       The CLERK         As directed by law, the Clerk of~
```

```
##  3 2017-01-03 H       The CLERK          Four hundred thirty-four Represe~
##  4 2017-01-03 H       The CLERK          Pursuant to law and precedent, t~
##  5 2017-01-03 H       Mrs McMORRIS RODG~ Whether you are from the Evergre~
##  6 2017-01-03 H       The CLERK          The Clerk now recognizes the gen~
##  7 2017-01-03 H       Mr CROWLEY         Madam Clerk, first I would like ~
##  8 2017-01-03 H       The CLERK          The names of the Honorable Paul ~
##  9 2017-01-03 H       Ms PLASKETT durin~ Madam Clerk, parliamentary inqui~
## 10 2017-01-03 H       The CLERK          The gentlewoman will state her p~
```

```
corpus115 <- quanteda::corpus(congress115)
head(docvars(corpus115), 10)
```

```
##              date chamber                          speaker
## text1  2017-01-03       H                        The CLERK
## text2  2017-01-03       H                        The CLERK
## text3  2017-01-03       H                        The CLERK
## text4  2017-01-03       H                        The CLERK
## text5  2017-01-03       H             Mrs McMORRIS RODGERS
## text6  2017-01-03       H                        The CLERK
## text7  2017-01-03       H                       Mr CROWLEY
## text8  2017-01-03       H                        The CLERK
## text9  2017-01-03       H Ms PLASKETT during the roll call
## text10 2017-01-03       H                        The CLERK
```

We can get some basic summary statistics by applying `summary()` on the corpus.

```
summary(corpus115, 10)
```

```
## Corpus consisting of 85480 documents, showing 10 documents:
##
##    Text Types Tokens Sentences       date chamber
##   text1    44     63         2 2017-01-03       H
##   text2   587    757        13 2017-01-03       H
##   text3    13     14         2 2017-01-03       H
##   text4    35     47         3 2017-01-03       H
##   text5   271    620        23 2017-01-03       H
##   text6    14     15         1 2017-01-03       H
##   text7   214    436        13 2017-01-03       H
##   text8   566    976        17 2017-01-03       H
##   text9     6      6         1 2017-01-03       H
##  text10     8      8         1 2017-01-03       H
##                          speaker
##                        The CLERK
##                        The CLERK
##                        The CLERK
##                        The CLERK
##              Mrs McMORRIS RODGERS
##                        The CLERK
##                       Mr CROWLEY
##                        The CLERK
##  Ms PLASKETT during the roll call
##                        The CLERK
##
## Source: /home/tpaskhalis/Decrypted/Git/VAM_Text_Analysis/code/* on x86_64 by tpaskhalis
## Created: Wed Nov 28 23:17:42 2018
## Notes:
```

Document-frequency matrix and summary statistics

We are removing stopwords pre-specified in the `stopwords()` function. For more details check the associated package `stopwords`.

```
stopwords::stopwords("english")
```

```
##   [1] "i"          "me"         "my"         "myself"     "we"
##   [6] "our"        "ours"       "ourselves"  "you"        "your"
##  [11] "yours"      "yourself"   "yourselves" "he"         "him"
##  [16] "his"        "himself"    "she"        "her"        "hers"
##  [21] "herself"    "it"         "its"        "itself"     "they"
##  [26] "them"       "their"      "theirs"     "themselves" "what"
##  [31] "which"      "who"        "whom"       "this"       "that"
##  [36] "these"      "those"      "am"         "is"         "are"
##  [41] "was"        "were"       "be"         "been"       "being"
##  [46] "have"       "has"        "had"        "having"     "do"
##  [51] "does"       "did"        "doing"      "would"      "should"
##  [56] "could"      "ought"      "i'm"        "you're"     "he's"
##  [61] "she's"      "it's"       "we're"      "they're"    "i've"
##  [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
##  [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
##  [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
##  [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
##  [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
##  [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
##  [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
## [101] "who's"      "what's"     "here's"     "there's"    "when's"
## [106] "where's"    "why's"      "how's"      "a"          "an"
## [111] "the"        "and"        "but"        "if"         "or"
## [116] "because"    "as"         "until"      "while"      "of"
## [121] "at"         "by"         "for"        "with"       "about"
## [126] "against"    "between"    "into"       "through"    "during"
## [131] "before"     "after"      "above"      "below"      "to"
## [136] "from"       "up"         "down"       "in"         "out"
## [141] "on"         "off"        "over"       "under"      "again"
## [146] "further"    "then"       "once"       "here"       "there"
## [151] "when"       "where"      "why"        "how"        "all"
## [156] "any"        "both"       "each"       "few"        "more"
## [161] "most"       "other"      "some"       "such"       "no"
## [166] "nor"        "not"        "only"       "own"        "same"
## [171] "so"         "than"       "too"        "very"       "will"
```

To create a document-frequency matrix, we will use `dfm()` function. Many of the parameters specified (such as `tolowe` and `stem`) below are the defaults, but it is often a good idea to be explicit about document pre-processing, as it starts gradually getting more attention in the text analysis literature.

```
dfm115 <- quanteda::dfm(corpus115,
                        tolower = TRUE,
                        stem = FALSE,
                        remove = stopwords("english"),
                        remove_punct = TRUE)
```

We can also group by chamber when creating a `dfm`.

```
grouped115 <- quanteda::dfm(corpus115,
                            tolower = TRUE,
                            stem = FALSE,
                            remove = stopwords("english"),
                            remove_punct = TRUE,
                            group = "chamber")
```

To see the most frequently used terms we use the `topfeatures()` function.

```
quanteda::topfeatures(dfm115, 50)
```

```
##        mr       act      bill    people   section      time   speaker
##    130798     82849     77991     69864     63667     62777     62150
## president    states     shall         1         b    united committee
##     61723     61142     58323     55531     50362     49069     48565
##         2       one     state       can    senate      year     house
##     48245     47879     45905     45320     43806     42563     41992
##       may     years   federal      just   support       new       law
##     41814     39695     38567     38251     38065     37821     37562
##    health secretary  american  national amendment   country        us
##     36821     36099     35688     35629     34858     34742     34069
##  congress       now      work     going   program       tax      also
##     34037     33395     33060     32491     32479     32311     32011
##      make      many  security     today      care      want    public
##     31795     31124     30981     30563     30408     30215     30154
##      vote
##     30058
```

We can also use `textstat_keyness()` function to compare how words are used across groups of documents. Here we are treating the Senate as a target group and House as the reference.

```
keyness <- quanteda::textstat_keyness(grouped115, target = "S")

head(keyness, 10)
```

```
##        feature      chi2 p n_target n_reference
## 1      senator 28036.512 0    24734        1650
## 2    president 27219.994 0    46650       15073
## 3       senate 22707.414 0    34326        9480
## 4   nomination 12278.465 0     9703         217
## 5        judge 10650.562 0    11509        1687
## 6     senators  7123.779 0     6256         403
## 7         call  5710.659 0     9828        3180
## 8  nominations  5417.783 0     4219          69
## 9      nominee  5214.335 0     4404         215
## 10       clerk  5159.416 0     9786        3512
```

```
tail(keyness, 10)
```

```
##          feature       chi2 p n_target n_reference
## 145059         b  -6762.598 0    12430       37932
## 145060  provided  -7020.434 0     3379       19006
## 145061 available  -7295.297 0     2892       18205
## 145062       h.r  -8771.948 0     2027       18047
## 145063     chair  -9841.184 0     2015       19518
## 145064   section  -9848.412 0    14874       48793
```

```
## 145065     shall -11566.537 0    12125       46198
## 145066 gentleman -18166.782 0      131       24736
## 145067        mr -20275.352 0    30582      100216
## 145068   speaker -44789.684 0      535       61615
```

Keywords-in-context

The idea to inspect the terms of interest within the smaller window of words surrounding it was one of the first to emerge in automatic analysis of text. Here, we will focus on a few issues that polarised US politics in the past two years. In order to do that we will use `kwic()` function and `textplot_xray()` for the graphical represenation of the results. Note that we need to apply `kwic()` to a corpus, rather than a dfm.

Let us start with the **Deferred Action for Childhood Arrivals** (also known as DACA), the immigration policy that has been subject of much debate under Trump's administration.

```
daca <- quanteda::kwic(corpus115, "daca", window = 5, valuetype = "fixed")

head(daca, 50)
```

```
##
##    [text1291, 665]              people who signed up for | DACA |
##    [text1829, 163]         promised to end the executive | DACA |
##    [text3163, 361]           also promised to remove the | DACA |
##   [text3918, 2058]            What is this proposal with | DACA |
##   [text3918, 2233]                    If you look at the | DACA |
##   [text3918, 2400]           this level of amnesty under | DACA |
##   [text3918, 2404]           under DACA. The President's | DACA |
##   [text3918, 2450]           weeks before he issued this | DACA |
##   [text3918, 2702]          almost certainly be sued for | DACA |
##   [text3918, 3006]               both of those policies, | DACA |
##   [text3918, 3654]                     in place now. But | DACA |
##   [text3918, 3833]             of amnesty. That includes | DACA |
##   [text3918, 3874]   his administration he would address | DACA |
##   [text3918, 3931] Immigration Services is still issuing | DACA |
##   [text3918, 3936]      DACA permits and still extending | DACA |
##   [text3918, 3987]               to freeze any action on | DACA |
##   [text3918, 4000]  executive order and invalidate every | DACA |
##   [text3918, 4257]       encouragement: the earlier that | DACA |
##  [text3918, 14010]             United States, let's end | DACA |
##    [text4450, 214]         many hundreds of thousands of | DACA |
##    [text4555, 104]                 for our DREAMers, for | DACA |
##     [text4715, 90]            Roque is a beneficiary of | DACA |
##    [text4715, 140]        our country has benefited from | DACA |
##    [text4715, 330]             the community. Because of | DACA |
##    [text4718, 124]                      math. Eliel is a | DACA |
##    [text4718, 134]                   repeat. Eliel is a | DACA |
##    [text4718, 170]     and hundreds of other hardworking | DACA |
##    [text4718, 197]                of our great country. | DACA |
##    [text4718, 223]                thing. Let's give our | DACA |
##     [text5910, 23]           Arrivals, commonly known as | DACA |
##     [text5910, 25]               commonly known as DACA. | DACA |
##     [text5910, 84]             . Taking any step against | DACA |
##     [text5910, 89]              DACA would not only hurt | DACA |
##   [text7757, 1162]                issued the order, the | DACA |
```

```
##   [text7757, 1248]            have the authority for the | DACA |
##   [text7757, 1305]                law. Subsequent to the | DACA |
##   [text9352, 2708]       Childhood Arrivals program, the | DACA |
##   [text11442, 648]            for Childhood Arrivals, or | DACA |
##   [text11442, 801]              for the DREAMers and the | DACA |
##   [text11442, 822]              like Texas, with 200,000 | DACA |
##  [text13710, 6456]               by Barack Obama in his | DACA |
##    [text14196, 60]              people who signed up for | DACA |
##   [text14196, 573]                  great? Then there is | DACA |
##   [text14196, 607]        Secretary Kelly says he thinks | DACA |
##   [text14448, 383]               would be used to enlist | DACA |
##   [text14448, 418]           our military, including our | DACA |
##   [text14448, 484]                  he began to push the | DACA |
##   [text14448, 491]                  through there. Well, | DACA |
##  [text14448, 1197]                    put an end to this | DACA |
##  [text14448, 3744]            principles? Vote down this | DACA |
##
##  . With the BRIDGE Act
##  action and potentially deport those
##  program. For this reason
##  and DAPA that President Obama
##  language that has been advanced
##  . The President's DACA acronym
##  acronym stands for Deferred Action
##  policy, he stood over
##  and later on for DAPA
##  and DAPA, are clearly
##  , the Deferred Action for
##  and DAPA. It needs
##  and DAPA and the Morton
##  permits and still extending DACA
##  permits. That is a
##  and DAPA. I would
##  permit and every DAPA permit
##  and DAPA are addressed by
##  , let's end DAPA,
##  children who need relief,
##  , and for women's rights
##  . He is a DREAMer
##  . As a teacher,
##  , hundreds of kids are
##  student. Let me repeat
##  student. He and his
##  students stay in America,
##  students are our new Americans
##  students and other hardworking taxpayers
##  . DACA recipients were brought
##  recipients were brought here to
##  would not only hurt DACA
##  recipients, it would hurt
##  order-- two of
##  program, and he said
##  order going out, President
##  program. We preserve funding
```

```
##  , program implemented under President
##  beneficiaries that is most urgent
##  beneficiaries out of 700,000 nationwide
##  program-- Deferred Action
##  and the hundreds of thousands
##  , the program where 800,000
##  is illegal, and,
##  aliens-- Deferred Action
##  personnel, into the United
##  recipients through there. Well
##  is unconstitutional. The Deferred
##  program. This Congress,
##  thing that rewards lawbreakers,
```

Interestingly, DACA is frequently mentioned in the context of other executive act, **Deferred Action for Parents of Americans and Lawful Permanent Residents (DAPA)**, that extended DACA to the parents of the 'Dreamers'. Let us now explore its context:

```
dapa <- quanteda::kwic(corpus115, "dapa", window = 5, valuetype = "fixed")

head(dapa, 50)
```

```
##
##    [text3918, 2060]                   this proposal with DACA and | DAPA |
##    [text3918, 2707]                        DACA and later on for | DAPA |
##    [text3918, 2822]                    Obama came with the policy | DAPA |
##    [text3918, 3008]                      those policies, DACA and | DAPA |
##    [text3918, 3015]             are clearly unconstitutional. And | DAPA |
##    [text3918, 3040]                         Judge Andrew Hanen. The | DAPA |
##    [text3918, 3069] had the clearest constitutional understanding | DAPA |
##    [text3918, 3835]                       . That includes DACA and | DAPA |
##    [text3918, 3876]                     he would address DACA and | DAPA |
##    [text3918, 3989]                        any action on DACA and | DAPA |
##    [text3918, 4004]                    every DACA permit and every | DAPA |
##    [text3918, 4259]                      the earlier that DACA and | DAPA |
##   [text3918, 14014]                            end DACA, let's end | DAPA |
##    [text7757, 1177]        openly and blatantly unconstitutional. | DAPA |
##    [text7757, 1235]                    have the authority for that | DAPA |
##    [text17228, 876]                         Permanent Residents(`` | DAPA |
##    [text17228, 948]                      would continue, DACA and | DAPA |
##    [text17228, 983]                      joined with Texas in the | DAPA |
##   [text17228, 1011]                been operative since 2012 while | DAPA |
##   [text17228, 1075]        in the pre-implementation challenge to | DAPA |
##   [text17228, 1105]                immigrants who were affected by | DAPA |
##   [text17228, 1138]                   would have been eligible for | DAPA |
##   [text20586, 1223]                         case come forward, the | DAPA |
##   [text20586, 1349]                  case in parallel fashion that | DAPA |
##    [text26865, 957]                          did with his DACA and | DAPA |
##   [text26865, 1922]                          to shut down DACA and | DAPA |
##  [text29519, 22081]                          to shut down DACA and | DAPA |
##   [text47970, 1001]                and Lawful Permanent Residents, | DAPA |
##   [text69450, 1014]                            and then later, the | DAPA |
##    [text75340, 831]                 DACA program and the proposed | DAPA |
##    [text75357, 139]              Residents, wrote that permitting | DAPA |
##    [text75357, 184]                            criminals. In fact, | DAPA |
```

```
##    [text75357, 308]                have voiced their support for | DAPA |
##    [text75357, 335]              after voicing his opposition to | DAPA |
##   [text80196, 1263]              Parental Accountability, or the | DAPA |
##   [text80196, 1266]                          or the DAPA Program. | DAPA |
##     [text80295, 90]                             Act, the DACA and | DAPA |
##
##  that President Obama so unconstitutionally
##  . Well, it was
##  , the Deferred Action for
##  , are clearly unconstitutional.
##  , Texas brought that case
##  policy is now at least
##  is unconstitutional and the President
##  . It needs to also
##  and the Morton memos.
##  . I would rescind the
##  permit. We have got
##  are addressed by this President
##  , and let's end the
##  , the Deferred Action for
##  program, and he knew
##  '') initiative that
##  are`` two separate
##  case before the Supreme Court
##  never went into effect.
##  . Further, the Fifth
##  , Texas, 809 F
##  ( up to 4.3 million
##  case, the Deferred Action
##  was litigated successfully. They
##  action, a higher percentage
##  and deport hundreds of thousands
##  and deport hundreds of thousands
##  , programs. These actions
##  Program. I felt that
##  program, which he claimed
##  -- the acronym for
##  was a program that would
##  because the program actually advances
##  , Mr. Duncan submitted
##  Program. DAPA would have
##  would have provided protections for
##  programs, the Voting Rights
```

Looking at the number of rows of the resultant objects, it appears that DAPA usually occurs in the context of the DACA discussion.

```
nrow(daca)
```

```
## [1] 3604
```

```
nrow(dapa)
```

```
## [1] 37
```

Plotting the kwic object using the current corpus is problematic due to the large number of documents. Let

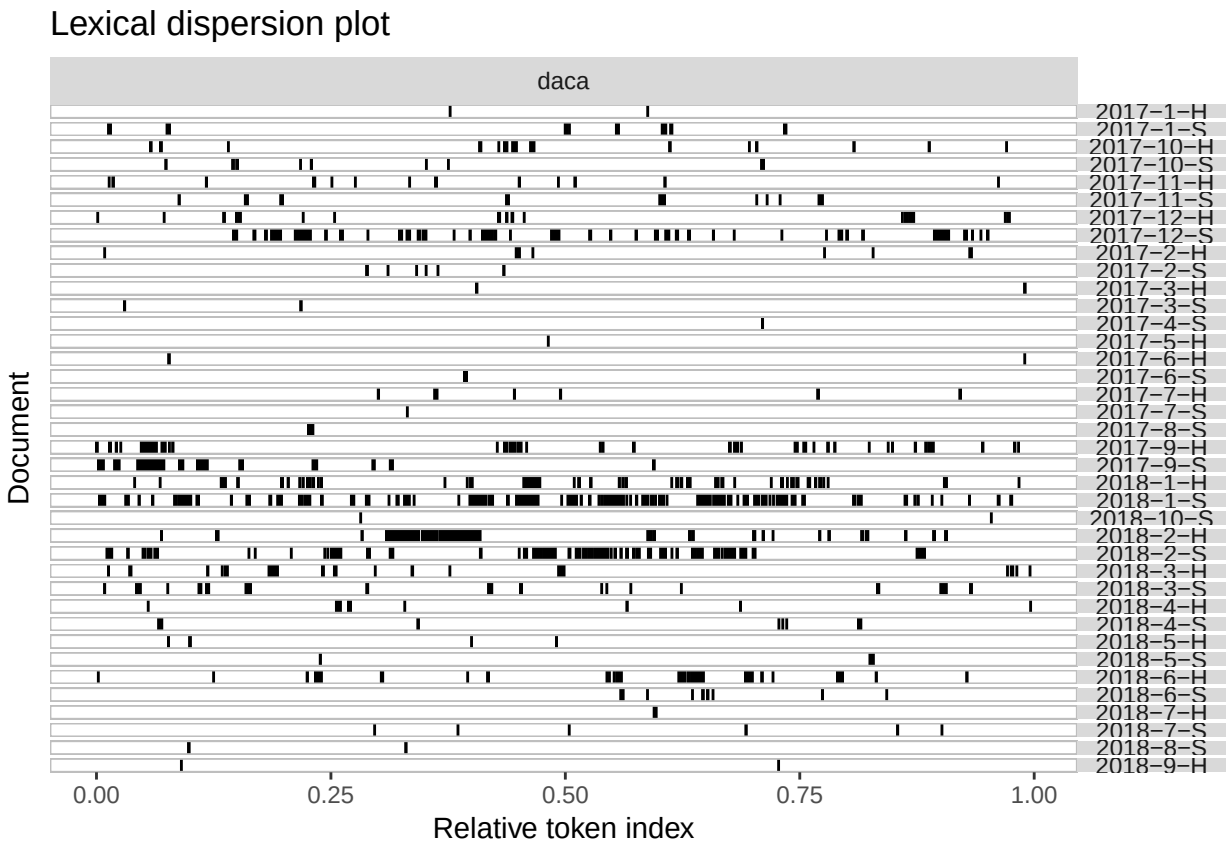
us make the original corpus more manageable by merging together all speeches within the same month in House and Senate. We pass `docid_field` parameter to get a more intuitive text labelling for the plot.

```
aggregate115 <- congress115 %>%
  dplyr::mutate(year = as.character(lubridate::year(date)),
                month = as.character(lubridate::month(date)),
                year_month_chamber = paste(year, month, chamber, sep = "-")) %>%
  dplyr::group_by(year_month_chamber) %>%
  dplyr::summarise(text = paste(text, collapse = " "))

aggregate115 <- quanteda::corpus(aggregate115,
                                 docid_field = "year_month_chamber")
```

Now we can apply the `textplot_xray()` function to get some insight into the distribution of the mentions of DACA over time and over chambers.

```
quanteda::textplot_xray(kwic(aggregate115, "daca", window = 5, valuetype = "fixed"))
```



From the plot we can see that the bulk of the discussion took place in winter 2017/18 with somewhat higher number of mentions in the Senate.

Dictionaries

Despite being perhaps the oldest analytical technique, dicitonaries are still frequently used by many researchers. To define a simple dictionary we will a `dictionary()` function.

```
dict <- quanteda::dictionary(
  list(trade = c("trade", "business", "corp*"),
```

```
       tax = c("tariff", "fiscal", "tax*"),
       regulation = c("law", "agreement", "deal", "regul*")))
```

To apply the dictionary we create another dfm and pass the dictionary as an argument. Another way to apply it is to use `dfm_lookup()` function to the already existing dfm.

```
dict115 <- quanteda::dfm(corpus115, dictionary = dict)
head(dict115)
```

```
## Document-feature matrix of: 6 documents, 3 features (77.8% sparse).
## 6 x 3 sparse Matrix of class "dfm"
##        features
## docs    trade tax regulation
##   text1     0   0          0
##   text2     0   0          2
##   text3     0   0          0
##   text4     1   0          1
##   text5     0   0          1
##   text6     0   0          0
```

```
# Or, equivalently
lookup115 <- quanteda::dfm_lookup(dfm115, dict, valuetype = "glob")
```

While helpful for certain kinds of analysis, a more useful approach might be to apply dictionary to a grouped dfm.

```
dict115 <- quanteda::dfm(grouped115, dictionary = dict)
head(dict115)
```

```
## Document-feature matrix of: 2 documents, 3 features (0.0% sparse).
## 2 x 3 sparse Matrix of class "dfm"
##     features
## docs trade   tax regulation
##    H 18819 37212      43785
##    S 15964 27905      31188
```

```
# It's also useful to see the proportions
quanteda::dfm_weight(dict115, scheme = "prop")
```

```
## Document-feature matrix of: 2 documents, 3 features (0.0% sparse).
## 2 x 3 sparse Matrix of class "dfm"
##     features
## docs     trade       tax regulation
##    H 0.1885369 0.3728060  0.4386571
##    S 0.2126917 0.3717841  0.4155242
```

Let us now use some automatic heuristics to create a dictionary from a given seed. We will use `trade` as our seed word and then proceed to explore the words most often used together with it, but exclude those used in the corpus more generally.

```
trade115 <- congress115 %>%
  dplyr::filter(grepl("trade", text))

trade115 <- quanteda::corpus(trade115)

# Extract most used terms in all the documents
top115 <- topfeatures(dfm115, 100)
```

```r
# Extract most used terms in trade-related documents
toptrade115 <- topfeatures(dfm(trade115, remove = stopwords("en"), remove_numbers = TRUE, remove_punct =

# As this is a named vector, we will apply names() function
autodict <- names(toptrade115)[!(names(toptrade115) %in% names(top115))]
autodict
```

```
##  [1] "funds"          "paragraph"      "u.s.c"          "amended"
##  [5] "inserting"      "striking"       "fiscal"         "d"
##  [9] "ii"             "date"           "code"           "assistance"
## [13] "expenses"       "amount"         "information"    "programs"
## [17] "activities"     "authorized"     "office"         "military"
## [21] "agency"         "term"           "described"      "necessary"
## [25] "pursuant"       "e"              "appropriations" "end"
## [29] "subparagraph"   "used"           "authority"      "administration"
## [33] "development"    "committees"     "foreign"        "plan"
## [37] "later"          "purposes"       "days"           "provide"
## [41] "respect"        "appropriate"    "appropriated"   "remain"
## [45] "requirements"   "system"         "enactment"
```

Challenge 2

Easy mode Explore the context of the words *gun* and *firearm* with the `kwic()` function. Are there any issues with the default setting? Treat them as `glob` to capture plural forms.

Medium Now plot the kwic objects on the aggregated corpus ot explore when and where these issues were discussed in the US Congress in the last two years.

Advanced To make the previous plot a bit nicer and easier to read, order the aggregated texts by year, month, chamber. To do this, you would need to modify the ordering of levels in the `aggregate115` dataset.

Subject Expert Create a dictionary related to gun violence and firearms control and apply to Congress debates.