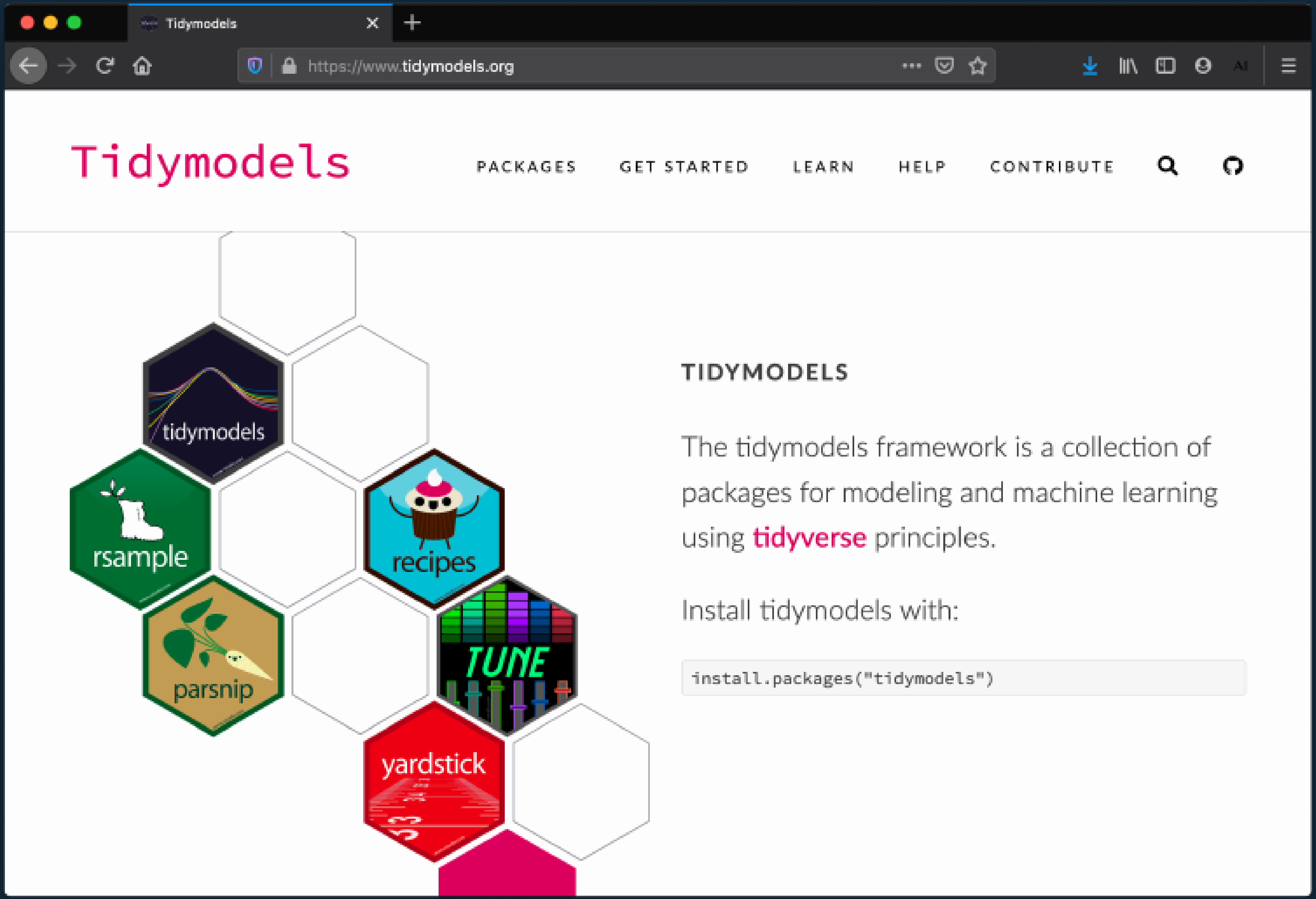
Creating features for machine learning from text

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)

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

## ✓ broom        0.7.10      ✓ 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
```

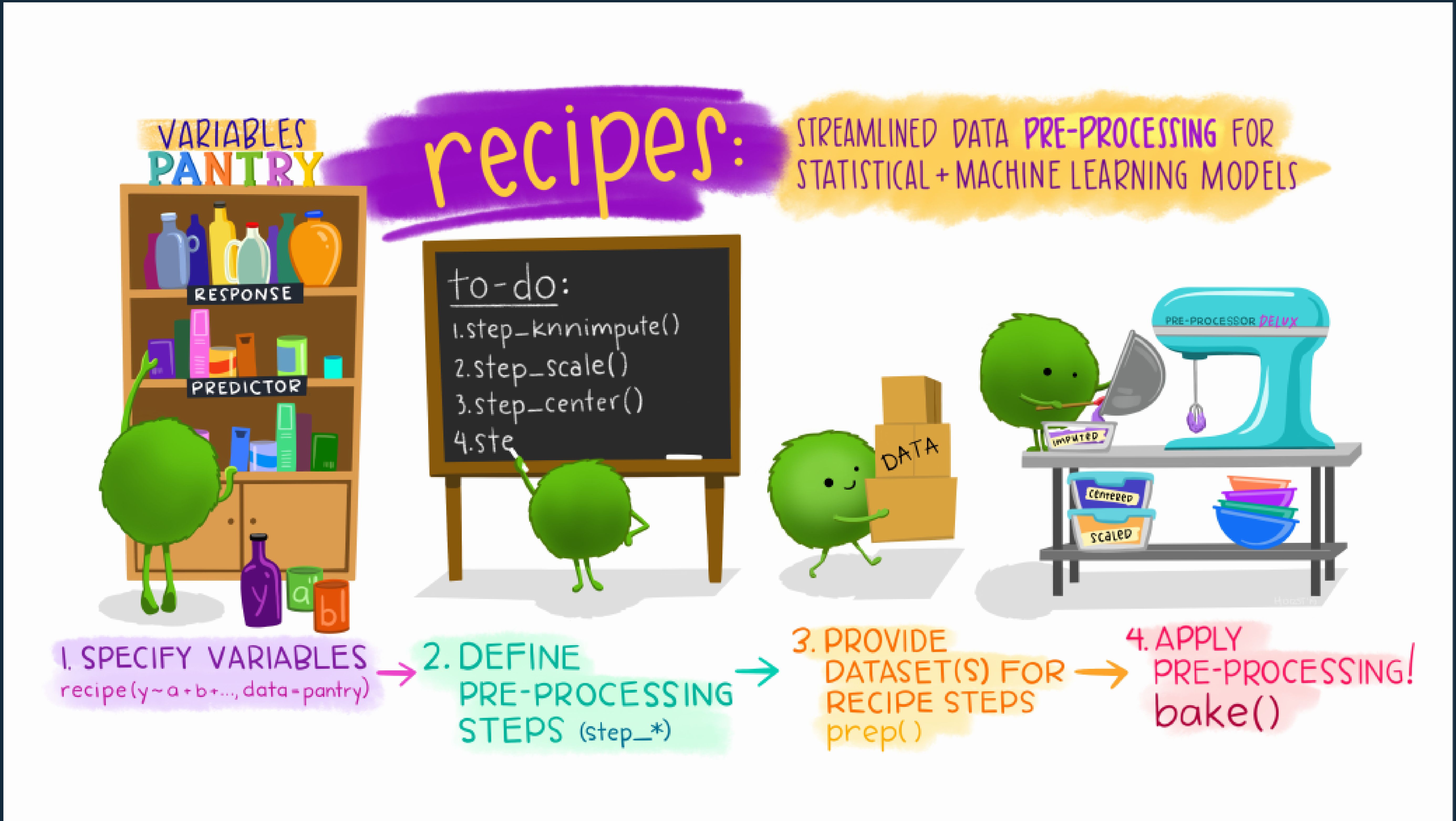
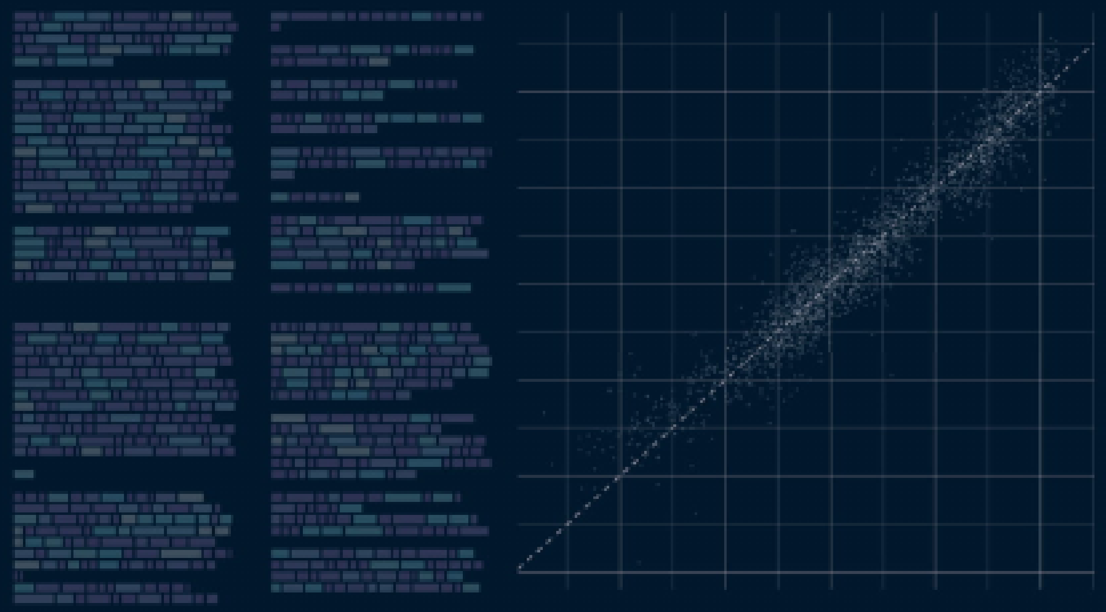


Illustration by Allison Horst

DATA SCIENCE SERIES

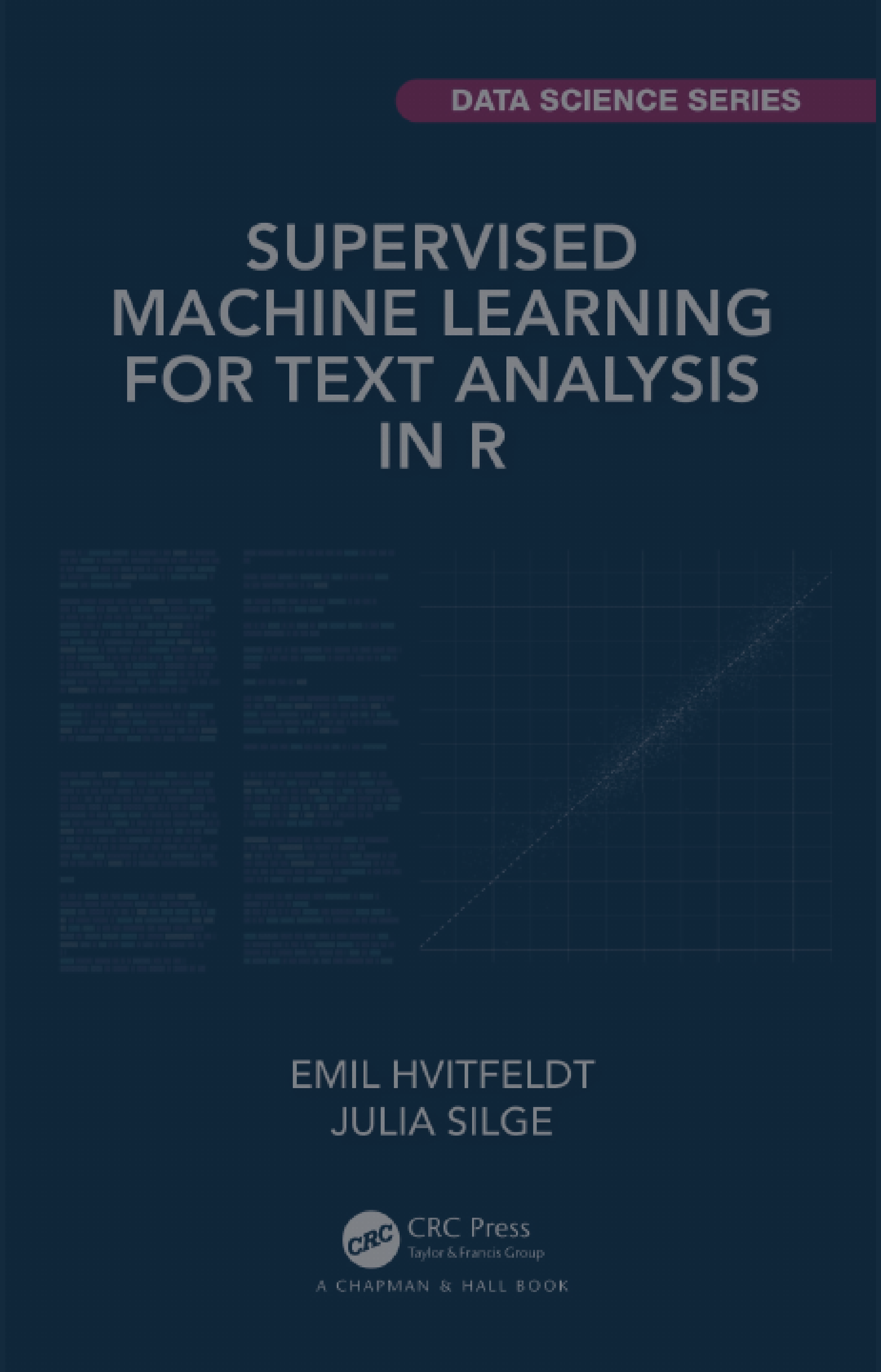
SUPERVISED MACHINE LEARNING FOR TEXT ANALYSIS IN R



EMIL HVITFELDT
JULIA SILGE

CRC Press
Taylor & Francis Group

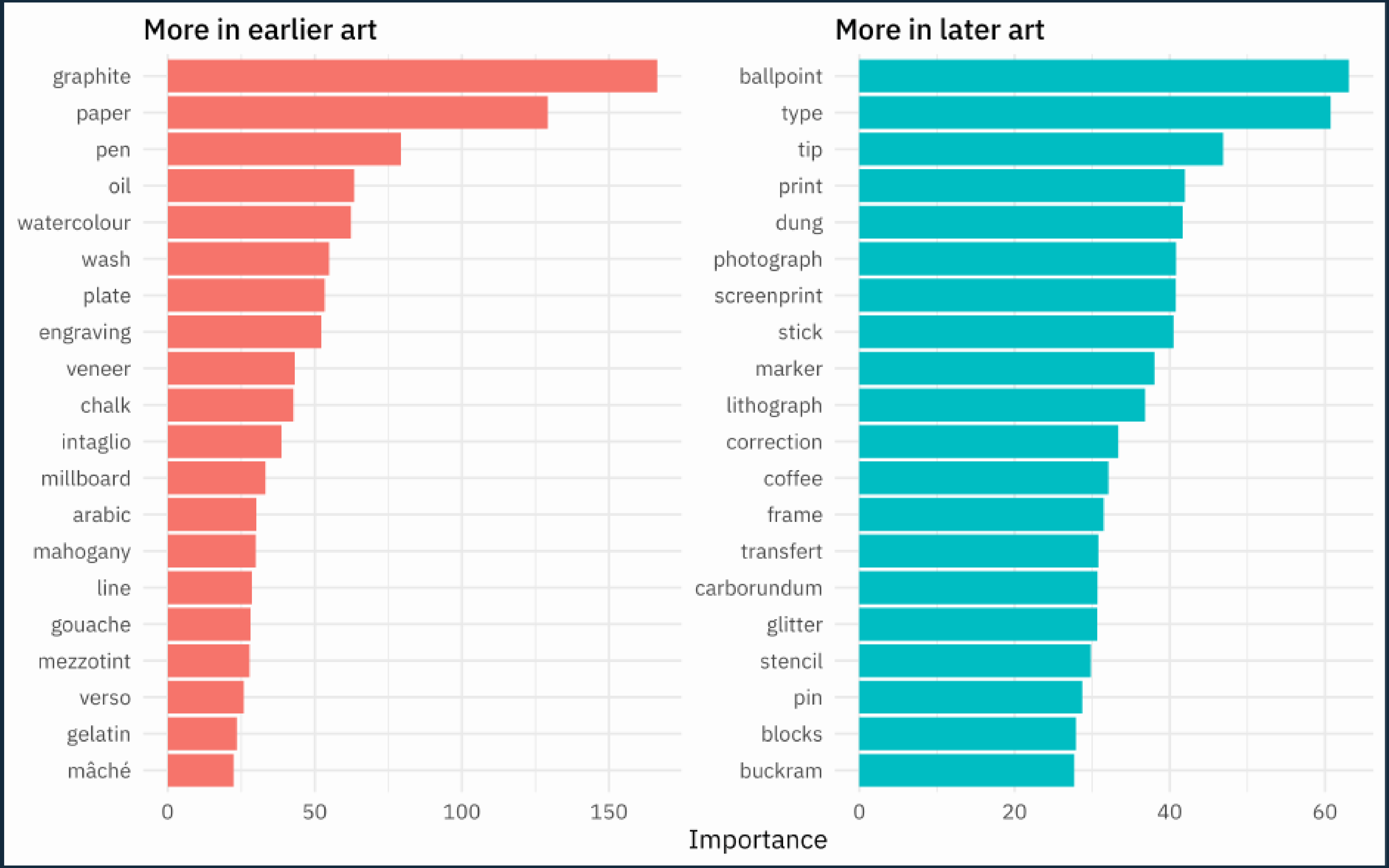
A CHAPMAN & HALL BOOK


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

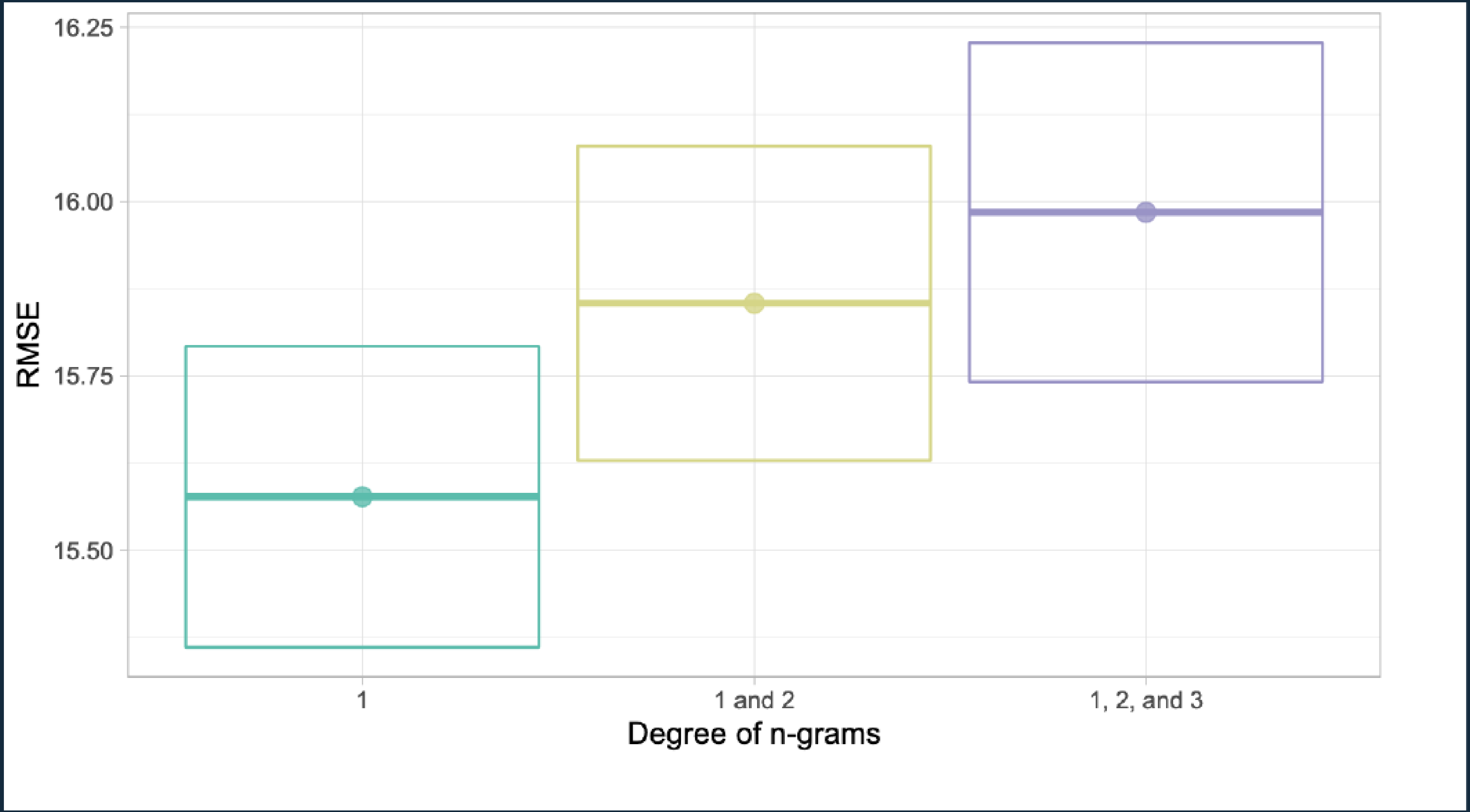
```
##  [1] "the"          "collared"     "peccary"      "also"
##  [5] "referred"     "to"           "as"           "a"
##  [9] "javelina"     "or"           "musk"         "hog"
## [13] "may"          "resemble"     "a"            "pig"
## [17] "however"      "peccaries"    "belong"       "to"
## [21] "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"
```

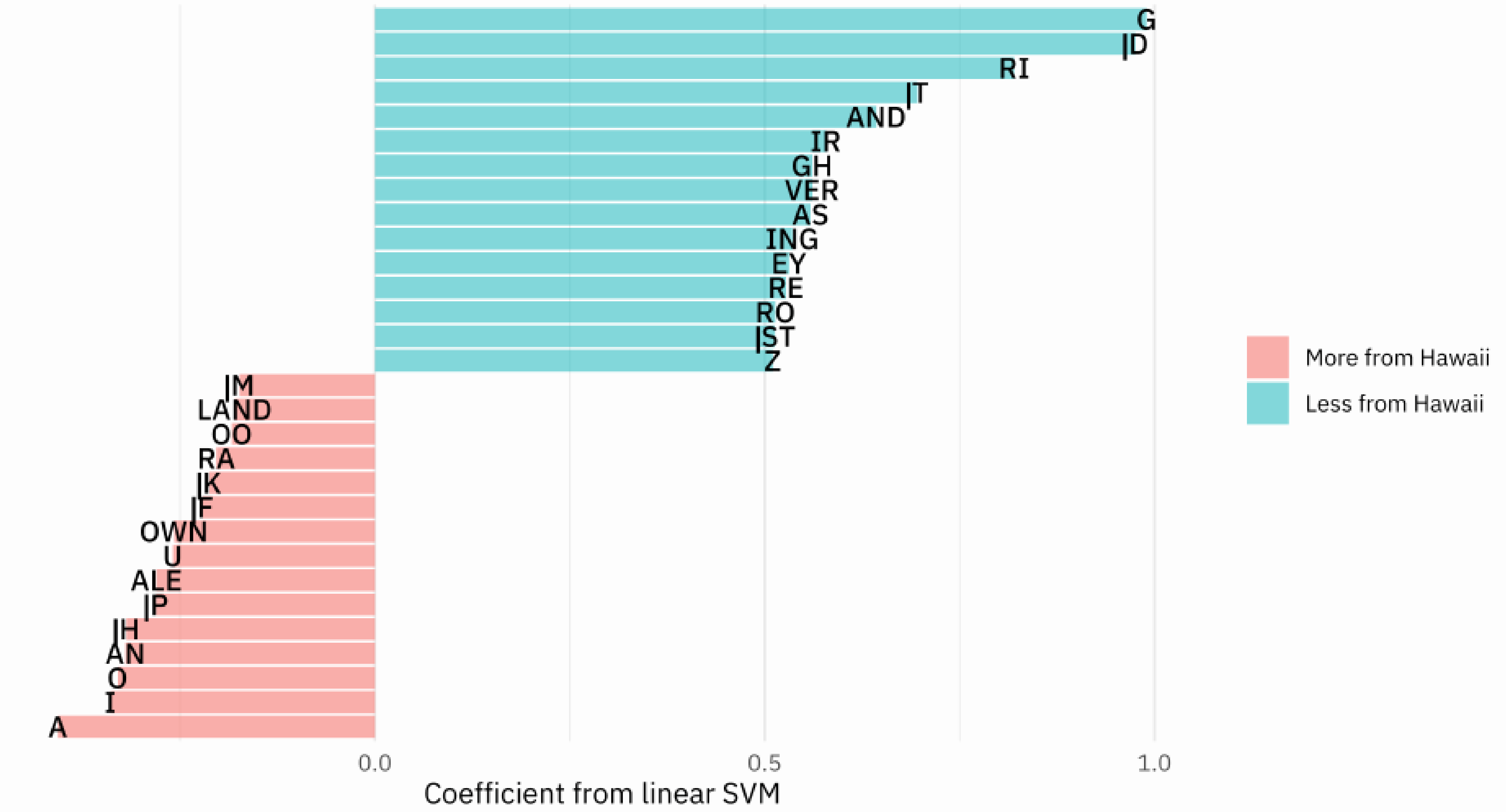
```
##  [1] "the collared peccary"        "collared peccary also"       "peccary also referred"
##  [4] "also referred to"            "referred to as"              "to as a"
##  [7] "as a javelina"               "a javelina or"               "javelina or musk"
## [10] "or musk hog"                 "musk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

Which subwords in a US Post Office name are used more in Hawaii?

Subwords like A, I, O, and AN are the strongest predictors of a post office being in Hawaii



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

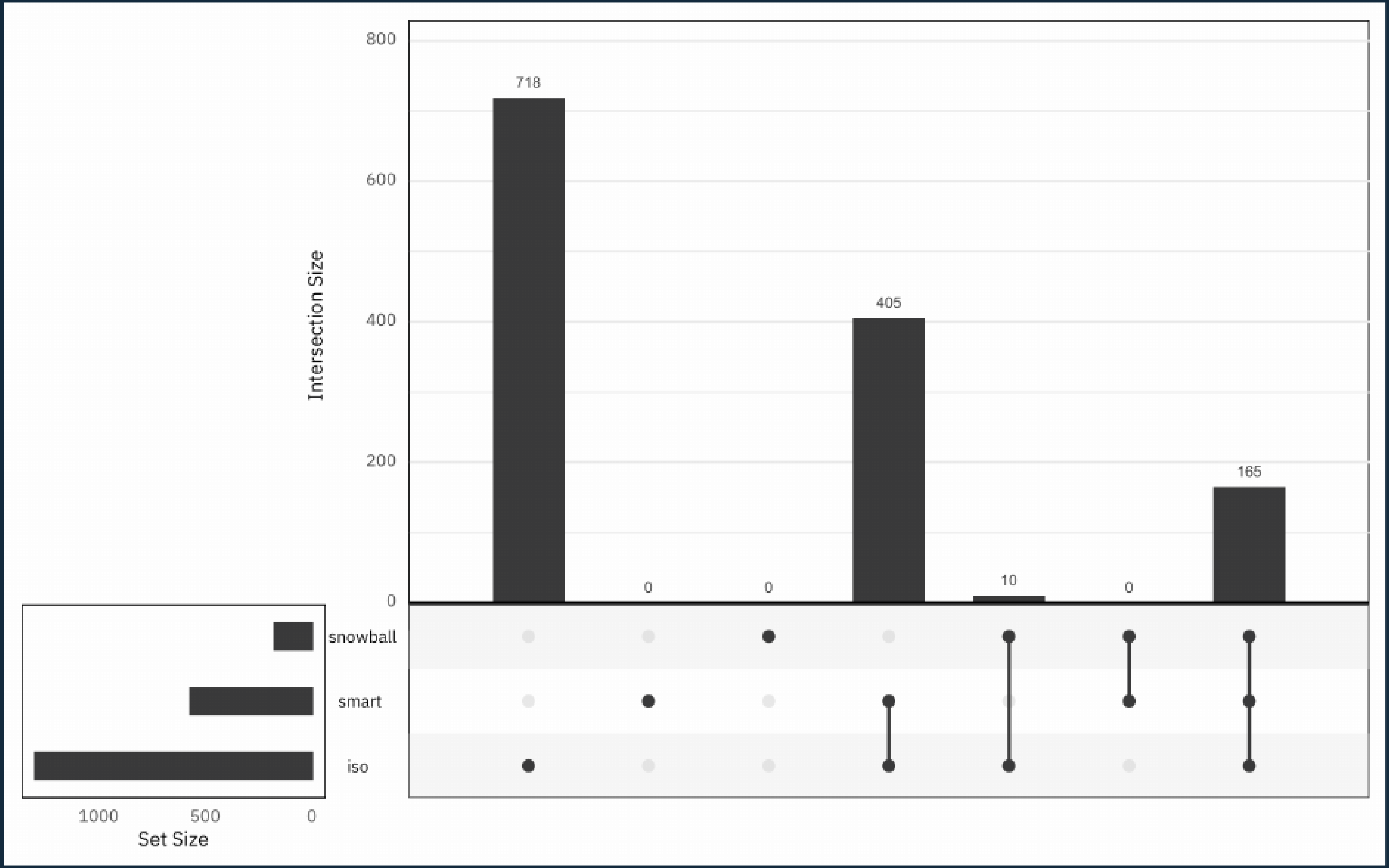
```r
library(textrecipes)
recipe(diet ~ text, data = animal_train) %>%
    step_tokenize(
        text,
        token = "ngrams",
        options = list(n = 3, n_min = 1)
    )
```

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

stop words

```
##   [1] "가"              "가까스로"          "가령"             "각"               "각각"             "각자"             "각종"             "갖고말하자면"      "같다"             "같이"             "개의치않고"        "거니와"
##  [13] "거바"            "거의"              "것"               "것과 같이"         "것들"             "게다가"           "게우다"           "겨우"             "견지에서"          "결과에 이르다"     "결국"             "결론을 낼 수 있다"
##  [25] "겸사겸사"        "고려하면"          "고로"             "곧"               "공동으로"          "과"               "과연"             "관계가 있다"       "관계없이"          "관련이 있다"       "관하여"           "관한"
##  [37] "관해서는"        "구"                "구체적으로"        "구토하다"          "그"               "그들"             "그때"             "그래"             "그래도"           "그래서"           "그러나"           "그러니"
##  [49] "그러니까"        "그러면"            "그러므로"          "그러한즉"          "그런 까닭에"       "그런데"           "그런즉"           "그럼"             "그럼에도 불구하고"  "그렇게 함으로써"    "그렇지"           "그렇지 않다면"
##  [61] "그렇지 않으면"    "그렇지만"          "그렇지않으면"      "그리고"            "그리하여"          "그만이다"          "그에 따르는"       "그위에"           "그저"             "그중에서"          "그치지 않다"       "근거로"
##  [73] "근거하여"        "기대여"            "기점으로"          "기준으로"          "기타"             "까닭으로"          "까악"             "까지"             "까지 미치다"       "까지도"           "꽈당"             "끙끙"
##  [85] "끼익"            "나"                "나머지는"          "남들"             "남짓"             "너"               "너희"             "너희들"           "네"               "넷"               "년"               "논하지 않다"
##  [97] "놀라다"          "누가 알겠는가"      "누구"             "다른"             "다른 방면으로"     "다만"             "다섯"             "다소"             "다수"             "다시 말하자면"     "다시말하면"        "다음"
## [109] "다음에"          "다음으로"          "단지"             "답다"             "당신"             "당장"             "대로 하다"         "대하면"           "대하여"           "대해 말하자면"     "대해서"           "댕그"
## [121] "더구나"          "더군다나"          "더라도"           "더불어"           "더욱더"           "더욱이는"          "도달하다"          "도착하다"          "동시에"           "동안"             "된바에야"          "된이상"
## [133] "두번째로"        "둘"                "둥둥"             "뒤따라"           "뒤이어"           "든간에"           "들"               "등"               "등등"             "딩동"             "따라"             "따라서"
## [145] "따위"            "따지지 않다"        "딱"               "때"               "때가 되어"         "때문에"           "또"               "또한"             "뚝뚝"             "라 해도"          "령"               "로"
## [157] "로 인하여"        "로부터"            "로써"             "륙"               "를"               "마음대로"          "마저"             "마저도"           "마치"             "막론하고"          "만 못하다"         "만약"
## [169] "만약에"          "만은 아니다"        "만이 아니다"       "만일"             "만큼"             "말하자면"          "말할것도 없고"     "매"               "매번"             "메쓰겁다"          "몇"               "모"
## [181] "모두"            "무렵"              "무릎쓰고"          "무슨"             "무엇"             "무엇때문에"        "물론"             "및"               "바꾸어말하면"      "바꾸어말하자면"     "바꾸어서 말하면"    "바꾸어서 한다면"
## [193] "바꿔 말하면"      "바로"              "바와같이"          "밖에 안된다"       "반대로"           "반대로 말하자면"    "반드시"           "버금"             "보는데서"          "보다더"           "보드득"           "본대로"
## [205] "봐"              "봐라"              "부류의 사람들"     "부터"             "불구하고"          "불문하고"          "붕붕"             "비걱거리다"        "비교적"           "비길수 없다"       "비로소"           "비록"
## [217] "비슷하다"        "비추어 보아"        "비하면"           "뿐만 아니라"       "뿐만아니라"        "뿐이다"           "삐걱"             "삐걱거리다"        "사"               "삼"               "상대적으로 말하자면" "생각한대로"
## [229] "설령"            "설마"              "설사"             "셋"               "소생"             "소인"             "쉬"               "쉿"               "습니까"           "습니다"           "시각"             "시간"
## [241] "시작하여"        "시초에"            "시키다"           "실로"             "심지어"           "아"               "아니"             "아니나다를가"      "아니라면"          "아니면"           "아니었다면"        "아래윗"
## [253] "아무거나"        "아무도"            "아야"             "아울러"           "아이"             "아이고"           "아이구"           "아이야"           "아이쿠"           "아하"             "아홉"             "안 그러면"
## [265] "않기 위하여"      "않기 위해서"        "알 수 있다"        "알았어"           "앗"               "앞에서"           "앞의것"           "야"               "약간"             "양자"             "어"               "어기여차"
## [277] "어느"            "어느 년도"          "어느것"           "어느곳"           "어느때"           "어느쪽"           "어느해"           "어디"             "어때"             "어떠한"           "어떤"             "어떤것"
## [289] "어떤것들"        "어떻게"            "어떻해"           "어이"             "어째서"           "어쨋든"           "어쩔수 없다"       "어찌"             "어찌됏든"          "어찌됏어"          "어찌하든지"        "어찌하여"
## [301] "언제"            "언젠가"            "얼마"             "얼마 안 되는 것"   "얼마간"           "얼마나"           "얼마든지"          "얼마만큼"          "얼마큼"           "엉엉"             "에"               "에 가서"
## [313] "에 달려 있다"     "에 대해"           "에 있다"          "에 한하다"         "에게"             "에서"             "여"               "여기"             "여덟"             "여러분"           "여보시오"          "여부"
## [325] "여섯"            "여전히"            "여차"             "연관되다"          "연이서"           "영"               "영차"             "옆사람"           "예"               "예를 들면"         "예를 들자면"       "예컨대"
## [337] "예하면"          "오"                "오로지"           "오르다"           "오자마자"          "오직"             "오호"             "오히려"           "와"               "와 같은 사람들"    "와르르"           "와아"
## [349] "왜"              "왜냐하면"          "외에도"           "요만큼"           "요만한 것"         "요만한걸"          "요컨대"           "우르르"           "우리"             "우리들"           "우선"             "우에 종합한것과같이"
## [361] "운운"            "월"                "위에서 서술한바와같이" "위하여"        "위해서"           "윙윙"             "육"               "으로"             "으로 인하여"       "으로서"           "으로써"           "을"
## [373] "응"              "응당"              "의"               "의거하여"          "의지하여"          "의해"             "의해되다"          "의해서"           "이"               "이 되다"          "이 때문에"         "이 밖에"
## [385] "이 외에"          "이 정도의"          "이것"             "이곳"             "이때"             "이라면"           "이래"             "이러이러하다"      "이러한"           "이런"             "이럴정도로"        "이렇게 많은 것"
## [397] "이렇게되면"      "이렇게말하자면"     "이렇구나"          "이로 인하여"       "이르기까지"        "이리하여"          "이만큼"           "이번"             "이봐"             "이상"             "이어서"           "이었다"
## [409] "이와 같다"        "이와 같은"          "이와 반대로"       "이와같다면"        "이외에도"          "이용하여"          "이유만으로"        "이젠"             "이지만"           "이쪽"             "이천구"           "이천육"
## [421] "이천칠"          "이천팔"            "인 듯하다"         "인젠"             "일"               "일것이다"          "일곱"             "일단"             "일때"             "일반적으로"        "일지라도"          "임에 틀림없다"
## [433] "입각하여"        "입장에서"          "잇따라"           "있다"             "자"               "자기"             "자기집"           "자마자"           "자신"             "잠깐"             "잠시"             "저"
## [445] "저것"            "저것만큼"          "저기"             "저쪽"             "저희"             "전부"             "전자"             "전후"             "점에서 보아"       "정도에 이르다"     "제"               "제각기"
## [457] "제외하고"        "조금"              "조차"             "조차도"           "졸졸"             "좀"               "좋아"             "좍좍"             "주룩주룩"          "주저하지 않고"     "줄은 몰랏다"       "줄은모른다"
## [469] "중에서"          "중의하나"          "즈음하여"          "즉"               "즉시"             "지든지"           "지만"             "지말고"           "진짜로"           "쪽으로"           "차라리"           "참"
## [481] "참나"            "첫번째로"          "쳇"               "총적으로"          "총적으로 말하면"   "총적으로 보면"     "칠"               "콸콸"             "쾅쾅"             "쿵"               "타다"             "타인"
## [493] "탕탕"            "토하다"            "통하여"           "툭"               "퉤"               "틈타"             "팍"               "팔"               "퍽"               "펄렁"             "하"               "하게될것이다"
## [505] "하게하다"        "하겠는가"          "하고 있다"         "하고있었다"        "하곤하였다"        "하구나"           "하기 때문에"       "하기 위하여"       "하기는한데"        "하기만 하면"       "하기보다는"        "하기에"
## [517] "하나"            "하느니"            "하는 김에"         "하는 편이 낫다"    "하는것도"          "하는것만 못하다"   "하는것이 낫다"     "하는바"           "하더라도"          "하도다"           "하도록시키다"      "하도록하다"
## [529] "하든지"          "하려고하다"        "하마터면"          "하면 할수록"       "하면된다"          "하면서"           "하물며"           "하여금"           "하여야"           "하자마자"          "하지 않는다면"     "하지 않도록"
## [541] "하지마"          "하지마라"          "하지만"           "하하"             "한 까닭에"         "한 이유는"         "한 후"            "한다면"           "한다면 몰라도"     "한데"             "한마디"           "한적이있다"
## [553] "한켠으로는"      "한항목"            "할 따름이다"       "할 생각이다"       "할 줄 안다"        "할 지경이다"       "할 힘이 있다"      "할때"             "할만하다"          "할망정"           "할뿐"             "할수있다"
## [565] "할수있어"        "할줄알다"          "할지라도"          "할지언정"          "함께"             "해도된다"          "해도좋다"          "해봐요"           "해서는 안된다"     "해야한다"          "해요"             "했어요"
## [577] "향하다"          "향하여"            "향해서"           "허"               "허걱"             "허허"             "헉"               "헉헉"             "헐떡헐떡"          "형식으로 쓰여"     "혹시"             "혹은"
## [589] "혼자"            "훨씬"              "휘익"             "휴"               "흐흐"             "흥"               "힘입어"
```

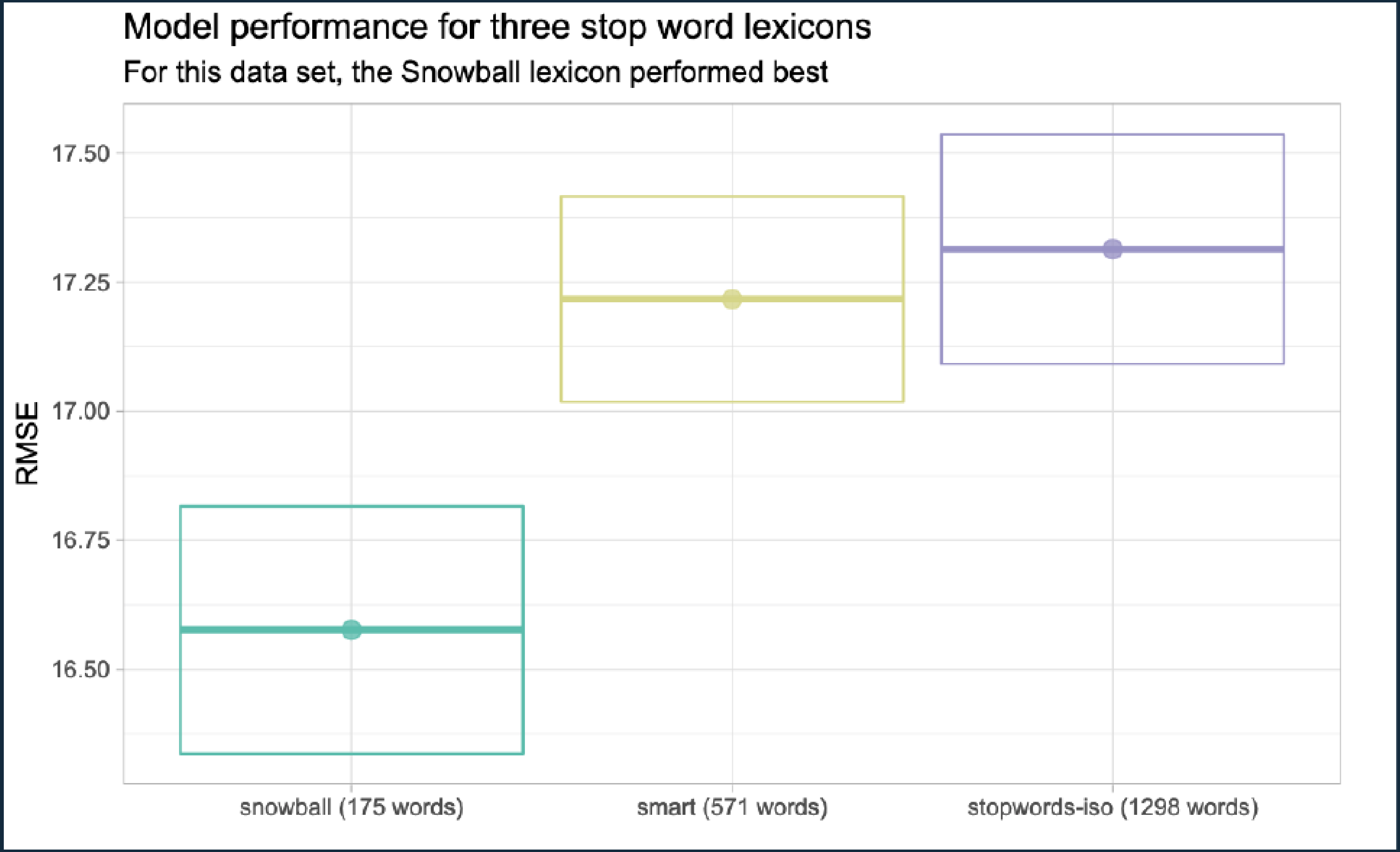
```
##   [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"
##   [8] "ourselves"  "you"        "your"       "yours"      "yourself"   "yourselves" "he"
##  [15] "him"        "his"        "himself"    "she"        "her"        "hers"       "herself"
##  [22] "it"         "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"
```

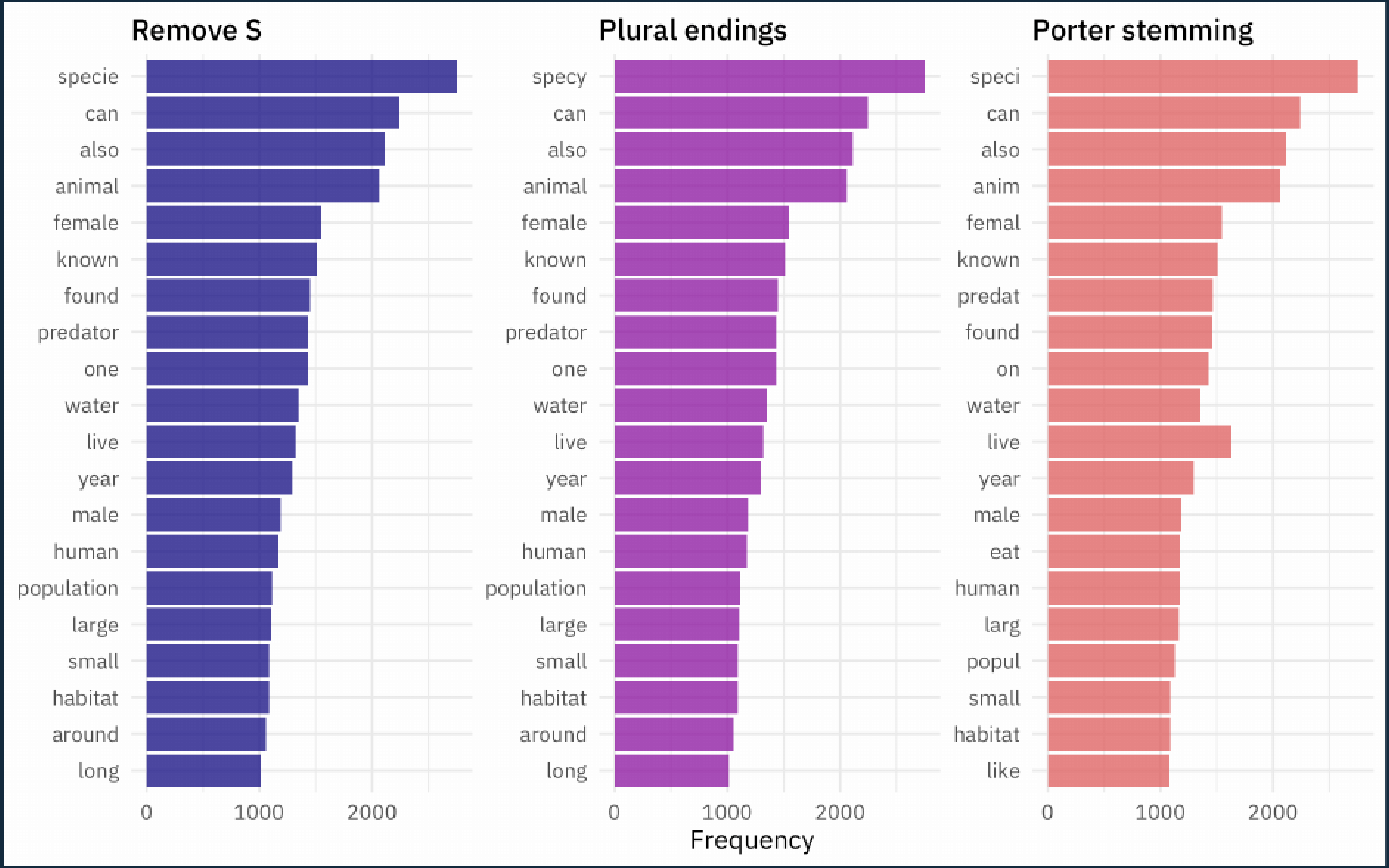
```r
recipe(diet ~ text, data = animal_train) %>%
    step_tokenize(text) %>%
    step_stopwords(text)
```

```
## 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
@juliasilge



Remove S
specie
can
also
animal
female
known
found
predator
one
water
live
year
male
human
population
large
small
habitat
around
long
0
1000
2000
Plural endings
specy
can
also
animal
female
known
found
predator
one
water
live
year
male
human
population
large
small
habitat
around
long
0
1000
2000
Frequency
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

@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.
```

```
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

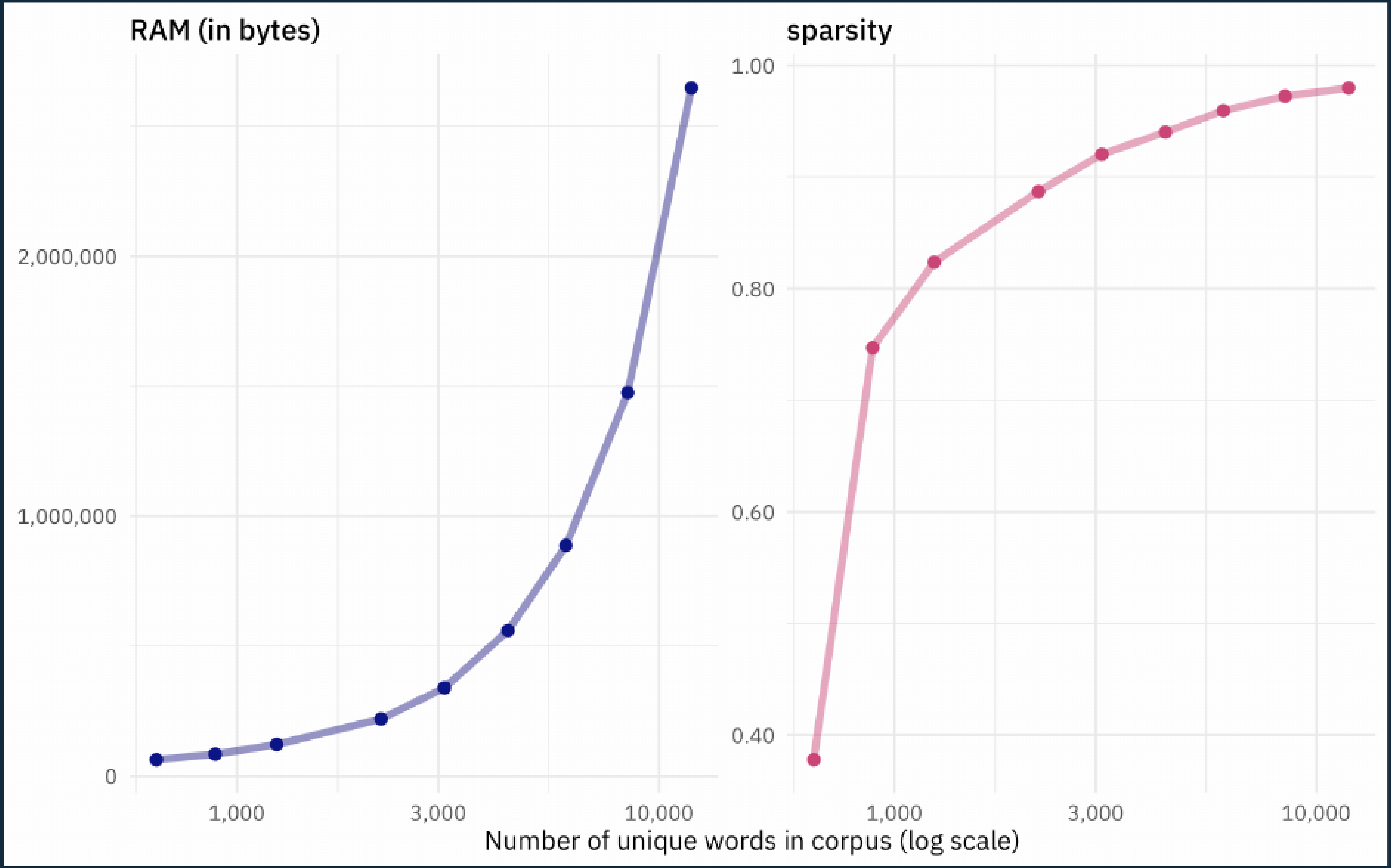
true positive rate

precision

true negative rate

```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