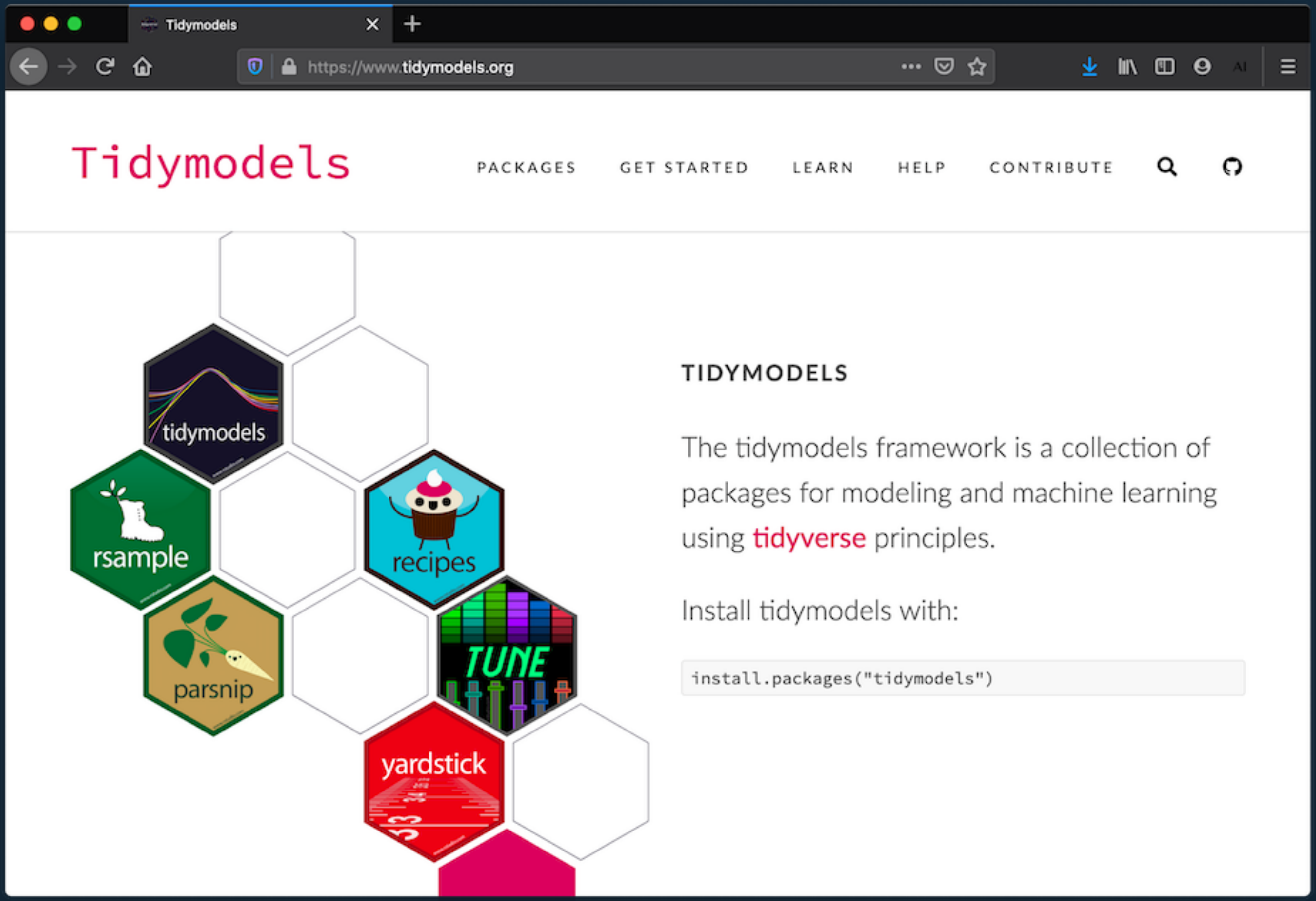
Creating features for machine learning from text

Julia Silge

"Aardvarks are small pig-like mammals that are found inhabiting a wide range of different habitats throughout Africa, south of the Sahara. They are mostly solitary and spend their days sleeping in underground burrows to protect them from the heat of the African sun, emerging in the cooler evening to search for food. Their name originates from the Afrikaans language in South Africa and means Earth Pig, due to their long snout and pig-like bo..."

##	a	able	african	after	all	along	also	although	an	and	animal	animals	are	areas	around	as	at	be	being	body	both	but	by	can	diet	different	due	eat	female	food
## 1	13	7	1	1	3	2	9	1	2	40	3	5	40	2	3	10	5	11	4	1	5	4	10	9	2	4	4	2	1	7
## 2	10	0	1	1	1	0	1	1	0	18	1	2	4	0	0	8	0	8	1	1	0	0	3	1	0	1	0	0	0	0
## 3	18	4	0	2	3	6	5	2	4	36	2	0	16	0	2	11	4	3	3	1	3	3	4	2	2	0	2	1	1	7
## 4	13	0	0	1	0	2	3	1	1	23	0	0	8	0	1	8	2	4	3	1	0	4	3	2	0	1	1	0	0	1
## 5	12	2	0	0	1	3	1	3	1	24	0	1	7	0	2	8	2	5	2	1	1	2	1	0	0	0	1	0	0	0
## 6	14	3	67	5	1	2	8	5	5	57	1	2	23	2	3	5	4	11	3	1	2	8	3	9	3	1	3	0	2	5
## 7	17	1	55	2	1	4	4	2	6	38	7	5	17	3	3	6	2	8	3	1	3	3	8	2	3	1	0	0	2	2
## 8	19	2	51	2	2	3	8	4	2	42	2	3	12	1	3	11	4	4	4	0	1	4	6	6	2	0	3	1	3	4
## 9	12	3	59	2	0	2	4	3	6	44	1	2	18	1	1	5	1	5	1	0	2	4	2	2	1	1	4	1	3	2
## 10	22	2	44	1	3	2	4	3	4	48	3	4	21	3	1	11	3	8	3	1	2	2	7	1	3	0	2	0	1	4
## 11	27	0	47	1	1	3	7	2	1	45	2	0	16	1	5	11	5	12	1	2	4	5	12	4	3	0	4	1	4	9
## 12	14	1	45	0	1	1	2	2	0	30	2	1	14	3	0	4	1	4	0	1	2	4	3	0	1	0	1	1	2	2
## 13	17	1	51	2	1	1	7	3	2	48	2	2	19	2	2	12	2	2	1	0	1	1	7	2	2	1	4	0	2	4
## 14	16	0	0	0	0	0	3	2	2	21	0	2	9	0	2	4	0	6	2	0	2	1	5	3	0	0	4	0	1	1
## 15	17	3	0	1	1	0	2	2	4	29	0	2	10	0	1	4	3	10	2	2	3	2	3	1	0	0	2	0	0	0
## 16	19	0	0	1	0	1	4	4	1	31	0	3	12	0	1	4	0	5	0	1	1	1	3	0	0	0	0	0	0	0
## 17	17	0	0	0	1	0	4	0	4	25	1	1	13	0	0	10	1	3	1	1	1	0	0	3	0	0	4	0	0	1
## 18	20	2	0	1	1	0	2	1	3	19	0	3	6	1	1	12	2	3	1	1	2	7	4	2	0	1	1	0	0	0
## 19	18	4	0	1	3	1	4	2	5	29	2	3	26	0	1	10	4	8	3	1	3	1	6	7	2	5	3	1	1	0
## 20	11	1	0	2	3	0	9	3	8	38	3	7	17	2	0	4	2	7	4	1	4	1	6	1	1	1	3	2	3	6
## 21	20	1	0	1	1	1	4	0	6	45	2	4	31	0	1	9	2	8	1	2	1	11	4	3	1	2	0	1	3	0
## 22	16	3	0	0	0	0	4	1	3	24	3	1	5	0	0	7	1	4	1	3	1	2	6	0	0	0	1	0	0	0
## 23	10	2	0	1	1	1	3	2	2	27	0	0	10	1	0	8	1	9	3	0	0	2	2	4	0	1	0	0	0	0
## 24	15	0	0	0	0	0	3	2	1	31	0	1	10	0	2	9	0	4	0	0	3	4	0	1	0	2	0	0	0	0
## 25	9	0	0	0	0	0	6	3	2	31	0	2	8	0	0	3	2	3	1	1	1	2	4	4	0	1	0	0	0	0
## 26	16	0	0	3	0	1	1	1	4	22	0	0	9	0	0	10	2	4	1	1	1	5	1	3	0	2	1	0	0	0
## 27	20	0	0	1	1	0	1	1	3	27	2	4	8	0	0	3	1	4	3	1	3	3	8	2	0	0	2	0	0	0
## 28	10	0	0	0	1	0	0	0	0	10	0	0	2	0	0	1	0	0	0	0	0	0	2	1	0	0	0	0	0	0
## 29	3	0	0	0	0	1	1	0	2	5	1	0	3	0	0	1	1	3	0	0	0	1	1	3	0	0	0	0	0	0
## 30	2	0	0	0	0	0	3	0	1	4	0	0	0	0	0	2	0	2	0	0	1	1	0	1	0	0	0	0	0	0


Tidymodels
https://www.tidymodels.org
Tidymodels
PACKAGES
GET STARTED
LEARN
HELP
CONTRIBUTE
tidymodels
rsample
recipes
parsnip
TUNE
yardstick
TIDYMODELS
The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.
Install tidymodels with:
install.packages("tidymodels")

@juliasilge

```r
library(tidymodels)
```

```
## Registered S3 method overwritten by 'tune':
##   method                   from
##   required_pkgs.model_spec parsnip
```

```
## ── Attaching packages ──────────────────────────────────────────── tidymodels 0.1.4 ──
```

```
## ✓ broom        0.7.9      ✓ rsample      0.1.0
## ✓ dials        0.0.10     ✓ tune         0.1.6
## ✓ infer        1.0.0      ✓ workflows    0.2.4
## ✓ modeldata    0.1.1      ✓ workflowsets 0.1.0
## ✓ parsnip      0.1.7      ✓ yardstick    0.0.8
## ✓ recipes      0.1.17
```

```
## ── Conflicts ───────────────────────────────────────────── tidymodels_conflicts() ──
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## • Dig deeper into tidy modeling with R at https://www.tmwr.org
```

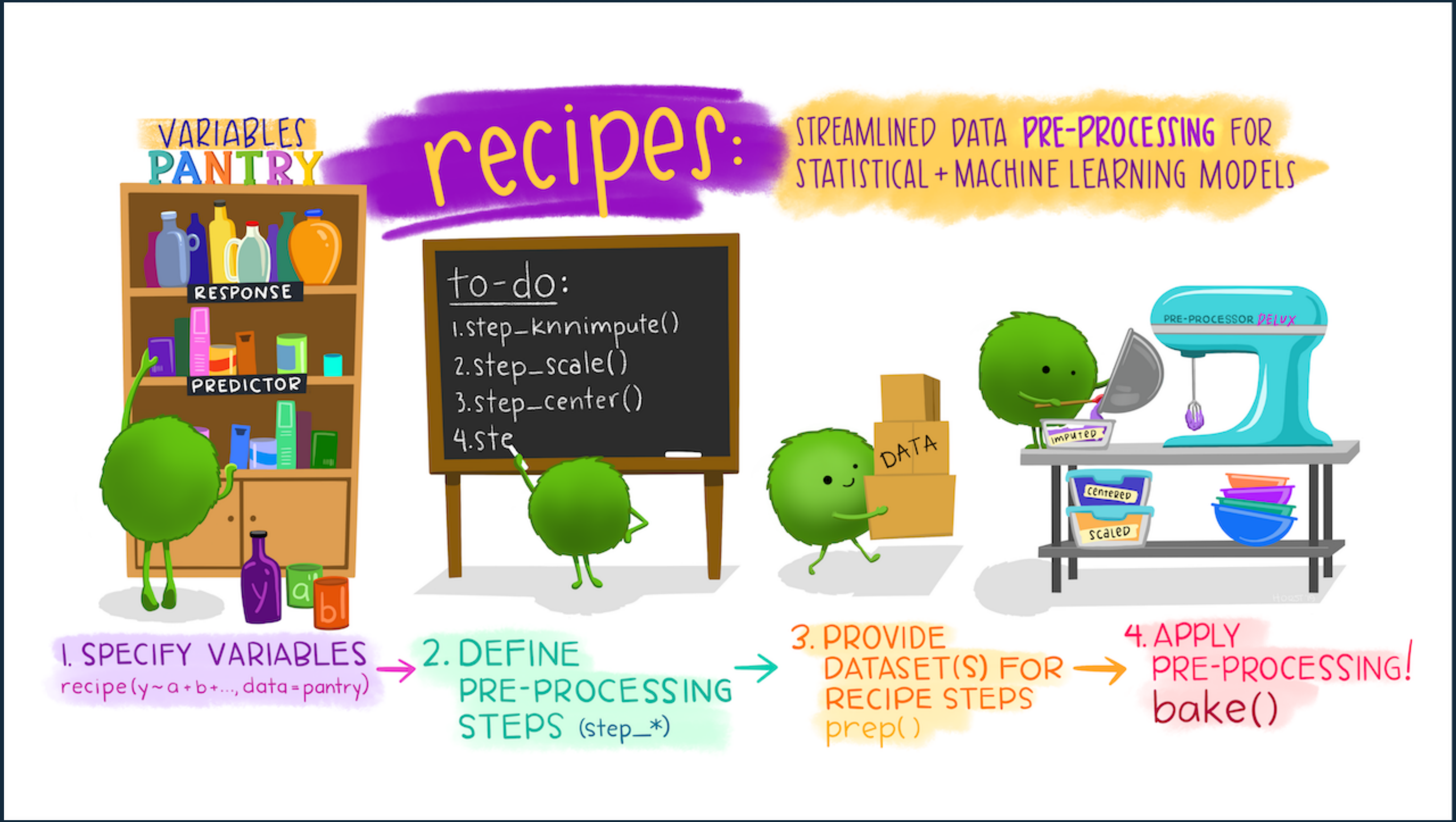
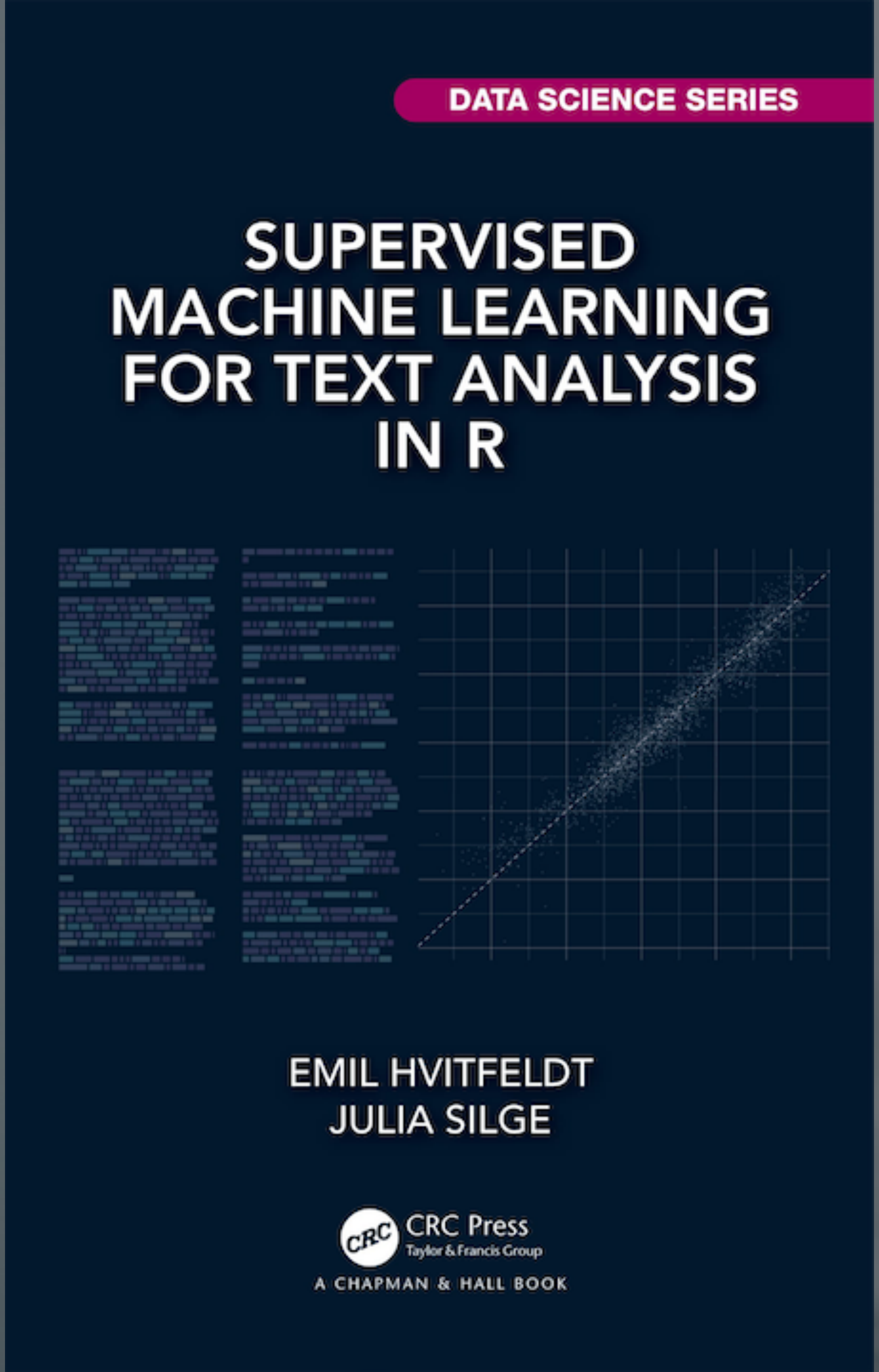
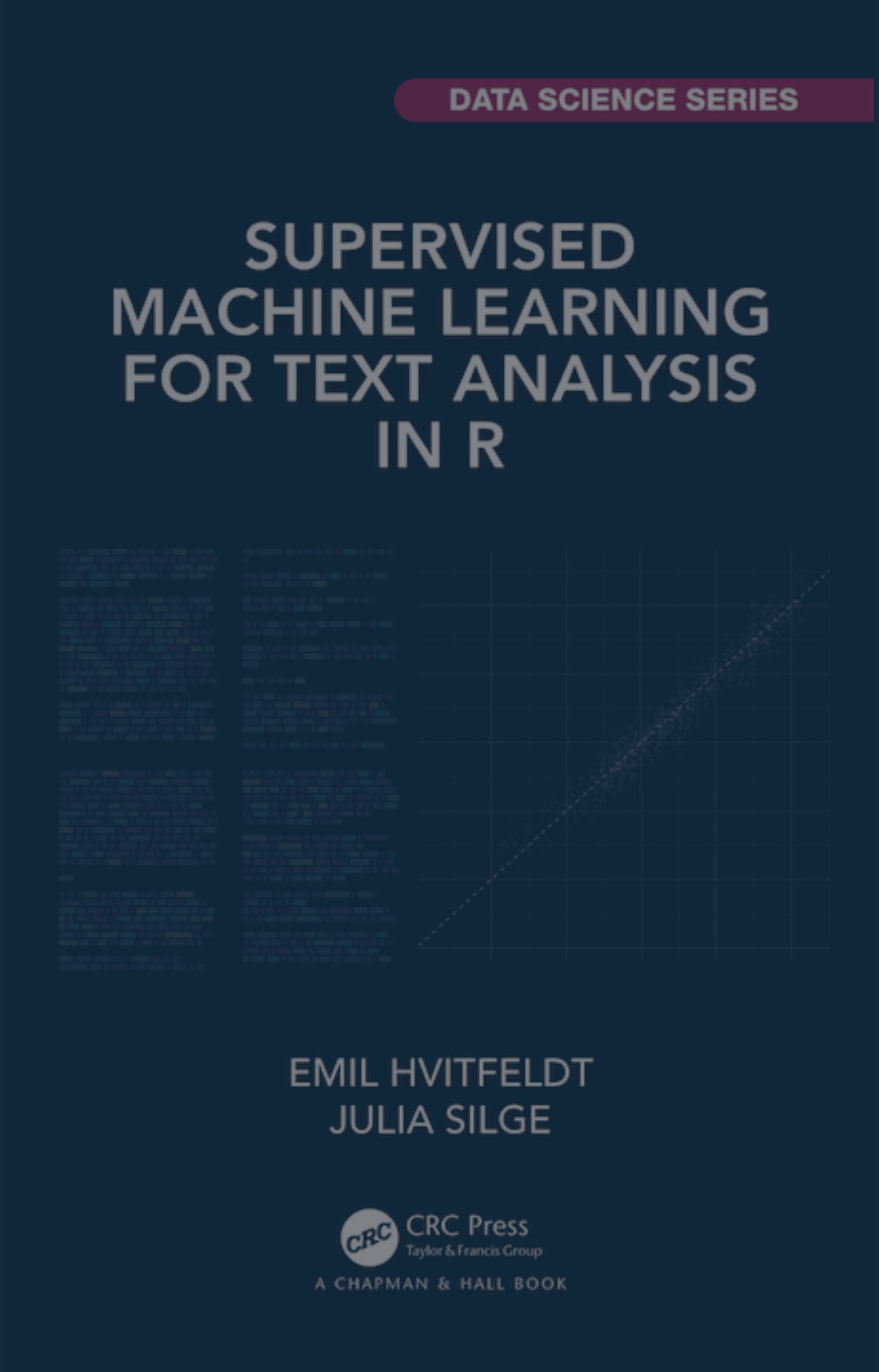


Illustration by Allison Horst

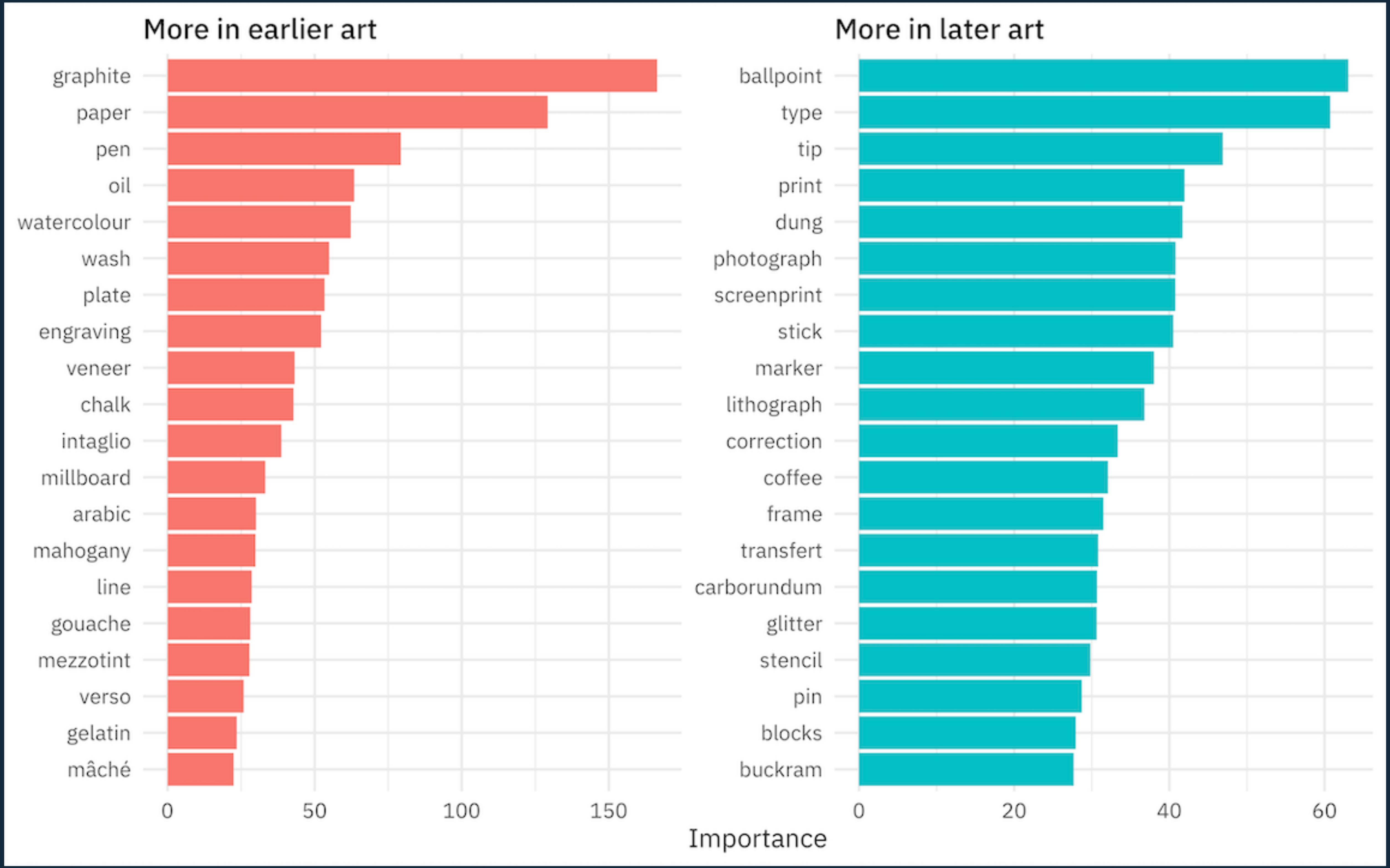

DATA SCIENCE SERIES
SUPERVISED MACHINE LEARNING FOR TEXT ANALYSIS IN R
EMIL HVITFELDT
JULIA SILGE
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



DATA SCIENCE SERIES
SUPERVISED MACHINE LEARNING FOR TEXT ANALYSIS IN R
EMIL HVITFELDT
JULIA SILGE
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

smltar.com

tokenization

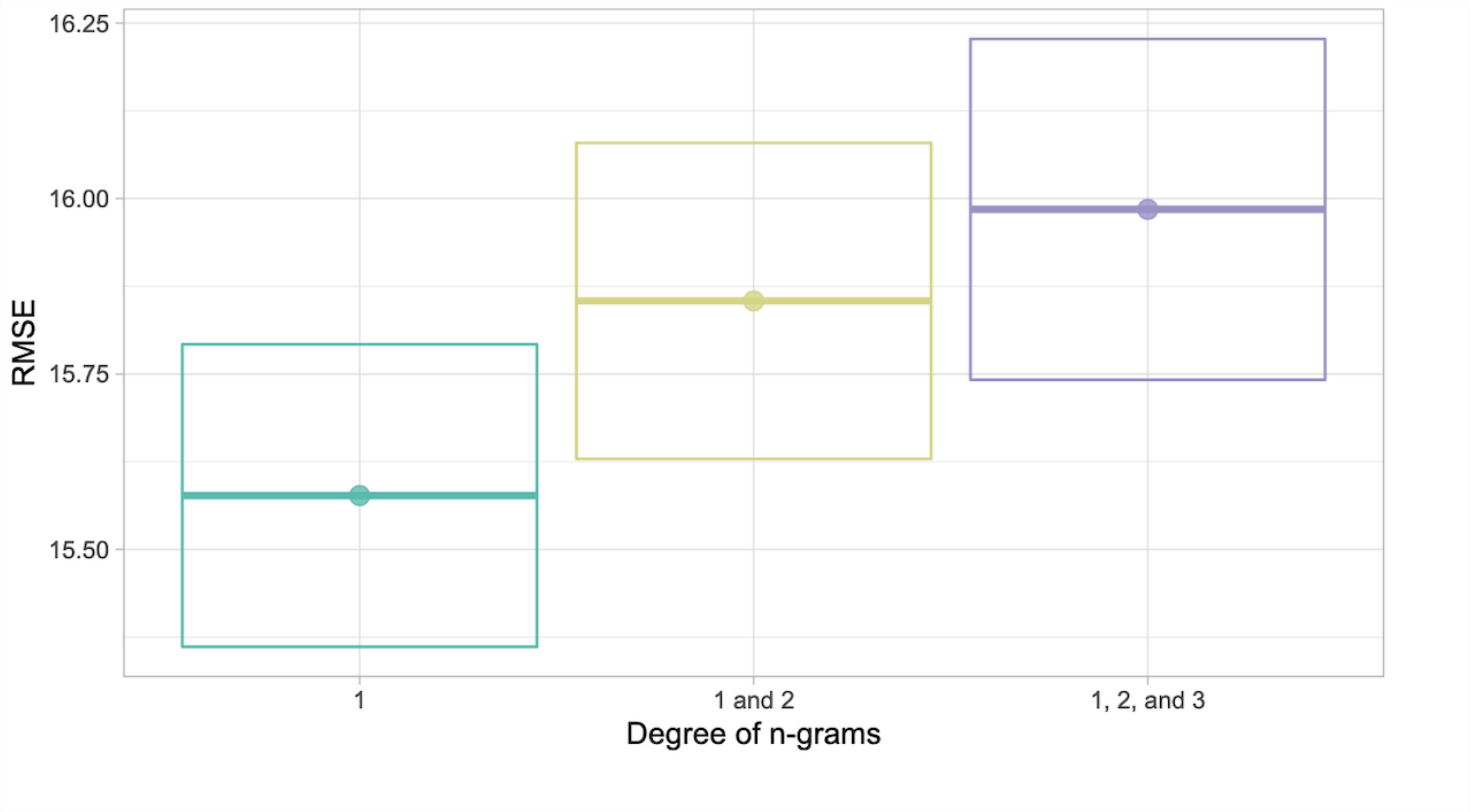
```
##  [1] "the"          "collared"     "peccary"      "also"
##  [5] "referred"     "to"           "as"           "a"
##  [9] "javelina"     "or"           "musk"         "hog"
## [13] "may"          "resemble"     "a"            "pig"
## [17] "however"      "peccaries"    "belong"       "to"
## [21] "a"            "completely"   "different"    "family"
## [25] "than"         "true"         "pigs"         "the"
## [29] "collared"     "peccary"      "belongs"      "to"
## [33] "the"          "tayassuidae"  "family"       "while"
## [37] "pigs"         "belong"       "to"           "the"
## [41] "suidae"
```



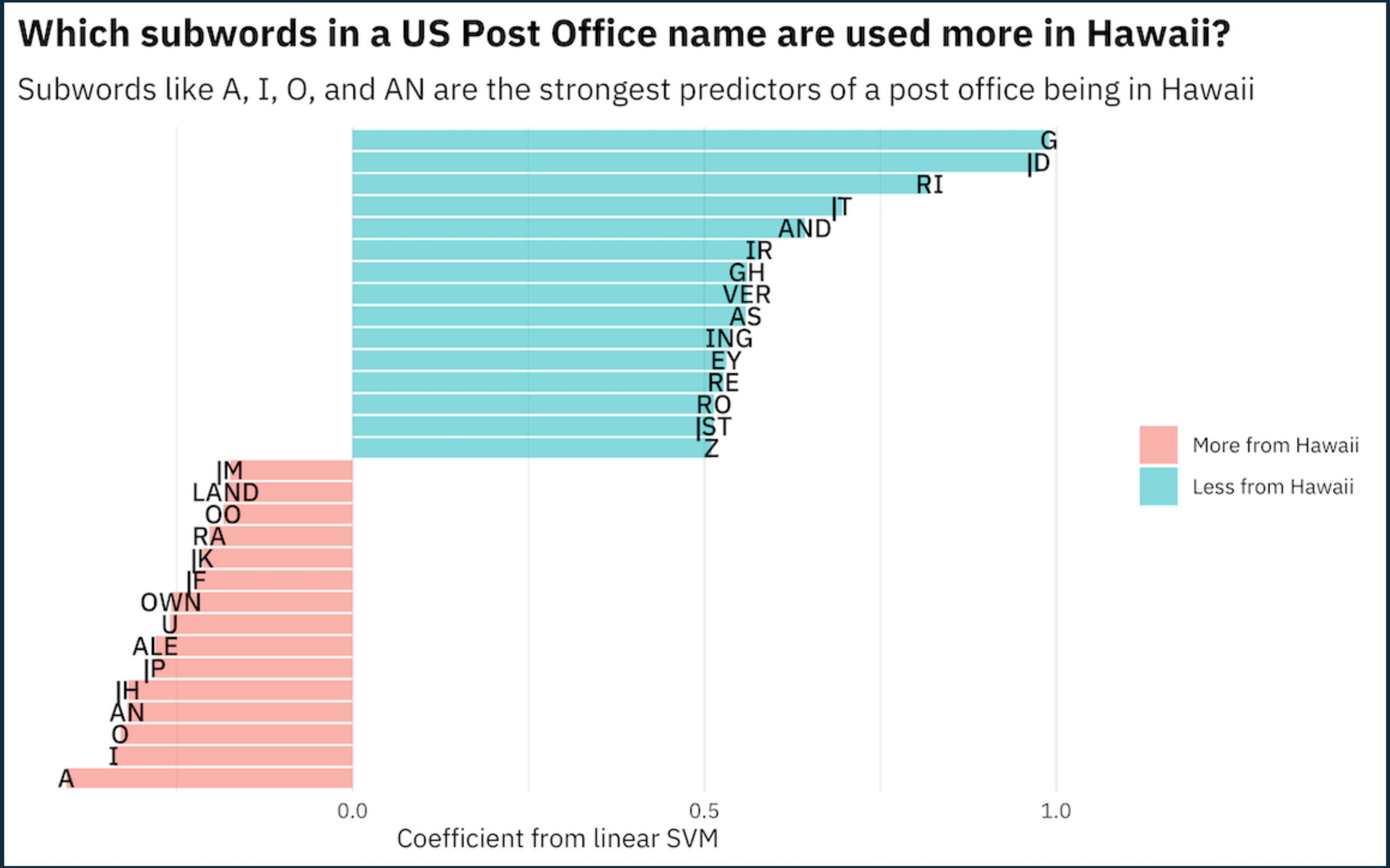
juliasilge.com/blog/tate-collection/

```
##  [1] "the collared"            "collared peccary"      "peccary also"
##  [4] "also referred"           "referred to"           "to as"
##  [7] "as a"                    "a javelina"            "javelina or"
## [10] "or musk"                 "musk hog"              "hog may"
## [13] "may resemble"            "resemble a"            "a pig"
## [16] "pig however"             "however peccaries"     "peccaries belong"
## [19] "belong to"               "to a"                  "a completely"
## [22] "completely different" "different family"      "family than"
## [25] "than true"               "true pigs"             "pigs the"
## [28] "the collared"            "collared peccary"      "peccary belongs"
## [31] "belongs to"              "to the"                "the tayassuidae"
## [34] "tayassuidae family"   "family while"          "while pigs"
## [37] "pigs belong"             "belong to"             "to the"
## [40] "the suidae"
```

```
##  [1] "the collared peccary"        "collared peccary also"       "peccary also referred"       "also referred to"
##  [5] "referred to as"              "to as a"                     "as a javelina"               "a javelina or"
##  [9] "javelina or musk"            "or musk hog"                 "musk hog may"                "hog may resemble"
## [13] "may resemble a"              "resemble a pig"              "a pig however"               "pig however peccaries"
## [17] "however peccaries belong"    "peccaries belong to"         "belong to a"                 "to a completely"
## [21] "a completely different"      "completely different family" "different family than"       "family than true"
## [25] "than true pigs"              "true pigs the"               "pigs the collared"           "the collared peccary"
## [29] "collared peccary belongs"    "peccary belongs to"          "belongs to the"              "to the tayassuidae"
## [33] "the tayassuidae family"      "tayassuidae family while"    "family while pigs"           "while pigs belong"
## [37] "pigs belong to"              "belong to the"               "to the suidae"
```


16.25
16.00
15.75
15.50
RMSE
1
1 and 2
1, 2, and 3
Degree of n-grams

```
##   [1] "the" "hec" "eco" "col" "oll" "lla" "lar" "are" "red" "edp" "dpe" "pec"
##  [13] "ecc" "cca" "car" "ary" "rya" "yal" "als" "lso" "sor" "ore" "ref" "efe"
##  [25] "fer" "err" "rre" "red" "edt" "dto" "toa" "oas" "asa" "saj" "aja" "jav"
##  [37] "ave" "vel" "eli" "lin" "ina" "nao" "aor" "orm" "rmu" "mus" "usk" "skh"
##  [49] "kho" "hog" "ogm" "gma" "may" "ayr" "yre" "res" "ese" "sem" "emb" "mbl"
##  [61] "ble" "lea" "eap" "api" "pig" "igh" "gho" "how" "owe" "wev" "eve" "ver"
##  [73] "erp" "rpe" "pec" "ecc" "cca" "car" "ari" "rie" "ies" "esb" "sbe" "bel"
##  [85] "elo" "lon" "ong" "ngt" "gto" "toa" "oac" "aco" "com" "omp" "mpl" "ple"
##  [97] "let" "ete" "tel" "ely" "lyd" "ydi" "dif" "iff" "ffe" "fer" "ere" "ren"
## [109] "ent" "ntf" "tfa" "fam" "ami" "mil" "ily" "lyt" "yth" "tha" "han" "ant"
## [121] "ntr" "tru" "rue" "uep" "epi" "pig" "igs" "gst" "sth" "the" "hec" "eco"
## [133] "col" "oll" "lla" "lar" "are" "red" "edp" "dpe" "pec" "ecc" "cca" "car"
## [145] "ary" "ryb" "ybe" "bel" "elo" "lon" "ong" "ngs" "gst" "sto" "tot" "oth"
## [157] "the" "het" "eta" "tay" "aya" "yas" "ass" "ssu" "sui" "uid" "ida" "dae"
## [169] "aef" "efa" "fam" "ami" "mil" "ily" "lyw" "ywh" "whi" "hil" "ile" "lep"
## [181] "epi" "pig" "igs" "gsb" "sbe" "bel" "elo" "lon" "ong" "ngt" "gto" "tot"
## [193] "oth" "the" "hes" "esu" "sui" "uid" "ida" "dae"
```



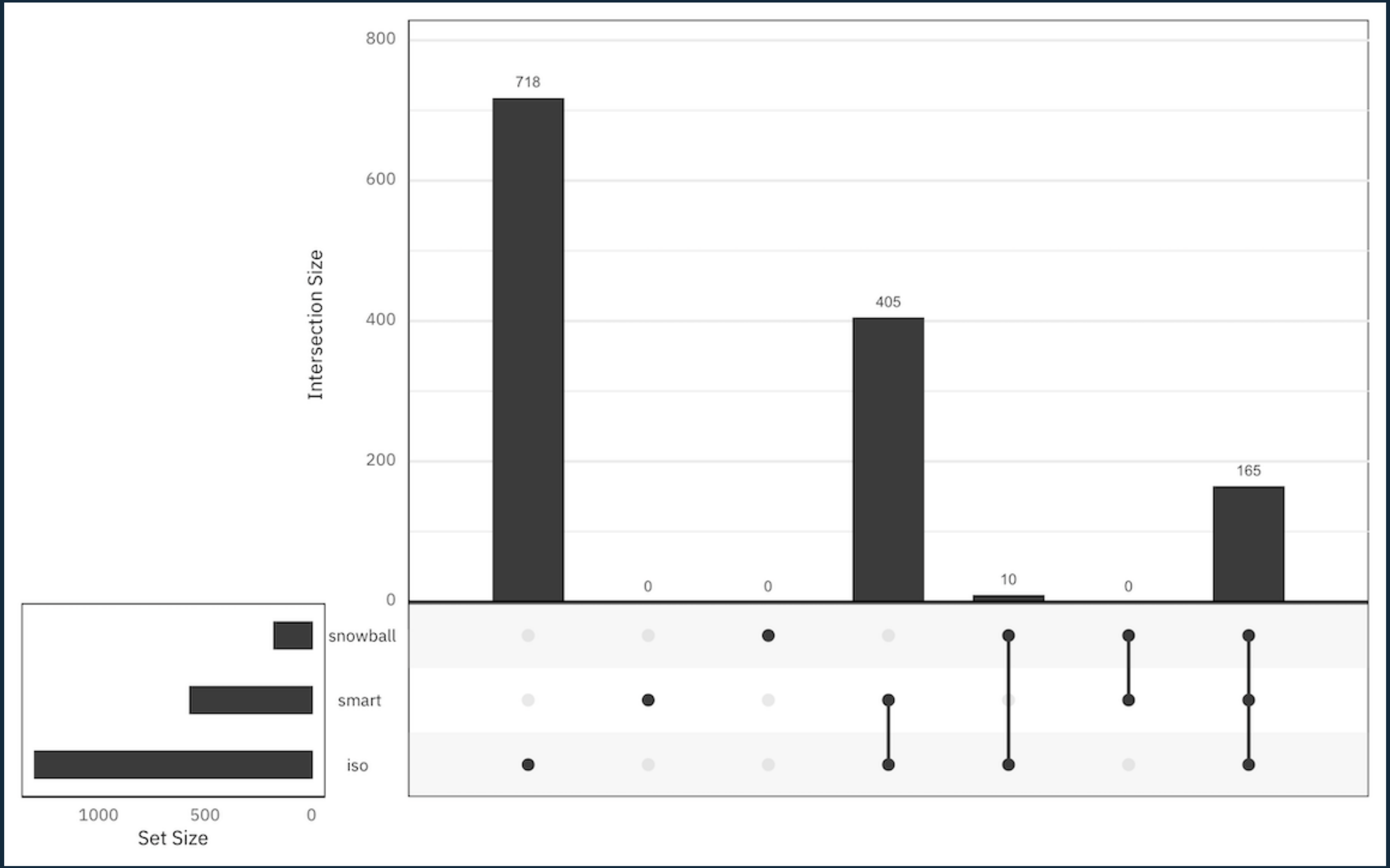
juliasilge.com/blog/hawaii-post-offices/

```r
library(textrecipes)
recipe(diet ~ text, data = animal_train) %>%
    step_tokenize(
        text,
        token = "ngrams",
        options = list(n = 3, n_min = 1)
    )
```

```
## Recipe
##
## Inputs:
##
##        role #variables
##     outcome          1
##   predictor          1
##
## Operations:
##
## Tokenization for text
```

stop words

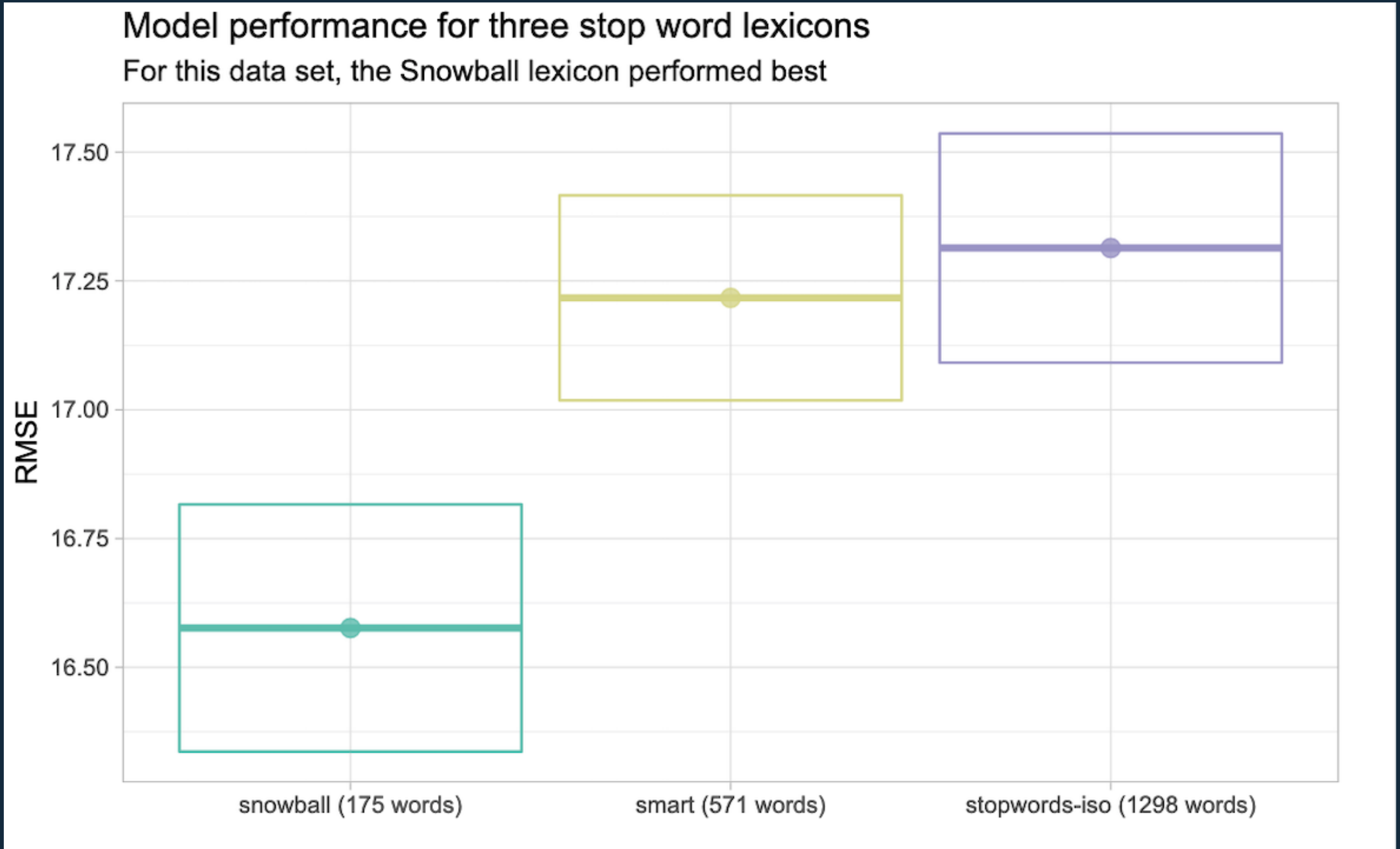
```
##   [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"
##   [8] "ourselves"  "you"        "your"       "yours"      "yourself"   "yourselves" "he"
##  [15] "him"        "his"        "himself"    "she"        "her"        "hers"       "herself"
##  [22] "it"         "its"        "itself"     "they"       "them"       "their"      "theirs"
##  [29] "themselves" "what"       "which"      "who"        "whom"       "this"       "that"
##  [36] "these"      "those"      "am"         "is"         "are"        "was"        "were"
##  [43] "be"         "been"       "being"      "have"       "has"        "had"        "having"
##  [50] "do"         "does"       "did"        "doing"      "would"      "should"     "could"
##  [57] "ought"      "i'm"        "you're"     "he's"       "she's"      "it's"       "we're"
##  [64] "they're"    "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"
##  [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"       "you'll"     "he'll"
##  [78] "she'll"     "we'll"      "they'll"    "isn't"      "aren't"     "wasn't"     "weren't"
##  [85] "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"     "won't"
##  [92] "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"
##  [99] "let's"      "that's"     "who's"      "what's"     "here's"     "there's"    "when's"
## [106] "where's"    "why's"      "how's"      "a"          "an"         "the"        "and"
## [113] "but"        "if"         "or"         "because"    "as"         "until"      "while"
## [120] "of"         "at"         "by"         "for"        "with"       "about"      "against"
## [127] "between"    "into"       "through"    "during"     "before"     "after"      "above"
## [134] "below"      "to"         "from"       "up"         "down"       "in"         "out"
## [141] "on"         "off"        "over"       "under"      "again"      "further"    "then"
## [148] "once"       "here"       "there"      "when"       "where"      "why"        "how"
## [155] "all"        "any"        "both"       "each"       "few"        "more"       "most"
## [162] "other"      "some"       "such"       "no"         "nor"        "not"        "only"
## [169] "own"        "same"       "so"         "than"       "too"        "very"       "will"
```


800
600
400
200
0
Intersection Size
718
0
0
405
10
0
165
snowball
smart
iso
1000
500
0
Set Size

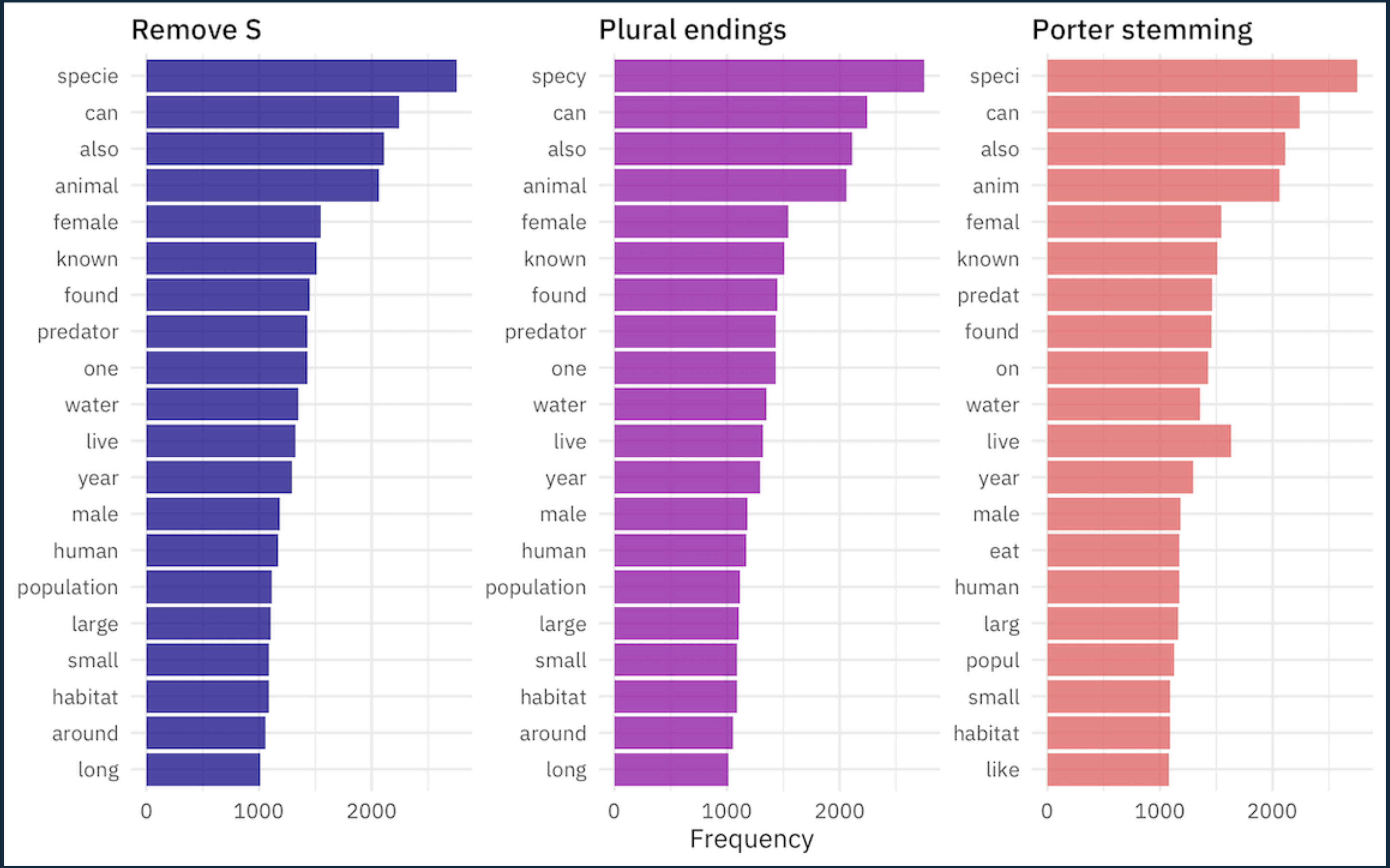
```
##  [1] "she's"    "he'd"
##  [3] "she'd"    "he'll"
##  [5] "she'll"   "shan't"
##  [7] "mustn't" "when's"
##  [9] "why's"    "how's"
```

```
recipe(diet ~ text, data = animal_train) %>%
    step_tokenize(text) %>%
    step_stopwords(text)
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          1
##
## Operations:
##
## Tokenization for text
## Stop word removal for text
```


Model performance for three stop word lexicons
For this data set, the Snowball lexicon performed best
RMSE
17.50
17.25
17.00
16.75
16.50
snowball (175 words)
smart (571 words)
stopwords-iso (1298 words)

stemming


Remove S
specie
can
also
animal
female
known
found
predator
one
water
live
year
male
human
population
large
small
habitat
around
long
Plural endings
specy
can
also
animal
female
known
found
predator
one
water
live
year
male
human
population
large
small
habitat
around
long
Porter stemming
speci
can
also
anim
femal
known
predat
found
on
water
live
year
male
eat
human
larg
popul
small
habitat
like
0
1000
2000
Frequency

@juliasilge

```r
tidy_animals %>%
    count(animal, word) %>%
    cast_dfm(animal, word, n)
```

```
## Document-feature matrix of: 610 documents, 16,840 features (98.13% sparse) and 0 docvars.
```

```r
tidy_animals %>%
    mutate(stem = wordStem(word)) %>%
    count(animal, stem) %>%
    cast_dfm(animal, stem, n)
```

```
## Document-feature matrix of: 610 documents, 12,045 features (97.62% sparse) and 0 docvars.
```

recall

true positive rate

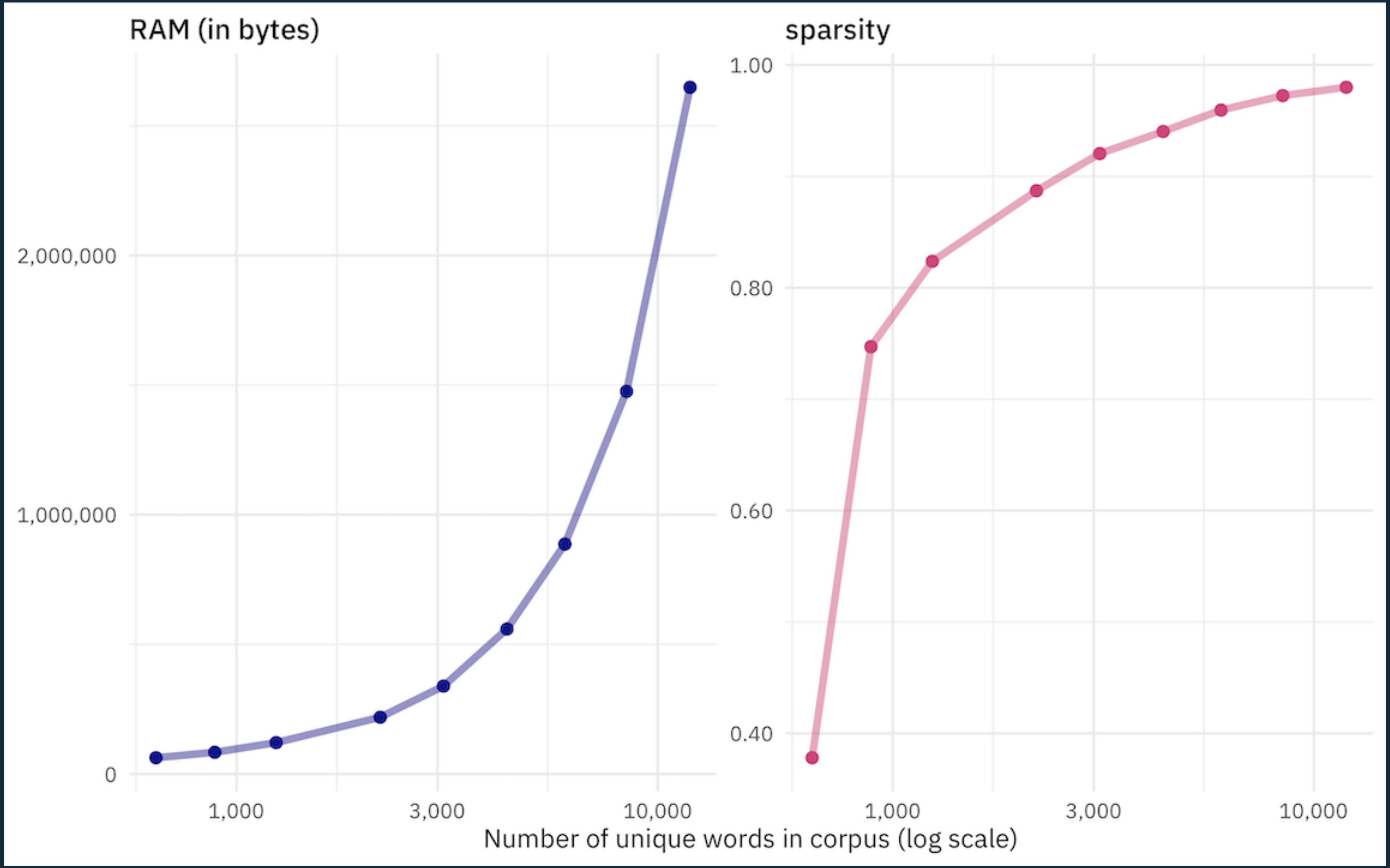
precision

true negative rate

@juliasilge

```r
recipe(diet ~ text, data = animal_train) %>%
    step_tokenize(
        text,
        token = "ngrams",
        options = list(
            n = 2, n_min = 1,
            stopwords = stopwords::stopwords(source = "snowball")
        )
    ) %>%
    step_tokenfilter(text, max_tokens = tune()) %>%
    step_tfidf(text)
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          1
##
## Operations:
##
## Tokenization for text
## Text filtering for text
## Term frequency-inverse document frequency with text
```


RAM (in bytes)
2,000,000
1,000,000
0
sparsity
1.00
0.80
0.60
0.40
1,000
3,000
10,000
1,000
3,000
10,000
Number of unique words in corpus (log scale)

@juliasilge

You shall know a word by the company it keeps.

— John Rupert Firth

word	*distance*
month	1
year	0.607
months	0.593
monthly	0.454
installments	0.446
payment	0.429
week	0.406
weeks	0.400
85.00	0.399
bill	0.396

word	*distance*
error	1
mistake	0.683
clerical	0.627
problem	0.582
glitch	0.580
errors	0.571
miscommunication	0.512
misunderstanding	0.486
issue	0.478
discrepancy	0.474

word	***distance***
error	1
errors	0.792
mistake	0.664
correct	0.621
incorrect	0.613
fault	0.607
difference	0.594
mistakes	0.586
calculation	0.584
probability	0.583

Fairness and word embeddings

- African American first names are associated with more unpleasant feelings than European American first names
- Women's first names are more associated with family and men's first names are more associated with career
- Terms associated with women are more associated with the arts and terms associated with men are more associated with science

Features from text machine learning in the real world

@juliasilge

Thank you!

Julia Silge

juliasilge.com | smltar.com

Photo by Sharon McCutcheon on Unsplash