Machine learning with text

Greg Ridgeway

2025-03-23

Table of contents

1 Introduction

In this section we will use `text2vec` to explore the language used in a collection of police reports describing officer-involved shootings (OIS). These reports contain unstructured narrative text. Our goal is to transform that text into a format we can analyze using tools from natural language processing (NLP). We will walk through a typical text analysis process: tokenizing the reports, building a vocabulary, constructing a document-term matrix, and applying TF-IDF to highlight the most distinctive terms. Along the way, we will also examine co-occurrence patterns.

To start, we are going to need a couple of R packages to facilitate our work. `text2vec` will do most of the work converting the documents into a form of data that we can analyze.

```
library(text2vec)
```

```
Warning: package 'text2vec' was built under R version 4.4.3
```

```
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(Matrix)
```

```
Warning: package 'Matrix' was built under R version 4.4.2
```

As for the source of our documents, the Philadelphia Police Department posts (reports)[https://www.phillypolice.com/accountability/ois/] on each officer-involved shooting. I have pulled the data off their website and packaged it into an .RData file. Loading it will create the data frame `ois`. Details on how to pull the data off of the PPD website are part of my (R4crim collection)[https://github.com/gregridgeway/R4crim?tab=readme-ov-file] of scripts.

```
load("data/PPD OIS.RData")
ois |> select(-text) |> head()
```

```
     id                                location
1 24-37      3450 Vista Street, Philadelphia, PA
2 24-36          3250 A Street, Philadelphia, PA
3 24-35 5450 Chancellor Street, Philadelphia, PA
4 24-32         2950 E. Street, Philadelphia, PA
5 24-31      3350 Willits Road, Philadelphia, PA
6 24-30    6150 Lebanon Avenue, Philadelphia, PA
                                       url       date       lon      lat
1   https://www.phillypolice.com/ois/24-37/ 2024-12-10 -75.03885 40.04022
2   https://www.phillypolice.com/ois/24-36/ 2024-11-12 -75.12714 39.99956
3   https://www.phillypolice.com/ois/24-35/ 2024-11-10 -75.23050 39.95692
4   https://www.phillypolice.com/ois/24-32/ 2024-10-11 -75.12024 39.99345
5   https://www.phillypolice.com/ois/24-31/ 2024-10-03 -75.00908 40.05383
6 https://www.phillypolice.com/ois/ps24-30/ 2024-10-02 -75.24450 39.98175
```

```
                                                addrmatch score      addrtype
1      3450 Vista St, Philadelphia, Pennsylvania, 19136   100 StreetAddress
2          3250 A St, Philadelphia, Pennsylvania, 19134   100 StreetAddress
3 5450 Chancellor St, Philadelphia, Pennsylvania, 19139   100 StreetAddress
4          2950 E St, Philadelphia, Pennsylvania, 19134   100 StreetAddress
5    3350 Willits Rd, Philadelphia, Pennsylvania, 19136   100 StreetAddress
6   6150 Lebanon Ave, Philadelphia, Pennsylvania, 19151   100  PointAddress
```

The data include an incident ID, the date of the shooting, the address and coordinates where the shooting occurred, and a URL to the incident report. There is also a column called `text` containing the full text of the officer-involved shooting report. Some can be long, but here's the first one as an example.

```
ois |> filter(id=="16-30") |> select(text) |> unlist() |> cat()
```

```
PS#16-30
9/16/16
On Friday, September 16, 2016, at approximately 11:18 P.M., a uniformed sergeant in a marked
Responding uniformed officers, in marked police vehicles, along with an officer from the Univ
The offender's firearm, a 9MM, semi-automatic pistol, with an obliterated serial number, load
The sergeant, the University of Pennsylvania Officer, along with the four civilians who were
The female from the parked vehicle was later pronounced deceased at Penn-Presbyterian Hospita
*** Information posted in the original summary reflects a preliminary understanding of what o
```

With this set of 133 reports, we will use a variety of data cleaning methods and machine learning methods to try to make sense of these documents.

2 Turning text into data with `text2vec`

```
# Create an iterator over tokens
#   tokens does not actually store data
#   just an efficient means for looping over documents
tokens <- itoken(ois$text,
                 progressbar = TRUE,
                 ids = ois$id)
# this gets the next batch of documents... for me around 14 documents
a <- tokens$nextElem()
a$ids
```

```
 [1] "24-37" "24-36" "24-35" "24-32" "24-31" "24-30" "24-29" "24-28" "24-27"
[10] "24-23" "24-22" "24-21" "24-20" "24-18"
```

```
a$tokens |> sapply(head)
```

```
     [,1]         [,2]         [,3]         [,4]         [,5]
[1,] "3400"       "3200"       "5400"       "29oo"       "3300"
[2,] "block"      "block"      "block"      "block"      "Willits"
[3,] "of"         "of"         "of"         "of"         "Road\nOn"
[4,] "Vista"      "A"          "Chancellor" "E."          "Thursday,"
[5,] "Street\nOn" "Street\nOn" "Street\nOn" "Street\nOn" "October"
[6,] "Tuesday,"   "Tuesday,"   "Sunday,"    "Friday,"    "3,"
     [,6]         [,7]           [,8]           [,9]          [,10]
[1,] "6100"       "2600"         "3900"         "2200"        "3000"
[2,] "block"      "block"        "block"        "block"       "block"
[3,] "of"         "of"           "of"           "of"          "of"
[4,] "Lebanon"    "Glenwood"     "Whittaker"    "S."          "Ruth"
[5,] "Avenue\nOn" "Avenue\nThe"  "Avenue\nThe"  "65th"        "Street\nThe"
[6,] "Wednesday," "Philadelphia" "Philadelphia" "Street\nThe" "Philadelphia"
     [,11]       [,12]          [,13]       [,14]
[1,] "6100"      "3500"         "2700"      "1500"
[2,] "block"     "block"        "block"     "block"
[3,] "of"        "of"           "of"        "of"
[4,] "West"      "F"            "North"     "North"
[5,] "Columbia"  "Street\nA"    "6th"       "57th"
[6,] "Avenue\nA" "Philadelphia" "Street\nA" "Street\nA"
```

```
# reset to beginning
tokens <- itoken(ois$text,
                 progressbar = TRUE,
                 ids = ois$id)


# Build vocabulary
#   these are the collection of words that I care about
#   skip stopwords (the, of, in, ...)
#   include two word phrases (2-gram or bigram),
#      "police officer", "full uniform", "black male", "drop weapon"
#   skip words that only show up in fewer than 10 documents
#   skip words that are in the majority of documents (police?, discharged?)
vocab <- tokens |>
  create_vocabulary(stopwords = stopwords::stopwords("en"),
```

```
                    ngram = 1:2) |>
  prune_vocabulary(term_count_min = 10,
                   doc_proportion_max = 0.5) |>
  filter(nchar(term) >=3) |>
  filter(!grepl("[0-9]", term))

# word_tokenizer(), default, keeps a lot of punctuation
vocab
```

```
Number of docs: 133
175 stopwords: i, me, my, myself, we, our ...
ngram_min = 1; ngram_max = 2
Vocabulary:
                  term term_count doc_count
                <char>      <int>     <int>
  1:               AM,         10        10
  2:        Avenue\nOn         10        10
  3:   District_Police         10        10
  4:         Hospital,         10         9
  5: Penn-Presbyterian         10         8
 ---
552:           firearm        112        63
553:               one        116        65
554:            posted        120        60
555:           vehicle        180        65
556:          offender        182        49
```

```
# redo with our own custom tokenizer
oisTokenizer <- function(text)
{
  text |>
    tolower() |>
    # remove abreviation punctuation (like 3 p.m.)
    gsub("([A-z])[,.]", "\\1", x=_) |>
    # remove some weird symbols
    gsub("[""()#]", "", x=_) |>
    strsplit("\\s+")
}

# reset to beginning
tokens <- itoken(ois$text,
                 tokenizer = oisTokenizer,
```

```
                progressbar = TRUE,
                ids = ois$id)

vocab <- tokens |>
  create_vocabulary(stopwords = stopwords::stopwords("en"),
                    ngram = 1:2) |>
  prune_vocabulary(term_count_min = 10,
                   doc_proportion_max = 0.5) |>
  filter(nchar(term) >=3) |>
  filter(!grepl("[0-9]", term))

vocab
```

```
Number of docs: 133
175 stopwords: i, me, my, myself, we, our ...
ngram_min = 1; ngram_max = 2
Vocabulary:
                  term term_count doc_count
                <char>      <int>     <int>
  1:          announced         10         8
  2:        approaching         10        10
  3:            assault         10         6
  4:            cameras         10        10
  5:            chelten         10         3
 ---
549:                two        100        55
550:               door        103        51
551: information_posted        120        60
552:             posted        120        60
553:           offender        270        55
```

```
vocab$term
```

```
  [1] "announced"                "approaching"
  [3] "assault"                  "cameras"
  [5] "chelten"                  "discharge_weapon"
  [7] "district_placed"          "district_police"
  [9] "drew_firearm"             "due"
 [11] "firearm_discharge"        "fled_scene"
 [13] "gave"                     "hospital_critical"
 [15] "hospital_police"          "individuals"
```

```
 [17] "june"                      "lane"
 [19] "location_officers"         "lost"
 [21] "lot"                       "notified"
 [23] "offender's_firearm"        "operating_unmarked"
 [25] "penn-presbyterian"         "penn-presbyterian_hospital"
 [27] "requested"                 "responding_officer"
 [29] "rounds_recovered"          "scene_injuries"
 [31] "seated"                    "shoulder"
 [33] "street_male"               "township_police"
 [35] "vehicle_officers"          "went"
 [37] "white"                     "yard"
 [39] "additional_officers"       "assignment"
 [41] "august"                    "basement"
 [43] "bike"                      "block_e"
 [45] "call_person"               "discharged_one"
 [47] "exit"                      "glass"
 [49] "greene"                    "homicide"
 [51] "inside_property"           "involved_shooting"
 [53] "listed_critical"           "missing_offender"
 [55] "monday"                    "moved"
 [57] "offenders"                 "officer_exited"
 [59] "officers'"                 "order"
 [61] "outside"                   "passenger_side"
 [63] "pm,_officer"               "radio_assignment"
 [65] "removed"                   "shooting_investigation"
 [67] "striking_officer"          "sustained_gunshot"
 [69] "treated_released"          "vehicle_male"
 [71] "vest"                      "victims"
 [73] "westbound"                 "arrest"
 [75] "bedroom"                   "blue"
 [77] "body_worn"                 "building"
 [79] "comply"                    "currently"
 [81] "deployed"                  "door_officer"
 [83] "driver's_door"             "driver's_seat"
 [85] "drove"                     "encountered"
 [87] "february"                  "firearm_officer"
 [89] "firearm_striking"          "inside_residence"
 [91] "kitchen"                   "large"
 [93] "listed_stable"             "lower"
 [95] "male_drop"                 "male_male"
 [97] "male_suspect"              "next"
 [99] "offender_fled"             "officer_involved"
[101] "officers_placed"           "pistol_loaded"
```

```
[103] "plainclothes_officers"      "pursued_male"
[105] "received"                   "roosevelt"
[107] "service"                    "shooting_victim"
[109] "shot_one"                   "side_door"
[111] "street_observed"            "third"
[113] "times_striking"             "torso"
[115] "township"                   "worn"
[117] "april"                      "arrived_location"
[119] "broad"                      "captured"
[121] "followed"                   "gun_officer"
[123] "made"                       "male_foot"
[125] "narcotics"                  "november"
[127] "number_two"                 "officer_placed"
[129] "officers_approached"        "ongoing"
[131] "opened"                     "right_hand"
[133] "room"                       "taser"
[135] "toyota"                     "tuesday"
[137] "uniform_operating"          "unmarked_vehicle"
[139] "arrested"                   "believed"
[141] "camera"                     "discharged_weapons"
[143] "driveway"                   "drop_weapon"
[145] "hands"                      "home"
[147] "male_transported"           "october"
[149] "offender_transported"       "officers_officer"
[151] "owner"                      "park"
[153] "pointed_firearm"            "reference"
[155] "reported_connection"        "reported_injuries"
[157] "september"                  "towards_officer"
[159] "weapon_striking"            "wounds"
[161] "able"                       "am,"
[163] "arm"                        "around"
[165] "called"                     "civilian"
[167] "district_officers"          "drop_knife"
[169] "found"                      "friday"
[171] "gsw"                        "hospital_pronounced"
[173] "injuries_police"            "live_rounds"
[175] "magazine"                   "man"
[177] "medical"                    "officers_exited"
[179] "one_officers"               "pointing"
[181] "retreated"                  "rifle"
[183] "saturn"                     "street_upon"
[185] "struggle"                   "swat_unit"
[187] "towards_officers"           "without"
```

```
[189] "woman"                    "attack"
[191] "block_south"              "connection"
[193] "connection_incident"      "cover"
[195] "detective"                "get"
[197] "gunshot_wound"            "observed_offender"
[199] "operator"                 "pm,_uniformed"
[201] "pulled"                   "returned_fire"
[203] "robbery"                  "sidewalk"
[205] "stated"                   "subsequently"
[207] "wanted"                   "warrant"
[209] "wearing"                  "admitted"
[211] "apartment"                "control"
[213] "corner"                   "critical_condition"
[215] "district_officer"         "driver's_side"
[217] "dropped"                  "four"
[219] "gunfire"                  "highway"
[221] "however"                  "information_district"
[223] "off-duty"                 "onto"
[225] "presbyterian_hospital"    "rear_passenger"
[227] "responding_officers"      "semi-automatic_pistol"
[229] "several_times"            "taken"
[231] "traveling"                "vehicle_officer"
[233] "walking"                  "window"
[235] "another"                  "charged"
[237] "chest"                    "coming"
[239] "custody"                  "discharged_firearms"
[241] "hospital_treatment"       "intersection"
[243] "officers_marked"          "ordered_male"
[245] "point_officer"            "presbyterian"
[247] "pursuit"                  "standing"
[249] "temple_hospital"          "along"
[251] "apprehended"              "arrival_officers"
[253] "bull"                     "leg"
[255] "male_fled"                "officer_observed"
[257] "officers_responded"       "pit_bull"
[259] "saw"                      "thursday"
[261] "unknown"                  "wednesday"
[263] "years_old"                "commands"
[265] "driving"                  "open"
[267] "patrol_car"               "ppd"
[269] "response_officer"         "road"
[271] "round"                    "search"
[273] "seat"                     "also"
```

```
[275] "attorney's"              "attorney's_office"
[277] "civilians"               "district_attorney's"
[279] "east"                    "injuries_reported"
[281] "saturday"                "sergeant"
[283] "defendant"               "fell_ground"
[285] "head"                    "multiple_times"
[287] "old"                     "took"
[289] "years"                   "block_n"
[291] "located"                 "officers_discharged"
[293] "officer's"               "treated"
[295] "uniformed_officers"      "waistband"
[297] "west"                    "discharge"
[299] "fled_foot"               "floor"
[301] "live"                    "marked_police"
[303] "result_incident"         "unmarked_police"
[305] "additional"              "block_north"
[307] "offender's"              "running"
[309] "weapons"                 "wound"
[311] "activated"               "affairs_officer-involved"
[313] "critical"                "holding"
[315] "identified"              "offender_offender"
[317] "person"                  "pit"
[319] "police_radio"            "pronounced_deceased"
[321] "pursued"                 "released"
[323] "semi-automatic"          "body"
[325] "front_door"              "gunshots"
[327] "listed"                  "missing"
[329] "officer_officer"         "ois"
[331] "pistol"                  "returned"
[333] "still"                   "street_officers"
[335] "striking_offender"       "caliber"
[337] "involved"                "loaded"
[339] "officer_discharged"      "officers_observed"
[341] "patrol_vehicle"          "streets"
[343] "temple_university"       "university_hospital"
[345] "away"                    "behind"
[347] "entered"                 "full_uniform"
[349] "injured"                 "injuries_result"
[351] "operating_marked"        "police_officers"
[353] "positioned"              "responded_radio"
[355] "striking_male"           "vehicles"
[357] "arrived"                 "back"
[359] "fire"                    "full"
```

```
[361] "number_one"                 "partner"
[363] "radio_call"                 "ran"
[365] "stop"                       "street_officer"
[367] "turned"                     "ground"
[369] "near"                       "one_time"
[371] "parked"                     "pointed"
[373] "shooting_investigations"    "toward"
[375] "transported_temple"         "treatment"
[377] "later"                      "uniform"
[379] "veteran_philadelphia"       "arrival"
[381] "black_male"                 "discharging_officer"
[383] "dogs"                       "drew"
[385] "fell"                       "gunshot"
[387] "outcome_internal"           "plainclothes"
[389] "stolen"                     "upon_arrival"
[391] "department_assigned"        "discharged_firearm"
[393] "incident_***"               "recovered_scene"
[395] "second"                     "attempted"
[397] "call"                       "deceased"
[399] "direction"                  "exited_vehicle"
[401] "firearms"                   "male's"
[403] "ordered"                    "police_district"
[405] "stopped"                    "marked_patrol"
[407] "sustained"                  "three"
[409] "university"                 "complainant"
[411] "continued"                  "driver"
[413] "response"                   "result"
[415] "stable"                     "stable_condition"
[417] "males"                      "officer_number"
[419] "shot"                       "heard"
[421] "struck"                     "observed_male"
[423] "police_department"          "responding"
[425] "right"                      "began"
[427] "da's_office"                "information_da's"
[429] "location"                   "unmarked"
[431] "da's"                       "investigations"
[433] "pronounced"                 "rounds"
[435] "armed"                      "left"
[437] "officer-involved_shooting"  "reported"
[439] "swat"                       "discharged_weapon"
[441] "drop"                       "hand"
[443] "multiple"                   "victim"
[445] "driver's"                   "duty_pending"
```

```
[447] "side"                      "administrative_duty"
[449] "officer-involved"          "outcome"
[451] "pending_outcome"           "placed_administrative"
[453] "pm,"                       "car"
[455] "north"                     "several"
[457] "south"                     "temple"
[459] "administrative"            "affairs"
[461] "fired"                     "internal"
[463] "internal_affairs"          "operating"
[465] "residence"                 "times"
[467] "police_officer"            "uniformed"
[469] "veteran"                   "duty"
[471] "handgun"                   "female"
[473] "knife"                     "number"
[475] "pending"                   "condition"
[477] "passenger"                 "upon"
[479] "department"                "approached"
[481] "dog"                       "point"
[483] "property"                  "***"
[485] "***_information"           "prior_charging"
[487] "charging_decision"         "decision"
[489] "incident_information"      "incident_may"
[491] "information_ppd's"         "investigation_leads"
[493] "investigation_prior"       "leads"
[495] "leads_new"                 "may_updated"
[497] "new_information"           "occurred_time"
[499] "office_provided"           "original_summary"
[501] "posted_original"           "posted_shortly"
[503] "ppd's_investigation"       "preliminary_understanding"
[505] "provided_information"      "reflects"
[507] "reflects_preliminary"      "shortly_incident"
[509] "summary"                   "summary_reflects"
[511] "understanding"             "understanding_occurred"
[513] "updated"                   "updated_investigation"
[515] "foot"                      "new"
[517] "original"                  "ppd's"
[519] "provided"                  "responded"
[521] "towards"                   "charging"
[523] "philadelphia_police"       "discharging"
[525] "preliminary"               "prior"
[527] "time_incident"             "unit"
[529] "placed"                    "radio"
[531] "shortly"                   "assigned"
```

```
[533] "fled"                    "marked"
[535] "inside"                  "police_vehicle"
[537] "rear"                    "gun"
[539] "black"                   "exited"
[541] "area"                    "scene"
[543] "patrol"                  "front"
[545] "philadelphia"            "shooting"
[547] "avenue"                  "suspect"
[549] "two"                     "door"
[551] "information_posted"      "posted"
[553] "offender"
```

```
# Create a vectorizer
#   helper function to convert streams of text into dta matrix
vectorizer <- vocab_vectorizer(vocab)

# Create the document-term matrix (DTM)
#   row represents a document
#   column represents a unique term (word or phrase)
#   cell contains the count (or weight) of that term in the document
oisDTM <- create_dtm(tokens, vectorizer)

# number of documents and words
dim(oisDTM)
```

```
[1] 133 553
```

```
# rows represent individual OIS shooting reports
rownames(oisDTM)[1:5]
```

```
[1] "24-37" "24-36" "24-35" "24-32" "24-31"
```

```
# columns are the words/phrases
colnames(oisDTM)[1:10]     # feature names
```

```
 [1] "announced"        "approaching"      "assault"          "cameras"
 [5] "chelten"          "discharge_weapon" "district_placed"  "district_police"
 [9] "drew_firearm"     "due"
```

```
# how many vocab words in document?
rowSums(oisDTM)
```

```
24-37 24-36 24-35 24-32 24-31 24-30 24-29 24-28 24-27 24-23 24-22 24-21 24-20
   56    73   113    54    66    65    69    66    45    88    87    85    99
24-18 24-17 24-15 24-14 24-13 24-12 24-10 24-09 24-08 24-07 24-06 24-05 24-04
  110    75   111    84    94   137    87   169   120    98   103   179   225
24-03 24-02 24-01 23-33 23-31 23-29 23-27 23-26 23-25 23-24 23-23 23-21 23-14
   99    85   101    78   147    80   141   102    98    94   111   102    97
23-13 23-10 23-04 22-27 22-26 22-24 22-22 22-15 22-14 22-10 22-09 22-08 22-07
   79    96   159   117   173   103    92   192   126   240    92   112   111
22-06 22-05 22-04 22-03 22-01 21-15 21-14 21-12 21-10 21-09 21-06 21-04 20-34
  180   135    85    95    82    82    82   201   150    67   102   176    63
20-33 20-32 20-31 20-30 20-29 20-26 20-24 20-23 20-20 20-15 20-12 20-08 20-07
  165   125   100   123   104   298   131   101   205   120   128   130   115
19-23 19-21 19-20 19-14 19-13 19-11 19-09 19-06 19-04 18-28 18-27 18-26 18-25
  141   117   157    99   154   127   167   140   135   136   135   135   128
18-22 18-19 18-17 18-16 18-12 18-08 18-02 18-01 17-37 17-36 17-30 17-28 17-25
  105   118   133   144    97    99   112   162   152   109    87    86   117
17-23 17-22 17-20 17-19 17-17 17-13 17-03 16-43 16-40 16-38 16-37 16-35 16-34
  136   113   104   119   129   126   148   151   158   126   164   150   101
16-33 16-32 16-30 16-29 16-28 16-19 16-18 16-16 16-13 16-12 16-11 16-10 16-07
  170   153   153   132   111    97   125   119   114   138   147   130   127
16-03 16-02 16-01
  192   142   124
```

```
# how many documents have these words?
colSums(oisDTM)[1:20]
```

```
       announced        approaching            assault            cameras
              10                 10                 10                 10
         chelten   discharge_weapon   district_placed   district_police
              10                 10                 10                 10
    drew_firearm                due firearm_discharge        fled_scene
              10                 10                 10                 10
           gave  hospital_critical   hospital_police        individuals
              10                 10                 10                 10
           june               lane  location_officers               lost
              10                 10                 10                 10
```

```r
# Most common words?
colSums(oisDTM) |>
  sort(decreasing = TRUE) |>
  head(10)
```

```
     offender information_posted              posted                door
          270                120                 120                 103
          two            suspect              avenue            shooting
          100                 99                  98                  93
 philadelphia              front
           89                 86
```

3 Term Frequency–Inverse Document Frequency

Term Frequency–Inverse Document Frequency (TF-IDF) gives weights to words in a document in a way that balances:

1. *Term Frequency (TF)* "This word must be important in this document"

- The more a word appears in a document, the more likely it is to be relevant to the document's content
- If the word "shooting" appears 12 times in a police report, it's probably central to that document

2. *Inverse Document Frequency (IDF)*: "But if it appears in *every* document, it's not very informative"

- Common words like "officer", "incident", or "said" might appear everywhere
- IDF *downweights* those high-frequency but low-discrimination terms
- It prefers terms that help *distinguish* one document from others

The formula for TF-IDF for document i and term j:

$$tfidf_{ij} = TF_{ij} \log \frac{N}{DF_j}$$

where - TF. This is the number of times term j appears in document i. It measures the importance of the term within a document - N = total number of documents - DF_j = number of documents containing term j

$IDF_{ij} = \log \frac{N}{DF_j}$ captures the rarity across documents. Note that if a word appears in all documents then $tfidf_{ij} = 0$. The combination of TF and IDF gives a measure of relevance and distinctiveness. A high $tfidf_{ij}$ means a term appears often in document i, but rarely in

other documents. It gives you terms that define a document. These are the terms that are useful for classification, clustering, or topic modeling.

3.1 Example

Assume there are $N = 100$ documents.

Term	TF in Doc A	DF across corpus	IDF	TF-IDF
"weapon"	5	10	2.3	11.5
"officer"	6	95	0.1	0.3
"said"	20	100	0	0

- **"weapon"** gets a high score, specific and relevant
- **"officer"** is common, downweighted
- **"said"** is everywhere, zeroed out

```
# TF-IDF: term frequency-inverse document frequency weights
#    downweights common words that appear in many documents
#    upweights rare words that are more informative or distinctive
# TF: How often a word appears in a document
# IDF: How rare that word is across all documents
#    TF-IDF = TF × log(N / DF)
#       N = total number of documents
#       DF = number of documents containing the term
tfidf_transformer <- TfIdf$new()
oisTFIDF <- tfidf_transformer$fit_transform(oisDTM)

oisTFIDF[1:10,1:10] |> as.matrix() |> round(2) |> t()
```

```
                 24-37 24-36 24-35 24-32 24-31 24-30 24-29 24-28 24-27 24-23
announced            0     0  0.00     0     0     0     0     0  0.00  0.00
approaching          0     0  0.00     0     0     0     0     0  0.00  0.00
assault              0     0  0.00     0     0     0     0     0  0.00  0.04
cameras              0     0  0.00     0     0     0     0     0  0.00  0.00
chelten              0     0  0.00     0     0     0     0     0  0.00  0.00
discharge_weapon     0     0  0.02     0     0     0     0     0  0.00  0.00
district_placed      0     0  0.00     0     0     0     0     0  0.06  0.03
district_police      0     0  0.00     0     0     0     0     0  0.00  0.00
drew_firearm         0     0  0.00     0     0     0     0     0  0.00  0.00
due                  0     0  0.00     0     0     0     0     0  0.00  0.00
```

```
# View top features by TF-IDF (note: no direct topfeatures method)
colSums(oisTFIDF) %>%
  sort(decreasing = TRUE) %>%
  head(10)
```

```
          offender            suspect                dog       philadelphia
         2.7175154          1.7096382          1.4692017          1.1719846
information_posted             posted              knife             avenue
         1.1005283          1.1005283          1.0848150          1.0585912
          shooting               door
         1.0452646          0.9737314
```

When scanning through each document, setting `skip_grams_window = 5` will treat any two terms that appear within a window of 5 tokens as co-occurring. For example, if the document has the phrase "the officer shot the suspect with a weapon" and we set `skip_grams_window = 5`, then for the word "shot" it will consider "the", "officer", "the", "suspect", "with" as co-occurring terms.

```
# Create a co-occurrence matrix (Feature Co-occurrence Matrix)
oisTCM <- itoken(ois$text,
                 tokenizer = oisTokenizer,
                 progressbar = FALSE,
                 ids = ois$id) |>
  create_tcm(vocab_vectorizer(vocab),
             skip_grams_window = 5)

# Convert to triplet format and extract top co-occurring pairs
oisPairs <- Matrix::summary(oisTCM) |>
  filter(i != j) |>
  rename(feature1 = i, feature2 = j, weight = x) |>
  left_join(data.frame(feature1 = 1:nrow(oisTCM),
                       term1 = colnames(oisTCM)))  |>
  left_join(data.frame(feature2 = 1:nrow(oisTCM),
                       term2 = colnames(oisTCM))) |>
  select(-feature1, -feature2) |>
  filter(term1 != term2) |>
  filter(!str_detect(term1, fixed(term2)) &
           !str_detect(term2, fixed(term1)))
```

```
Joining with `by = join_by(feature1)`
Joining with `by = join_by(feature2)`
```

```
oisPairs |>
  arrange(desc(weight)) |>
  slice_head(n = 10)
```

```
   weight                term1                 term2
1      30    information_ppd's   ppd's_investigation
2      30            leads_new       new_information
3      30  investigation_leads             leads_new
4      30  investigation_leads updated_investigation
5      30          may_updated updated_investigation
6      30         incident_may           may_updated
7      30         incident_may      shortly_incident
8      30       posted_shortly      shortly_incident
9      30              shortly                posted
10     30 incident_information    information_posted
```

```
oisPairs |>
  arrange(desc(weight)) |>
  filter(weight >= 30)
```

```
   weight                     term1                  term2
1      30         information_ppd's    ppd's_investigation
2      30                 leads_new        new_information
3      30       investigation_leads              leads_new
4      30       investigation_leads  updated_investigation
5      30               may_updated  updated_investigation
6      30              incident_may            may_updated
7      30              incident_may       shortly_incident
8      30            posted_shortly       shortly_incident
9      30                   shortly                 posted
10     30      incident_information     information_posted
11     30      incident_information          time_incident
12     30             occurred_time          time_incident
13     30 preliminary_understanding understanding_occurred
14     30 preliminary_understanding   reflects_preliminary
15     30                  reflects            preliminary
16     30      reflects_preliminary       summary_reflects
17     30          original_summary       summary_reflects
18     30                   summary                original
```

19	30	original_summary	posted_original
20	30	original	posted
21	30	posted_original	information_posted
22	30	office_provided	provided_information
23	30	understanding	preliminary
24	30	posted_shortly	information_posted
25	30	leads	new
26	30	reflects	summary
27	30	decision	charging
28	30	information_ppd's	provided_information
29	30	occurred_time	understanding_occurred
30	30	investigation_prior	ppd's_investigation