TEXT MINING

EXPLORATORY DATA ANALYSIS TO MACHINE LEARNING

TIDYTEXT

HELLO

I’m Julia Silge

Data Scientist at Stack Overflow

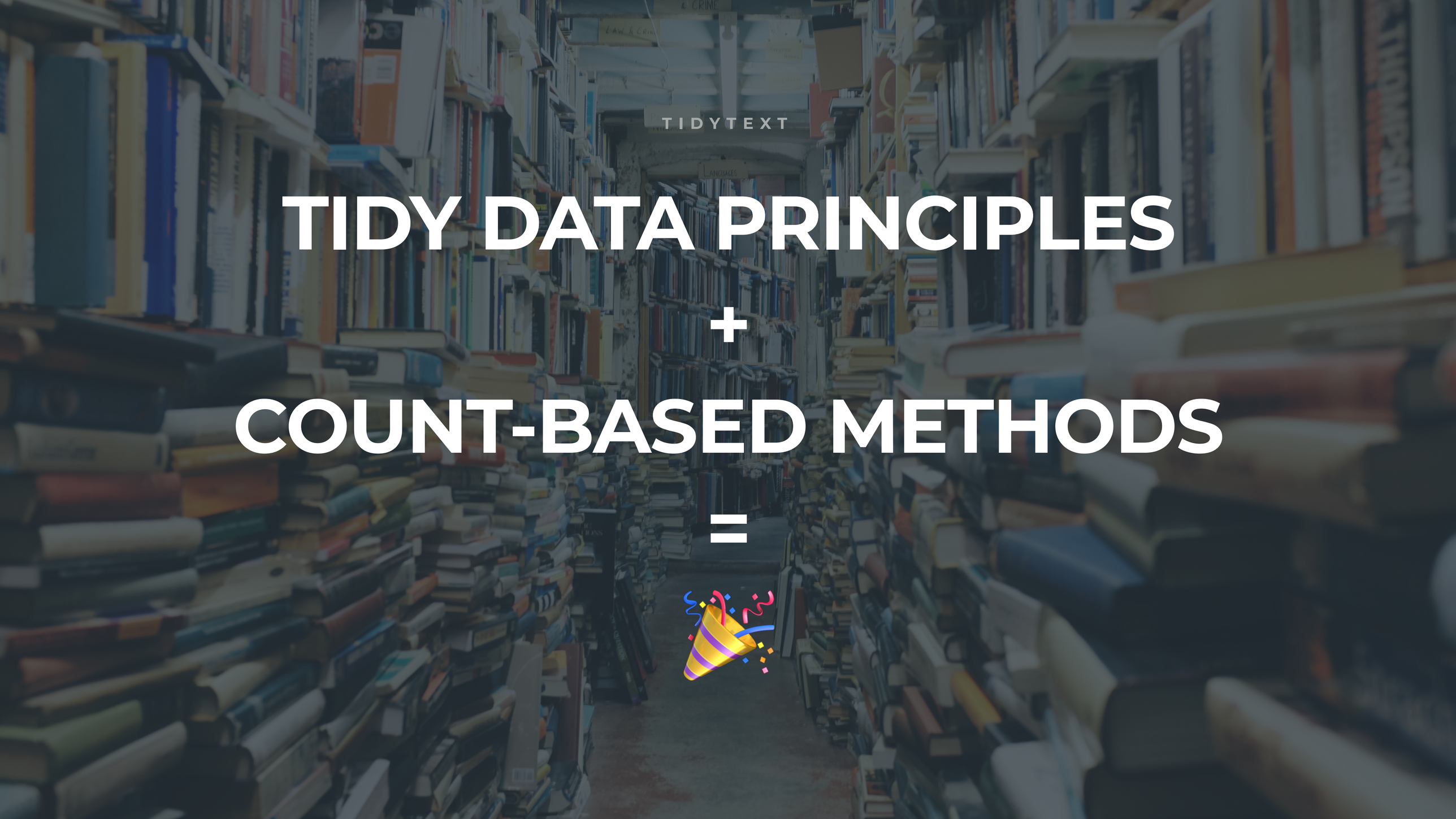
@juliasilge

https://juliasilge.com/

TEXT DATA IS INCREASINGLY IMPORTANT

TEXT DATA IS INCREASINGLY IMPORTANT

NLP TRAINING IS SCARCE ON THE GROUND



TIDY DATA PRINCIPLES

+

COUNT-BASED METHODS

=

🎉

tidytext: Text mining using dplyr, ggplot2, and other tidy tools

Authors: Julia Silge, David Robinson

Using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use. Much of the infrastructure needed for text mining with tidy data frames already exists in packages like dplyr, broom, tidyr and ggplot2. In this package, we provide functions and supporting data sets to allow conversion of text to and from tidy formats, and to switch seamlessly between tidy tools and existing text mining packages. Check out our book to learn more about text mining using tidy data principles.

https://github.com/juliasilge/tidytext

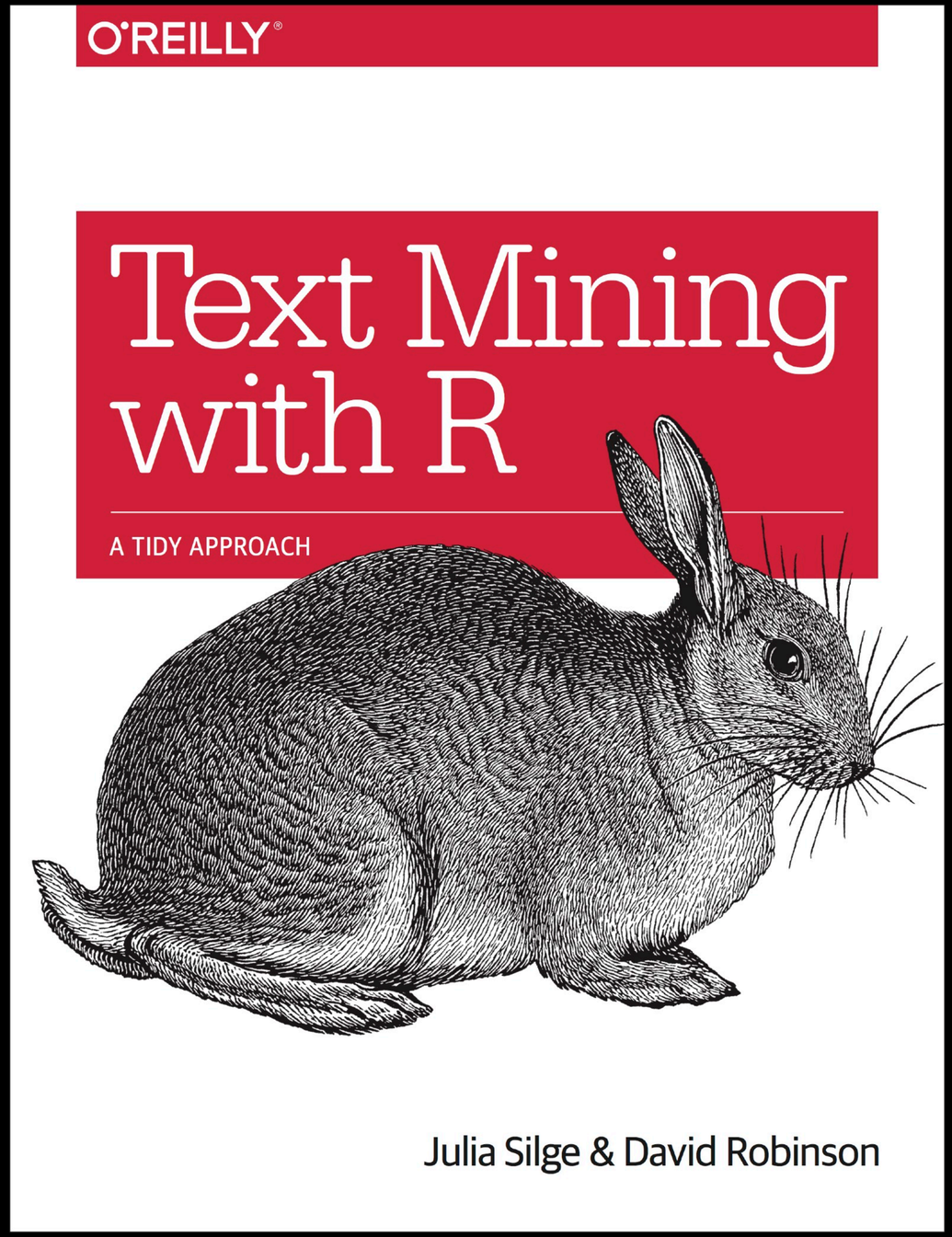
tidytext: Text mining using dplyr, ggplot2, and other tidy tools

Authors: Julia Silge, David Robinson

build passing | build passing | CRAN 0.2.2 | coverage 81% | DOI 10.5281/zenodo.3355453
JOSS 10.21105/joss.00037 | downloads 27K/month | downloads 683K

Using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use. Much of the infrastructure needed for text mining with tidy data frames already exists in packages like dplyr, broom, tidyr and ggplot2. In this package, we provide functions and supporting data sets to allow conversion of text to and from tidy formats, and to switch seamlessly between tidy tools and existing text mining packages. Check out our book to learn more ab mining using tidy data principles.

https://github.com/juliasilge/tidytext



https://tidytextmining.com/

EXPLORATORY DATA ANALYSIS

N-GRAMS AND MORE WORDS

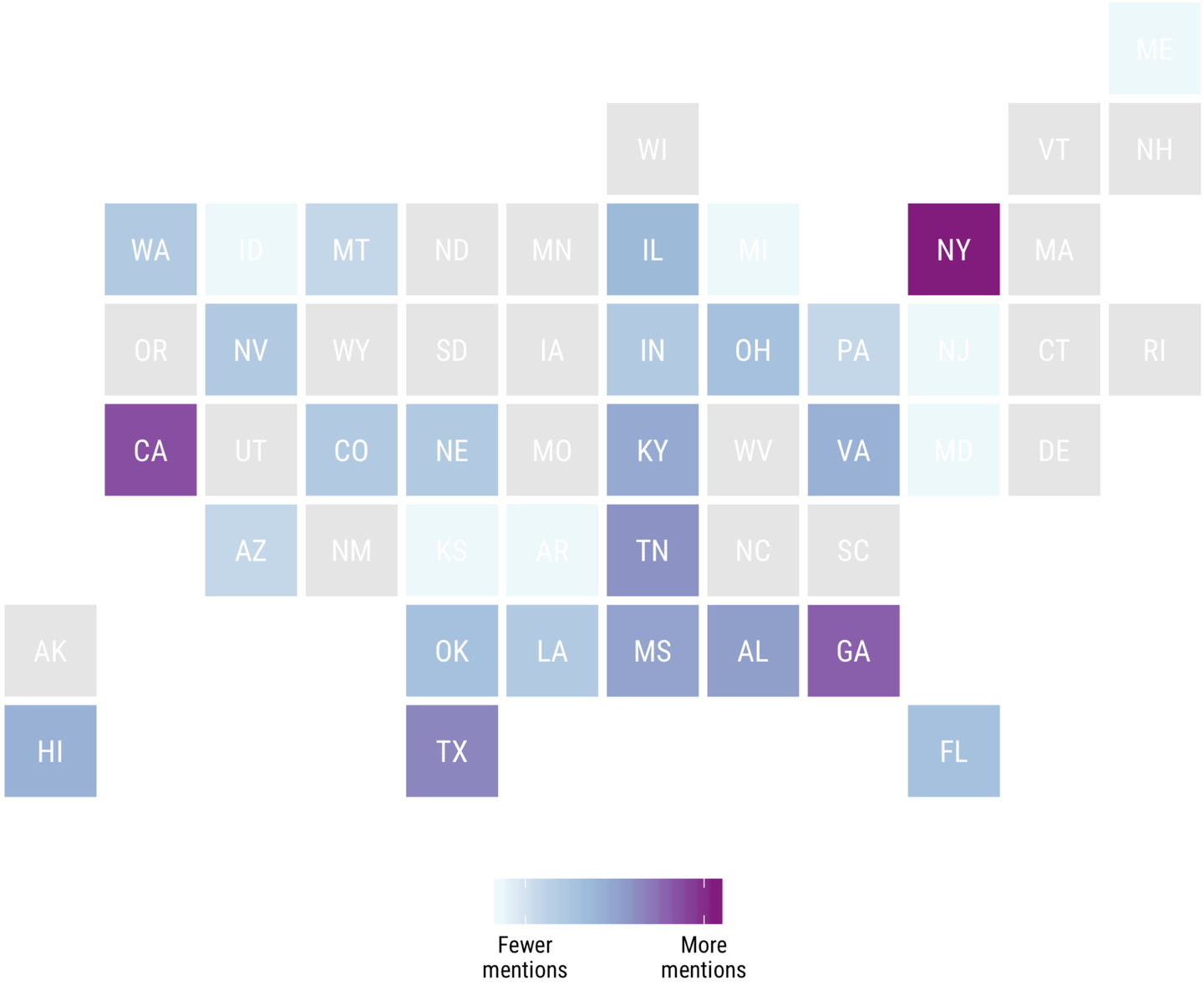
MACHINE LEARNING

TIDYTEXT

EXPLORATORY DATA ANALYSIS

Which States Are Mentioned Most in Song Lyrics?

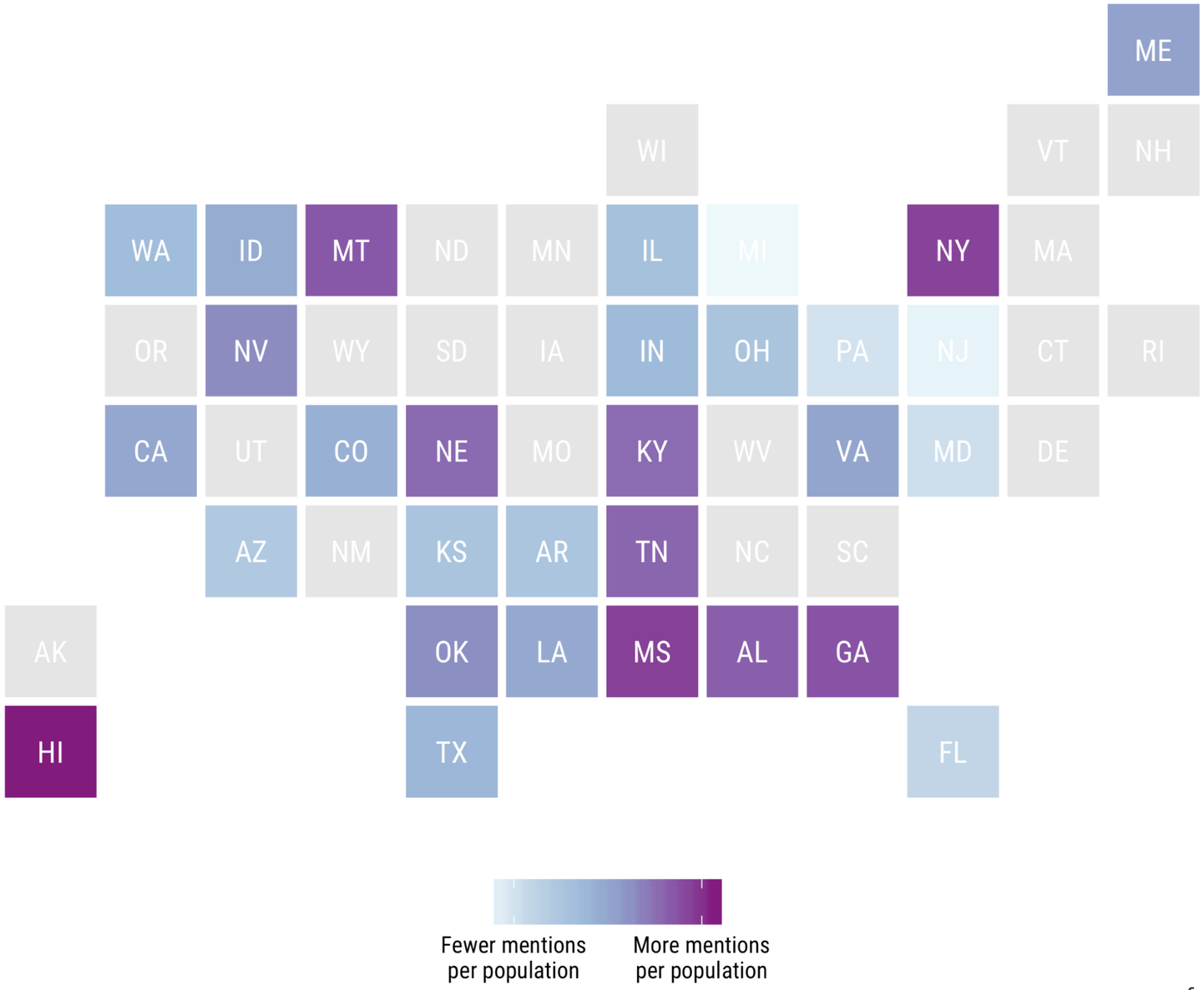
States like California are mentioned most often



from the Washington Post's Wonkblog

Which States Are Mentioned Most in Song Lyrics?

States like Hawaii and Montana are mentioned more often relative to their population

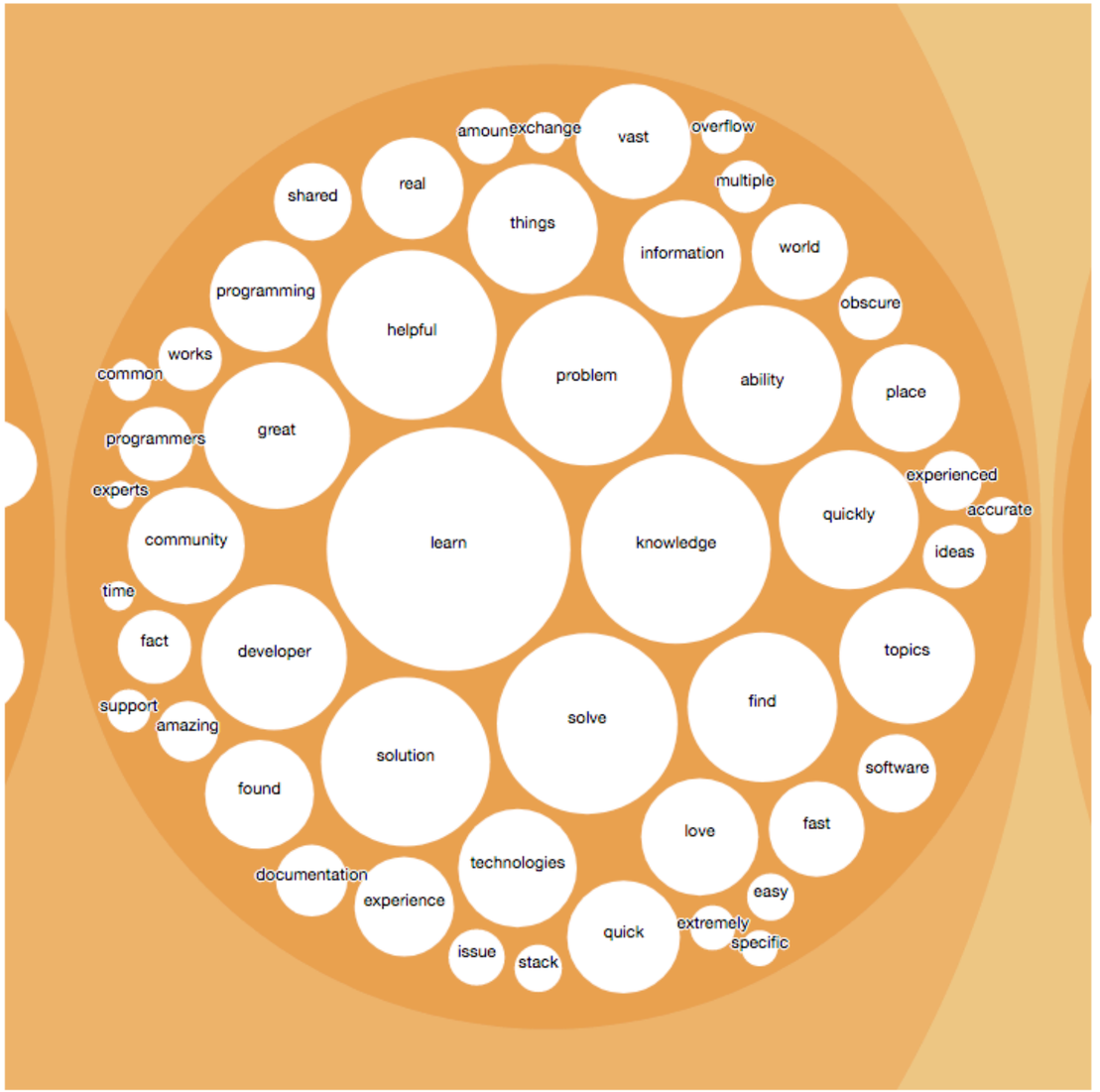


from the Washington Post's Wonkblog

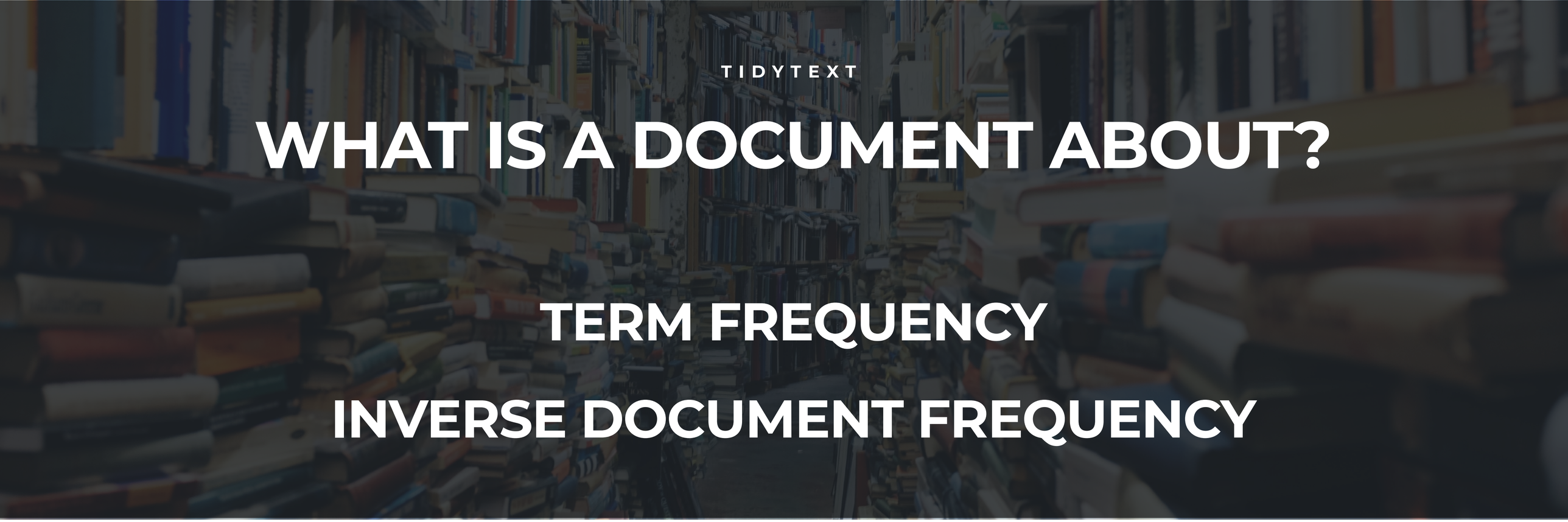
What do users say about Stack Overflow?

On the 2018 Stack Overflow Developer Survey, we asked users what the best, worst, most exciting, and most annoying things about Stack Overflow are.

Here are their answers.



D3 visualization on Glitch



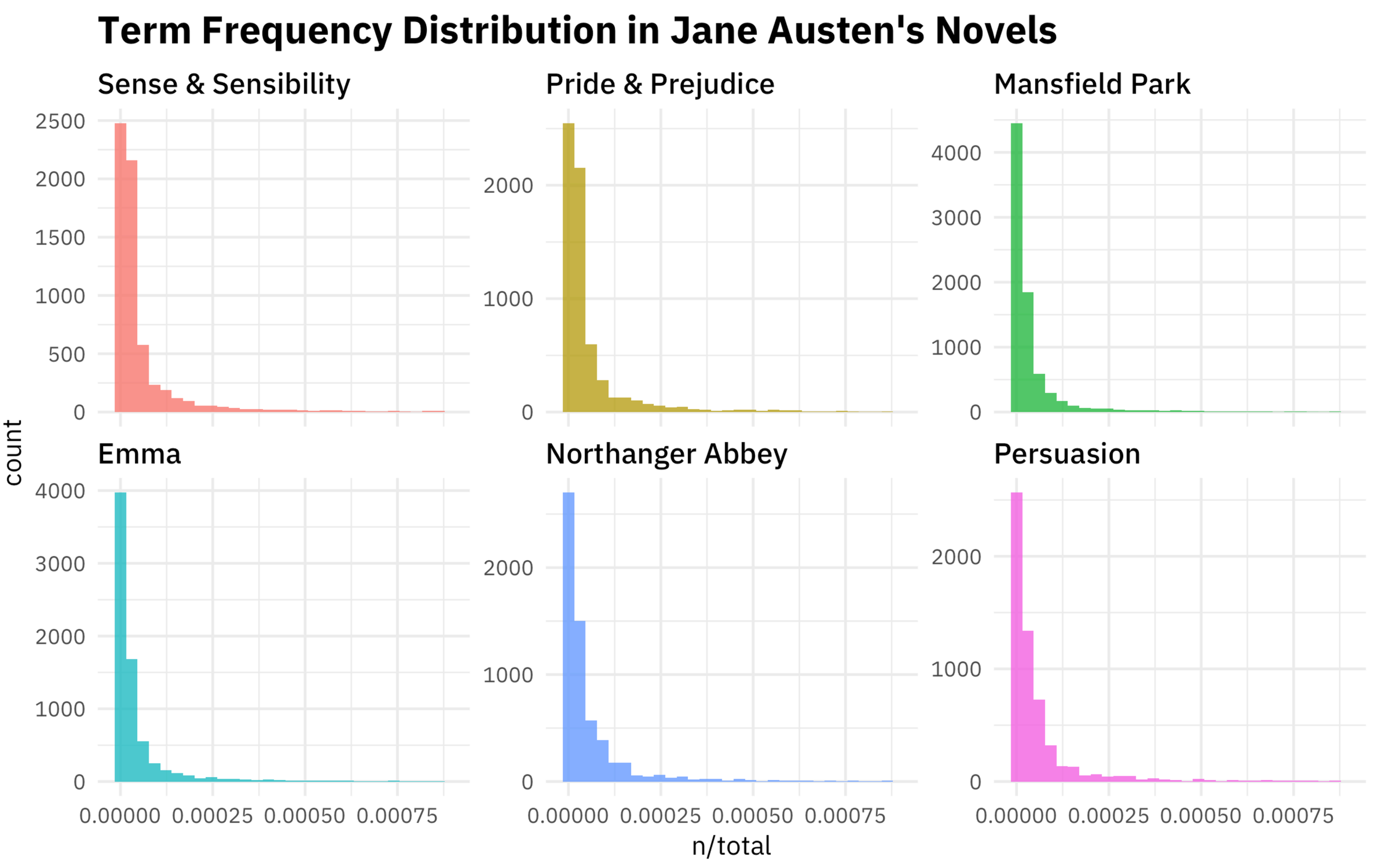
TIDYTEXT

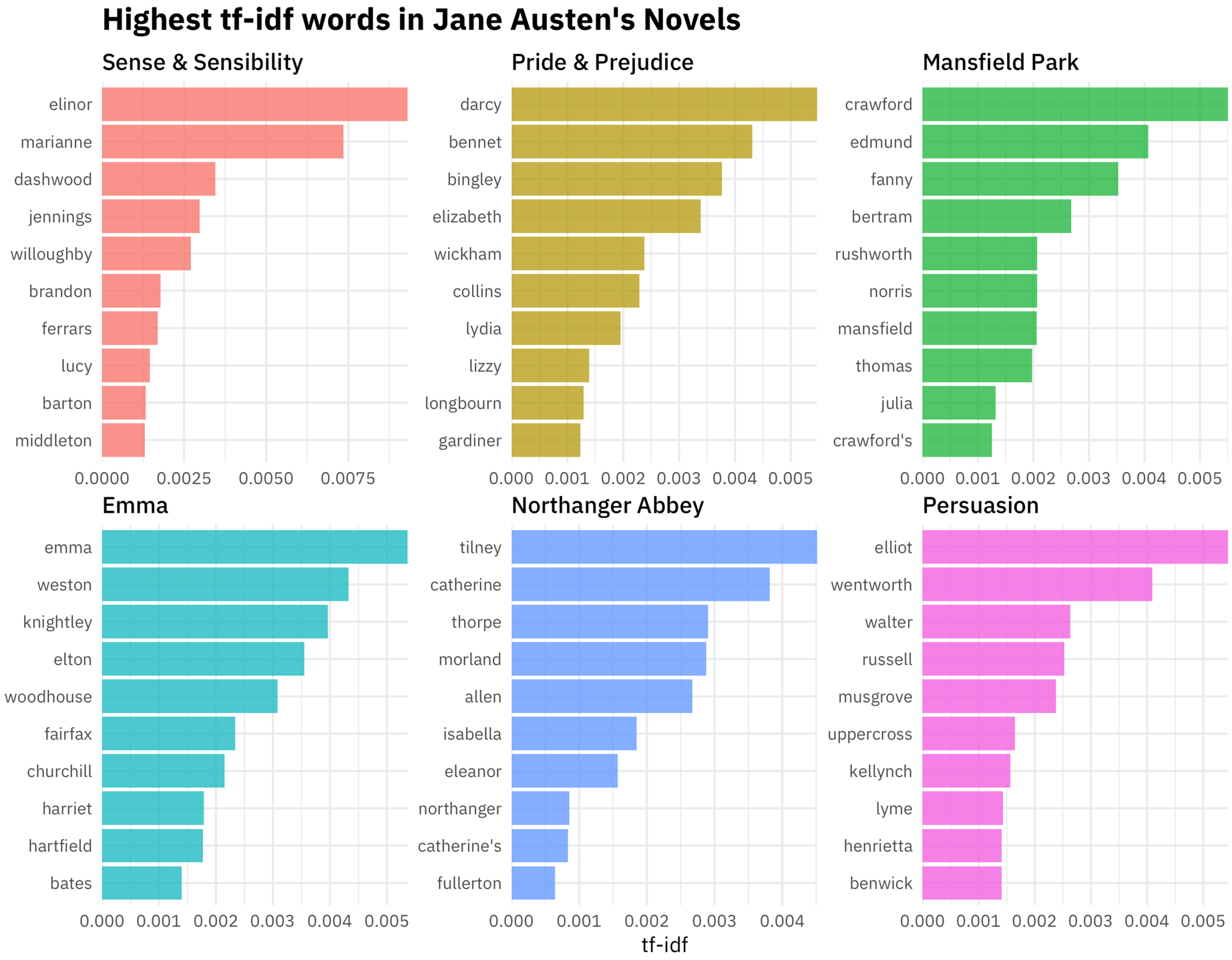
WHAT IS A DOCUMENT ABOUT?

TERM FREQUENCY

INVERSE DOCUMENT FREQUENCY

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

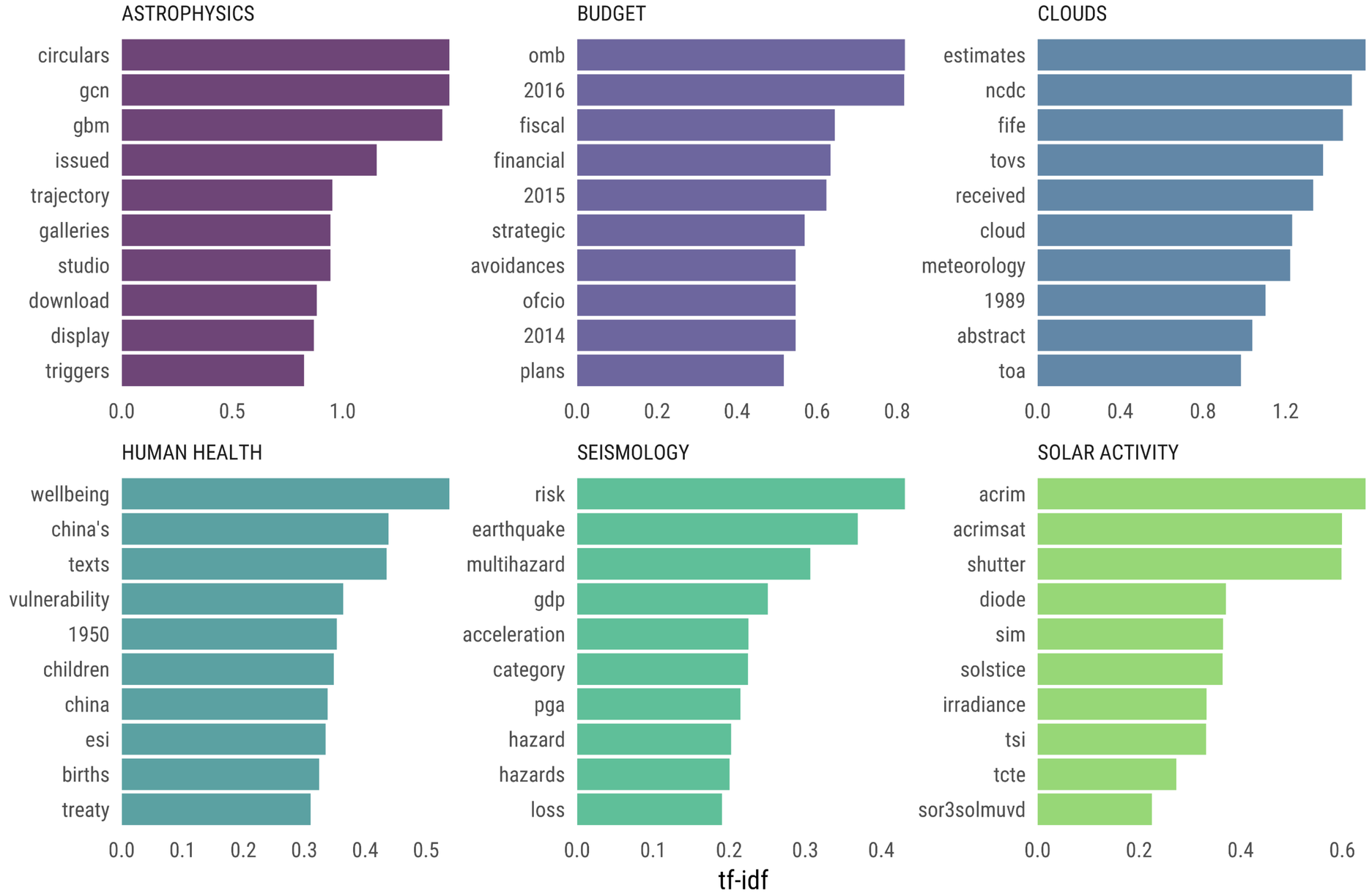

Term Frequency Distribution in Jane Austen's Novels
Sense & Sensibility
Pride & Prejudice
Mansfield Park
Emma
Northanger Abbey
Persuasion
count
n/total
0
500
1000
1500
2000
2500
3000
4000
0.00000
0.00025
0.00050
0.00075


Highest tf-idf words in Jane Austen's Novels
Sense & Sensibility
elinor
marianne
dashwood
jennings
willoughby
brandon
ferrars
lucy
barton
middleton
0.0000
0.0025
0.0050
0.0075
Pride & Prejudice
darcy
bennet
bingley
elizabeth
wickham
collins
lydia
lizzy
longbourn
gardiner
0.000
0.001
0.002
0.003
0.004
0.005
Mansfield Park
crawford
edmund
fanny
bertram
rushworth
norris
mansfield
thomas
julia
crawford's
0.000
0.001
0.002
0.003
0.004
0.005
Emma
emma
weston
knightley
elton
woodhouse
fairfax
churchill
harriet
hartfield
bates
0.000
0.001
0.002
0.003
0.004
0.005
Northanger Abbey
tilney
catherine
thorpe
morland
allen
isabella
eleanor
northanger
catherine's
fullerton
0.000
0.001
0.002
0.003
0.004
Persuasion
elliot
wentworth
walter
russell
musgrove
uppercross
kellynch
lyme
henrietta
benwick
0.000
0.001
0.002
0.003
0.004
0.005
tf-idf

- As part of the NASA Datanauts program, I worked on a project to understand NASA datasets
- Metadata includes title, description, keywords, etc

Highest tf-idf words in NASA Metadata Description Fields

Distribution of tf-idf for words from datasets labeled with select keywords

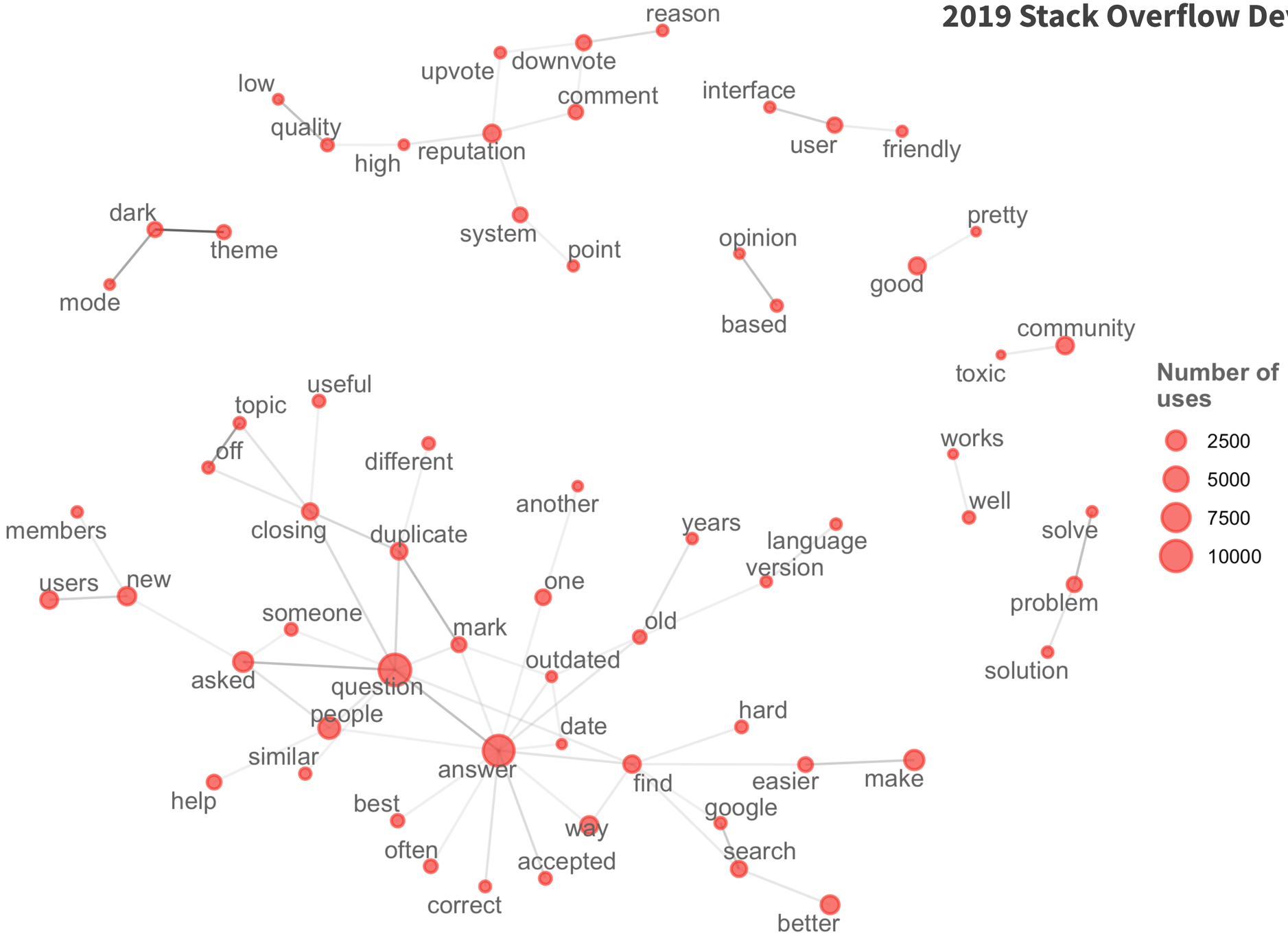


NASA metadata from https://data.nasa.gov/data.json

TAKING TIDY TEXT TO THE NEXT LEVEL

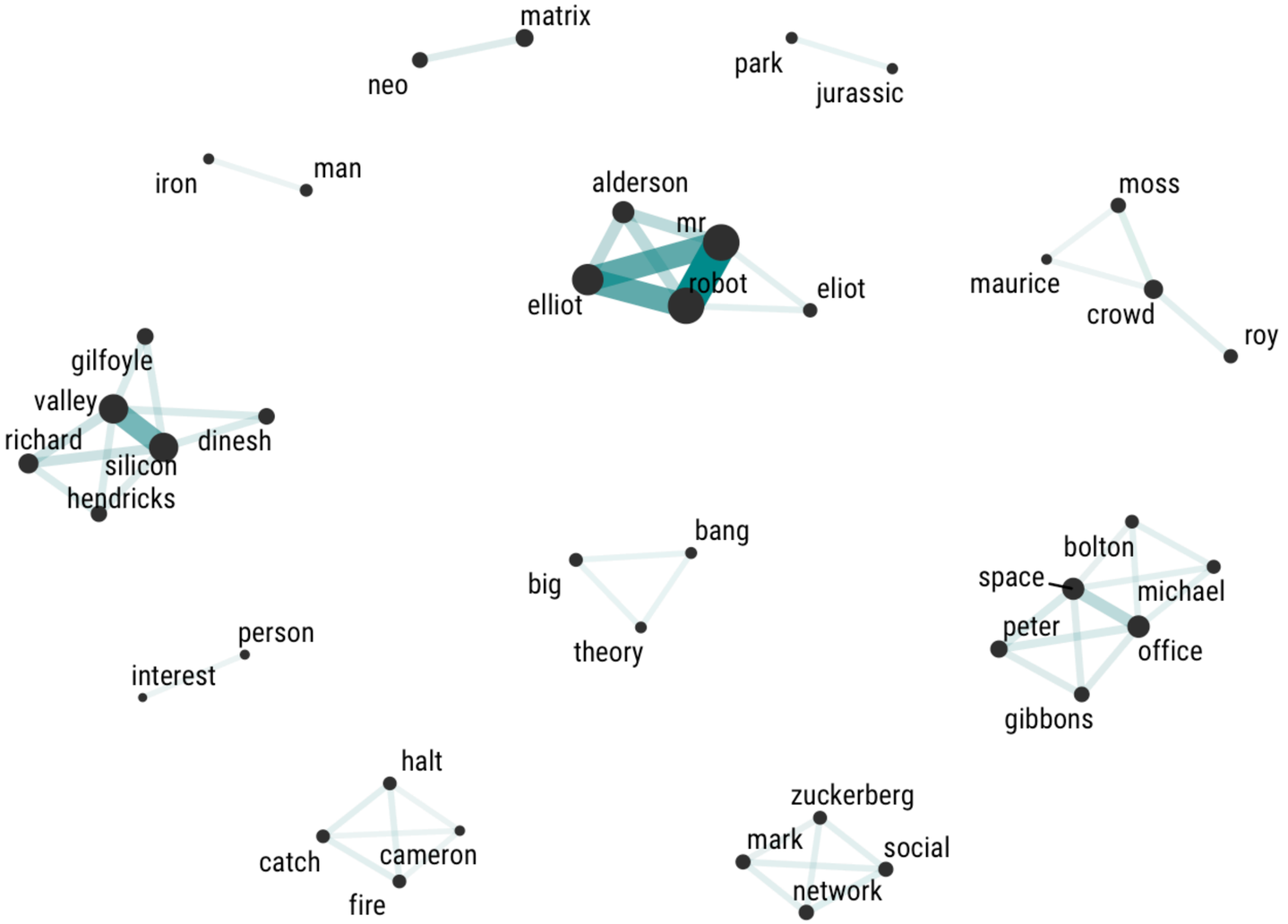
N-GRAMS, NETWORKS, & NEGATION

2019 Stack Overflow Developer Survey

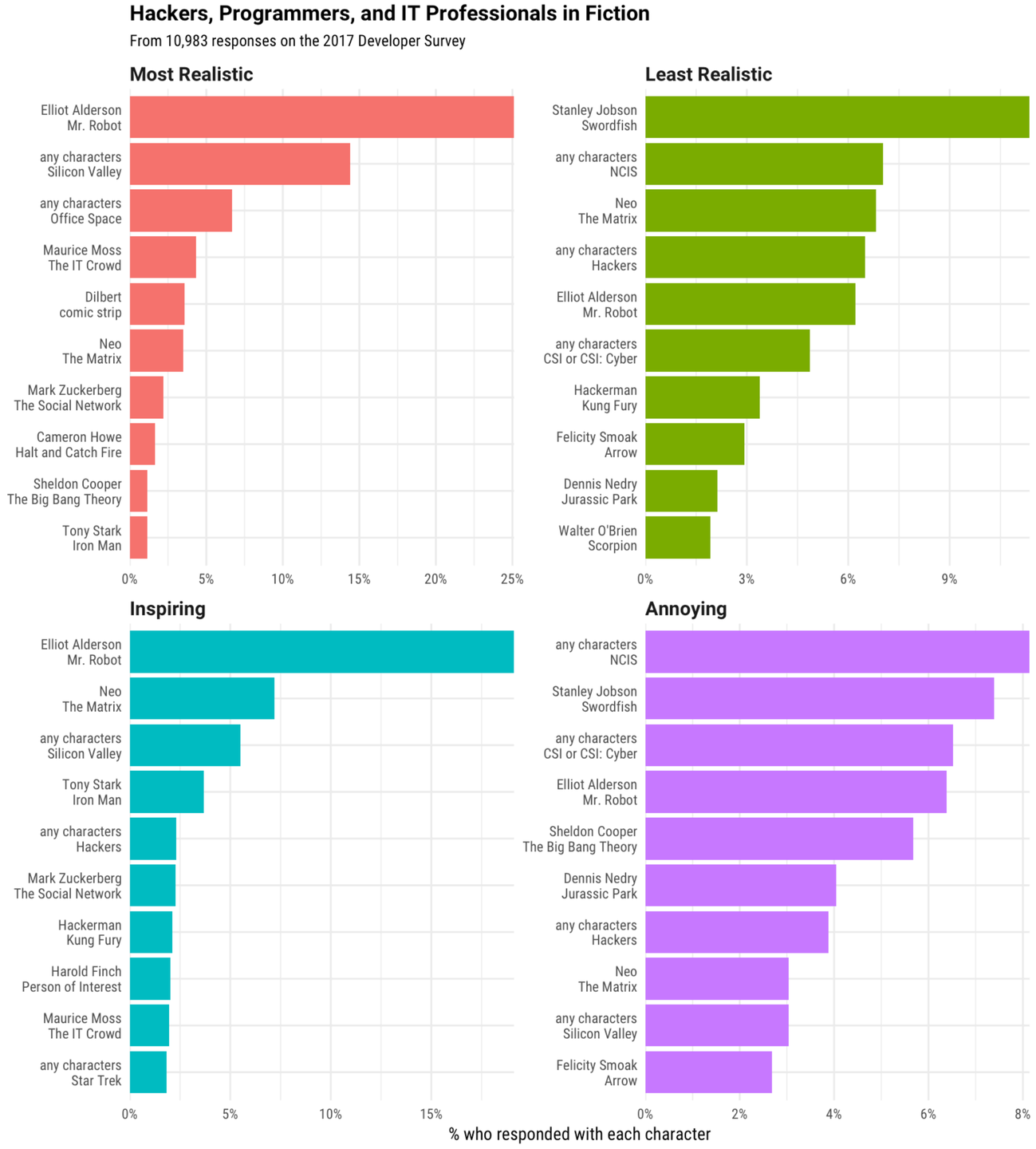


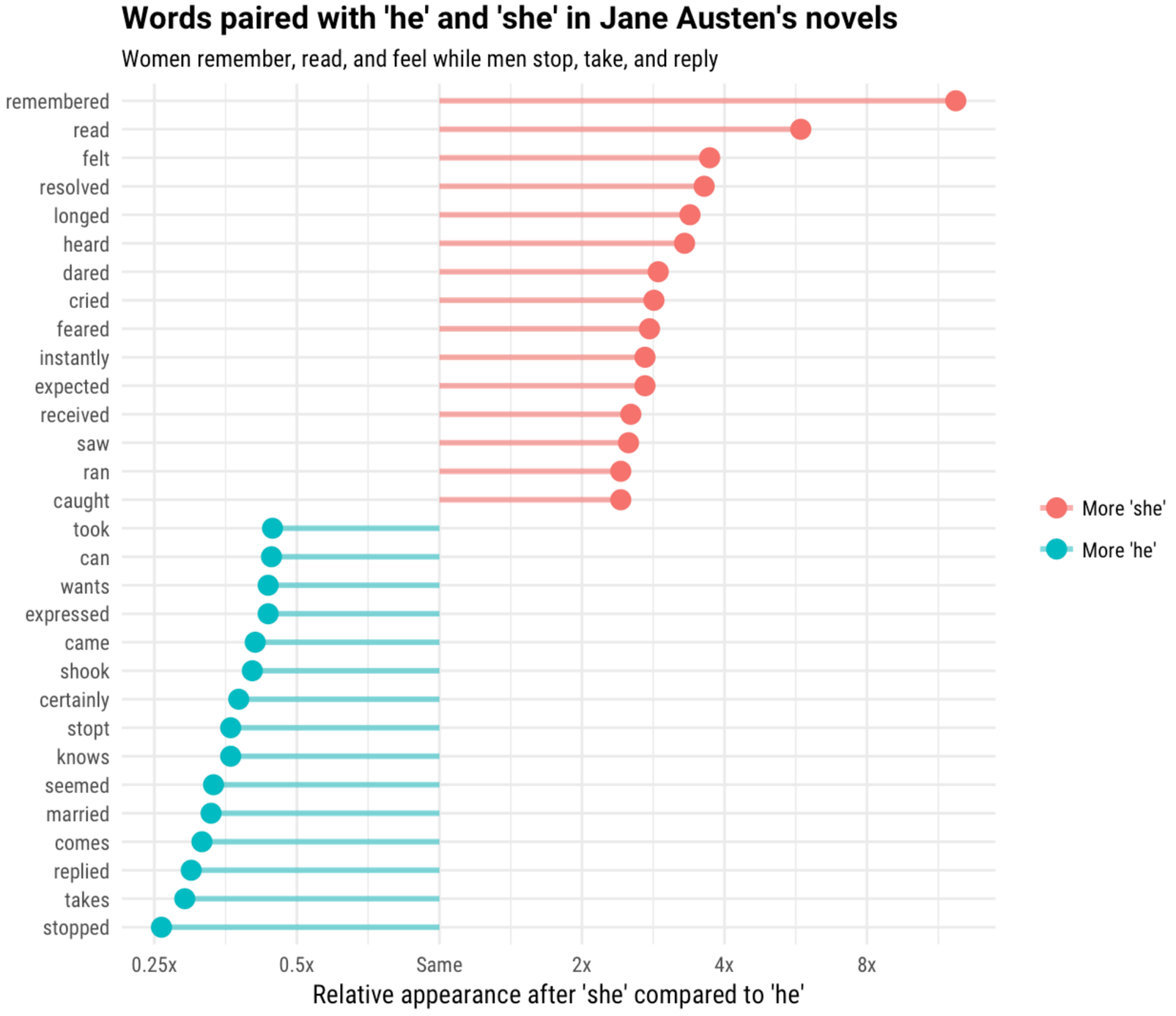
Most Realistic Fictional Characters

As identified by respondents in the 2017 Stack Overflow Survey



Words with larger points are used more often, and heavier connections indicate words are used more often together


Hackers, Programmers, and IT Professionals in Fiction
From 10,983 responses on the 2017 Developer Survey
Most Realistic
Elliot Alderson Mr. Robot
any characters Silicon Valley
any characters Office Space
Maurice Moss The IT Crowd
Dilbert comic strip
Neo The Matrix
Mark Zuckerberg The Social Network
Cameron Howe Halt and Catch Fire
Sheldon Cooper The Big Bang Theory
Tony Stark Iron Man
0%
5%
10%
15%
20%
25%
Least Realistic
Stanley Jobson Swordfish
any characters NCIS
Neo The Matrix
any characters Hackers
Elliot Alderson Mr. Robot
any characters CSI or CSI: Cyber
Hackerman Kung Fury
Felicity Smoak Arrow
Dennis Nedry Jurassic Park
Walter O'Brien Scorpion
0%
3%
6%
9%
Inspiring
Elliot Alderson Mr. Robot
Neo The Matrix
any characters Silicon Valley
Tony Stark Iron Man
any characters Hackers
Mark Zuckerberg The Social Network
Hackerman Kung Fury
Harold Finch Person of Interest
Maurice Moss The IT Crowd
any characters Star Trek
0%
5%
10%
15%
Annoying
any characters NCIS
Stanley Jobson Swordfish
any characters CSI or CSI: Cyber
Elliot Alderson Mr. Robot
Sheldon Cooper The Big Bang Theory
Dennis Nedry Jurassic Park
any characters Hackers
Neo The Matrix
any characters Silicon Valley
Felicity Smoak Arrow
0%
2%
4%
6%
8%
% who responded with each character


Words paired with 'he' and 'she' in Jane Austen's novels
Women remember, read, and feel while men stop, take, and reply
remembered
read
felt
resolved
longed
heard
dared
cried
feared
instantly
expected
received
saw
ran
caught
took
can
wants
expressed
came
shook
certainly
stopt
knows
seemed
married
comes
replied
takes
stopped
0.25x
0.5x
Same
2x
4x
8x
Relative appearance after 'she' compared to 'he'
More 'she'
More 'he'

ThePudding

She Giggles, He Gallops

Analyzing gender tropes in film with screen direction from 2,000 scripts.

By Julia Silge

+

Russell Goldenberg Amber Thomas Hanah Anderson

The top 800 words paired with “she” or “he”

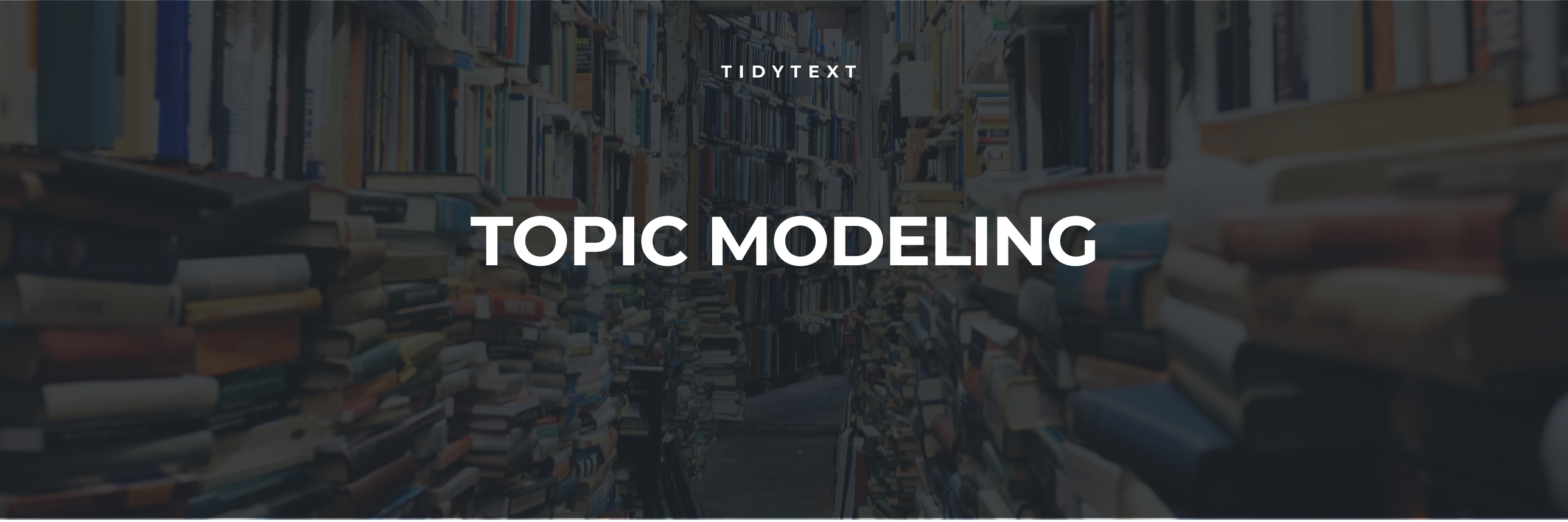
Underlined words contain examples of their usage in screen direction.

snuggles giggles squeals sobs weeps blushes clings rocks shrieks hugs shrinks gasps responds trembles pets flinches arches skips utters shudders startles buries swats murmurs resists
hovers caresses awakens shivers screams dances beats absently flees cleans stirs straddles cries moans bites realises mouths accepts wore smiles laughs wrote serves scoots liked
arranges scampers storms twirls softens ignores softly faints wonders fades sags hesitates casts applies hisses fiddles kisses sings awkwardly smokes stretches sips unbuttons stiffens
hurriedly hurries dries lool early refuses tiptoes lingers beams pivots curls glides strokes meant abruptly retrieves bursts
rakes relents reluctantly fr trails frowns retreats gonna licks touches reacts nearly sighs backs embraces squirms panics
yelps ends allows flashes rs shakes instinctively replies told freezes resumes creeps calms gives rushes sails tentatively
thrashes becomes types a s blinks dabs meets rests regards tilts attacks darts eyes brushes descends gently nervously
returns forces closes fixes s halts wakes bolts slaps fumbles quickly wears clasps faces feeds barely shrugs believes
floats left whacks blows i slices runs leans sounds washes swallows cranes observes accidentally marches rifles
squeezes begins kneels tu grimaces frees put calls glares tucks like plops scurries whispers tries remains actually
continues pinches tells lets yawns disappears heads looks chokes discovers plugs springs watches cautiously opens clutches studies dropped massages obeys suddenly loves scowls
crosses packs scribbles spits sneaks puts likes just lashes topples hangs lies angles starts claps adds flicks slowly cups angrily really stops pours jabs traces unzips crawls died
grasps slugs steadies breathes glances pushes directs inches hands reads comes unfolds winds attempts rubs snorts walks drifts sinks hides goes swivels feels keeps sprays sways
means ducks races repeats used steps sends finishes talking trips locks waits gets snatches chooses obviously takes decides plucks slides moves peers holds claws already stomps
swipes asks admires met said stands buttons owns says sets strips must wanted focuses fell unwraps edges unlocks remembers shines reaches jumps always places climbs stumbles
handles leafs needed passed surreptitiously concentrates helps interrupts reels scrambles steels blocks clenches floors gags splashes rips notices knocks enters finally rises listens
quietly jerks bumps wants lays kicks flies makes picks throws casually scans winces might dangles hefts flails waves dresses realizes passes catches plants thinks knows rings empties
hears sweeps may signs wrenches swims shares started recovers hops props tightens indicates hear dashes got peels will silently probably grows whips finds spins pulls switches
lights assumes ties digs hopes wheels lines performs settles tenses sniffs stabs wanders downs pounds notes slams twists shows weaves bends bounces curses expects hastily pokes
can sees acts scrolls brings arrives needs follows get zips manages proceeds deposits hardly strikes exhales still smells came tears yanks lifts knew shifts presses grabs excuses
straightens hustles speeds recognizes carries pays also falls stays tastes drags never speaks dials turned slows relaxes brought dumps sticks changes know come lowers flips bought
sleeps greets registers succeeds now even withdraws cracks writhes saw nudges scratches hobbles steals pauses collapses offers puffs grinds braces thought expected levels gave hits
drops spots lurches clambers ever eventually snaps writes searches gazes approaches selects eases dips cocks groans lives works winks swerves grips fills rummages loved joins
regains wiggles talks beckons gulps maneuvers zooms stalks seizes vanishes points thrusts hauls leaps hit adjusts heard shoots refers deals honks rams releases clears nears pries

She twirls, looking at herself in the mirror...
She twirls toward the door, grabbing her purse.
She twirls off. He chases her, beer and entries in hand.
She twirls a wooden katana as part of a floor routine.
She twirls a seductive finger around his tie. They kiss.
She twirls her finger around Milo's palm.
She twirls the string of her basketball shorts...

TAKING TIDY TEXT TO THE NEXT LEVEL

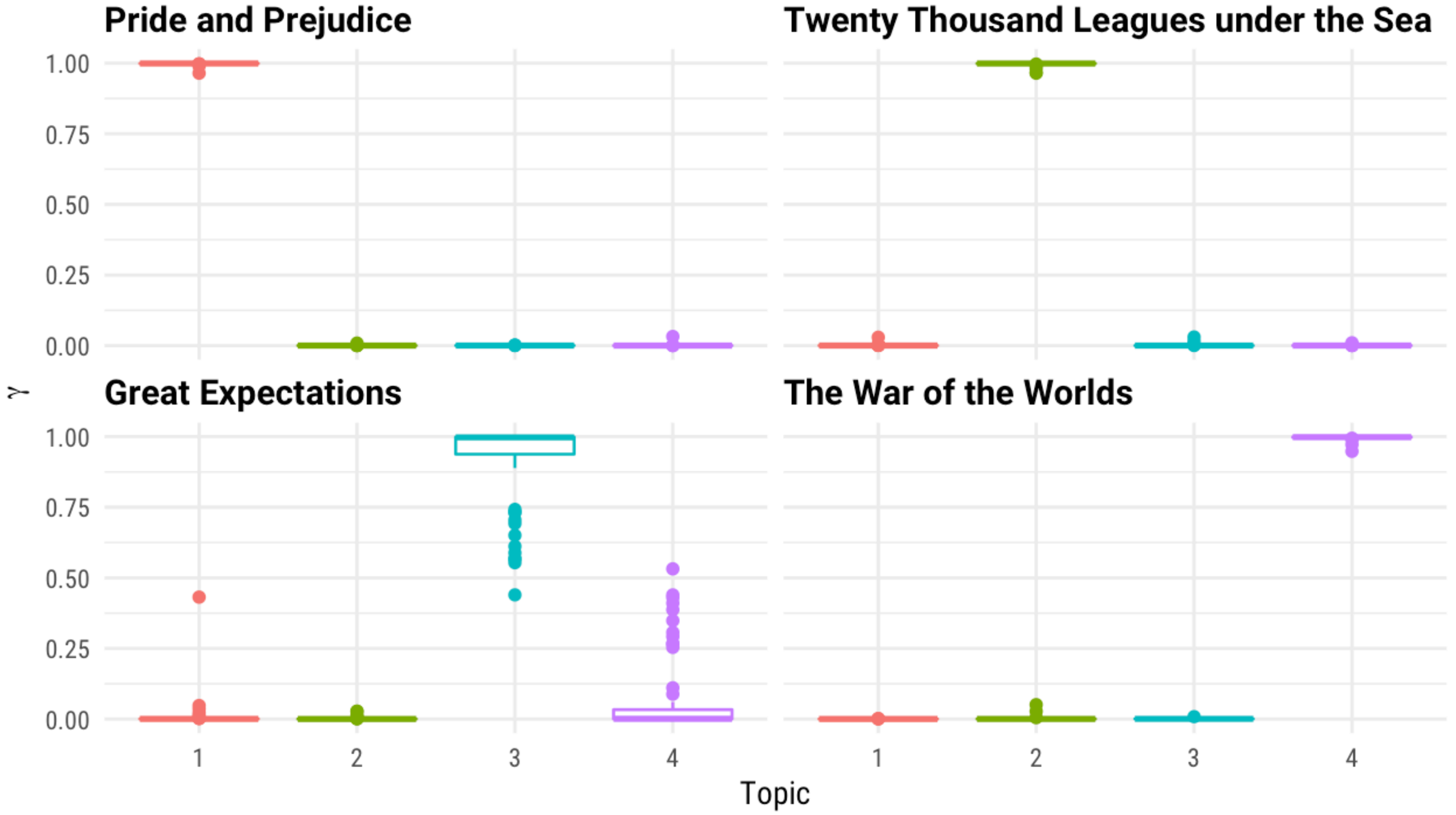
TOPIC MODELING



TIDYTEXT

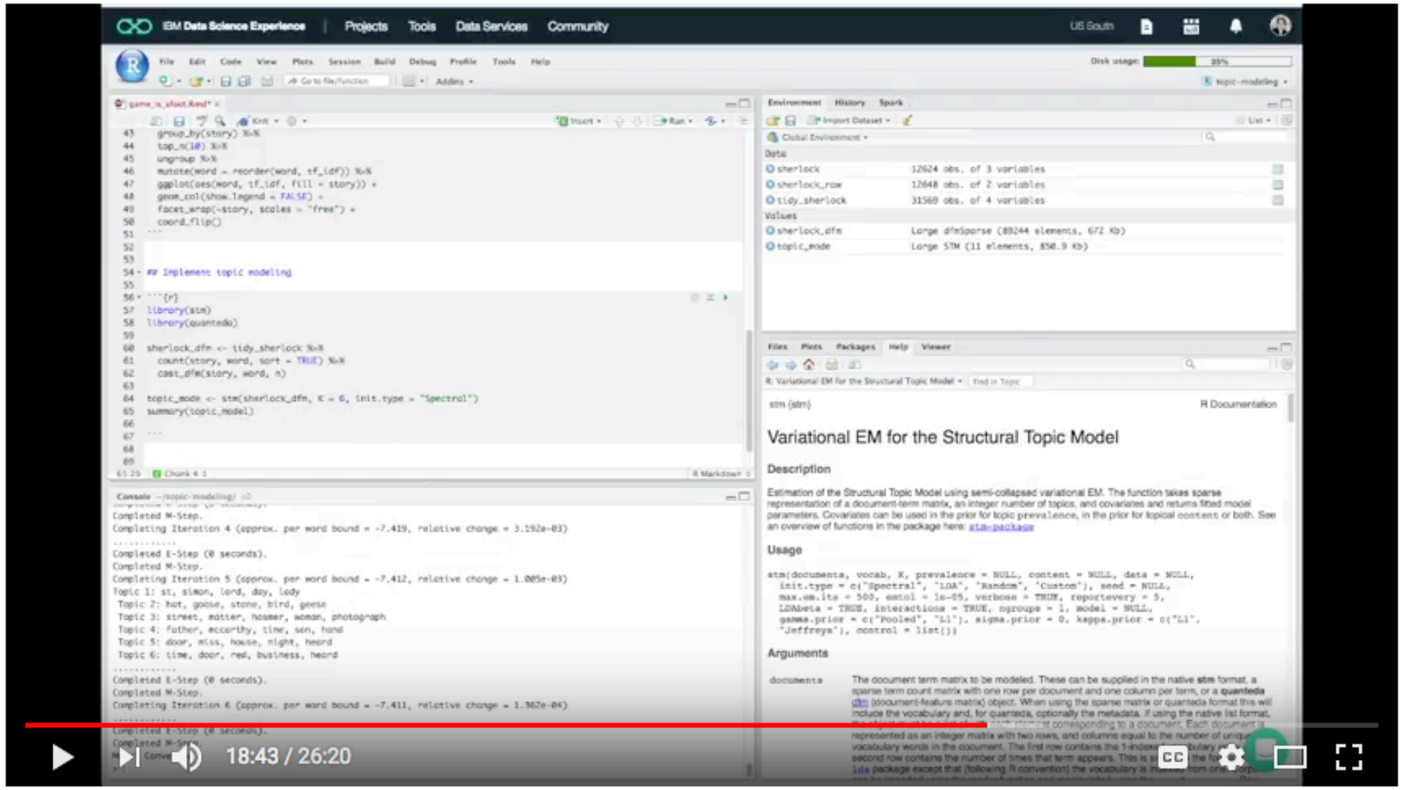
TOPIC MODELING

- Each DOCUMENT = mixture of topics
- Each TOPIC = mixture of words


Pride and Prejudice
Twenty Thousand Leagues under the Sea
Great Expectations
The War of the Worlds
1.00
0.75
0.50
0.25
0.00
γ
1
2
3
4
Topic

Top tags for each LDA topic

For questions with >80% probability for that topic



Topic modeling with R and tidy data principles

1,372 views

42 0 SHARE

Julia Silge
Published on Dec 18, 2017

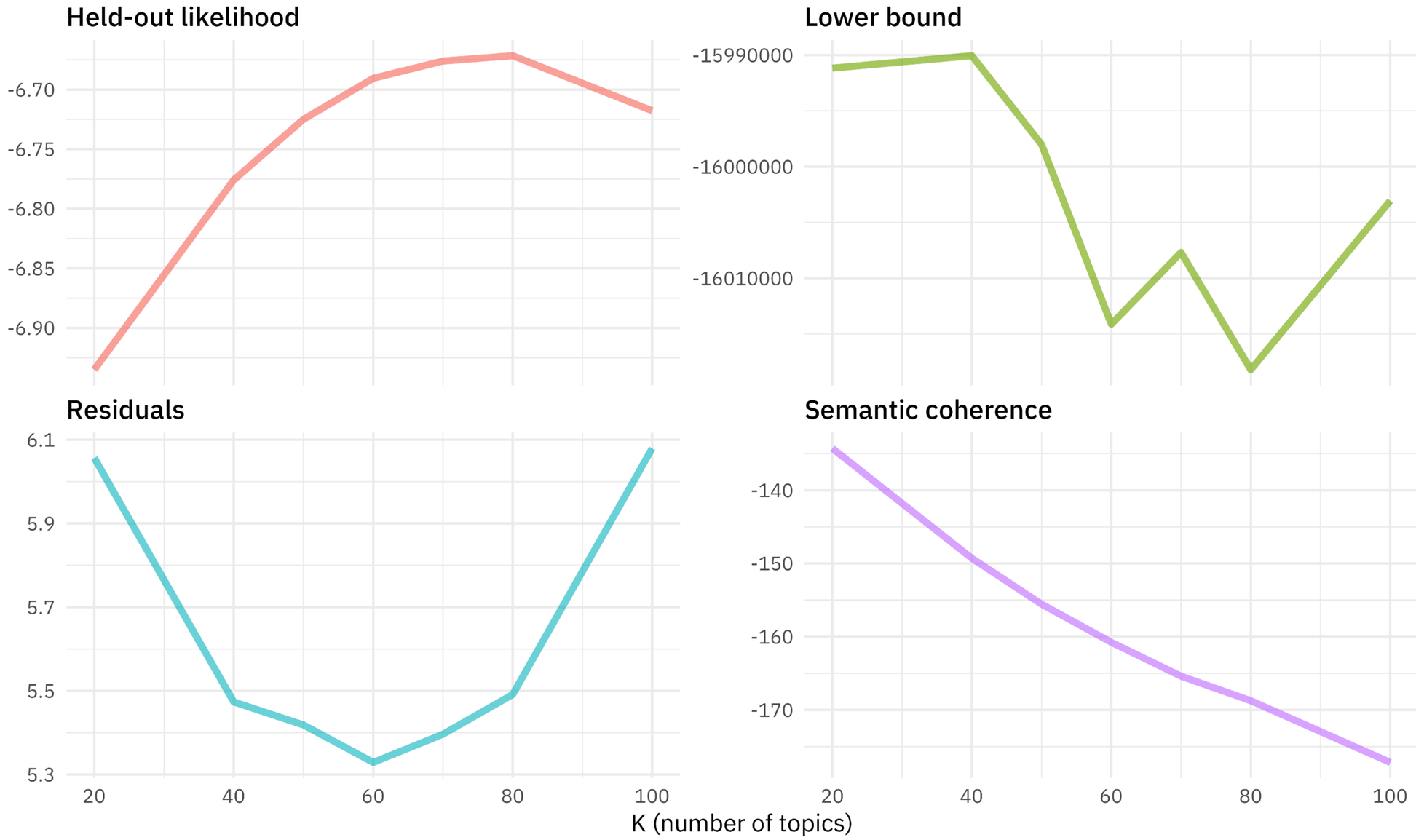
SUBSCRIBE 59

Watch along as I demonstrate how to train a topic model in R using the tidytext and stm packages on a collection of Sherlock Holmes stories. In this video, I'm working in IBM Cloud's Data Science Experience environment.

SHOW MORE

Model diagnostics by number of topics

These diagnostics indicate that a good number of topics would be around 60



TAKING TIDY TEXT TO THE NEXT LEVEL

TEXT CLASSIFICATION

TIDYTEXT

TRAIN A GLMNET MODEL

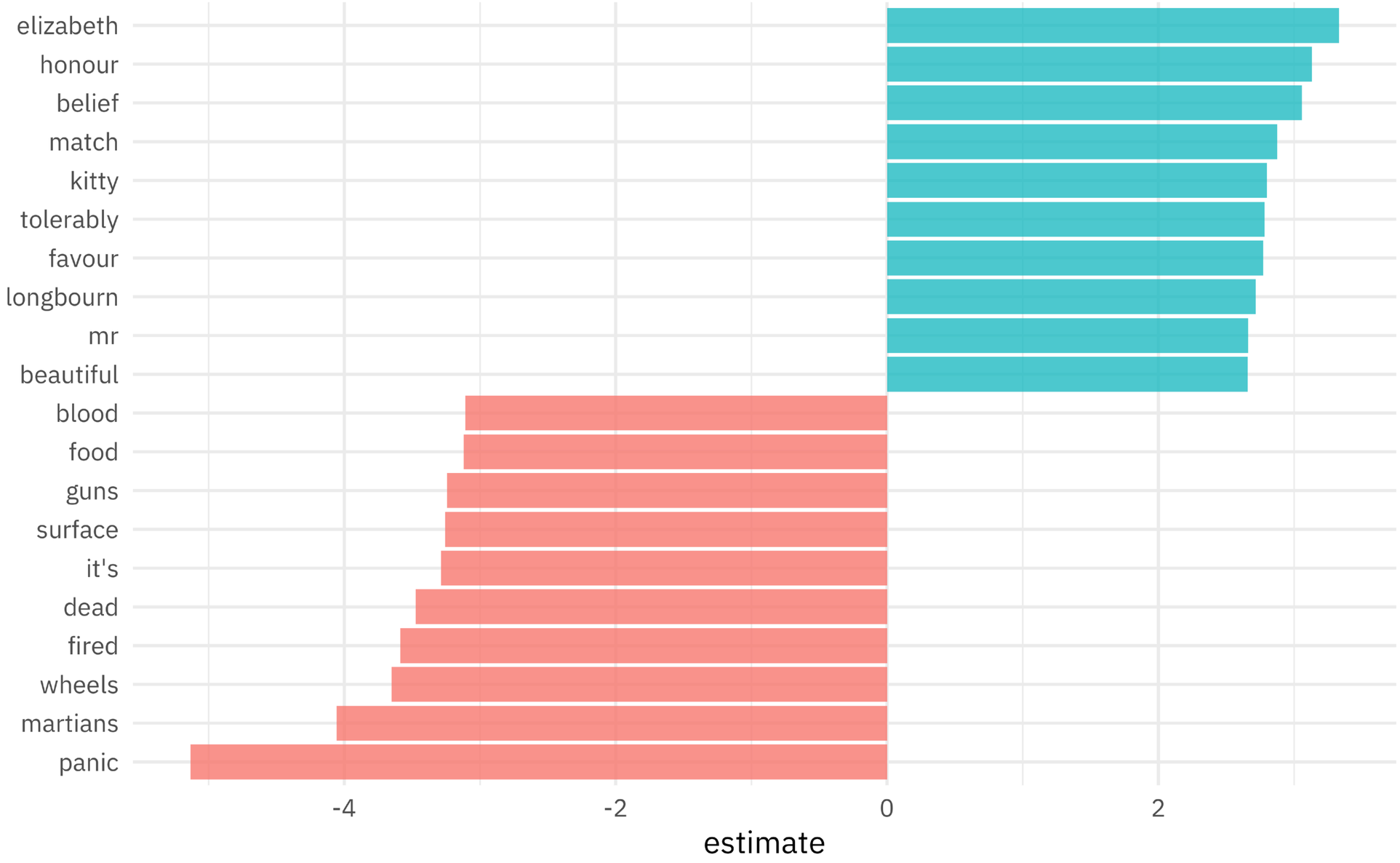
TIDYTEXT

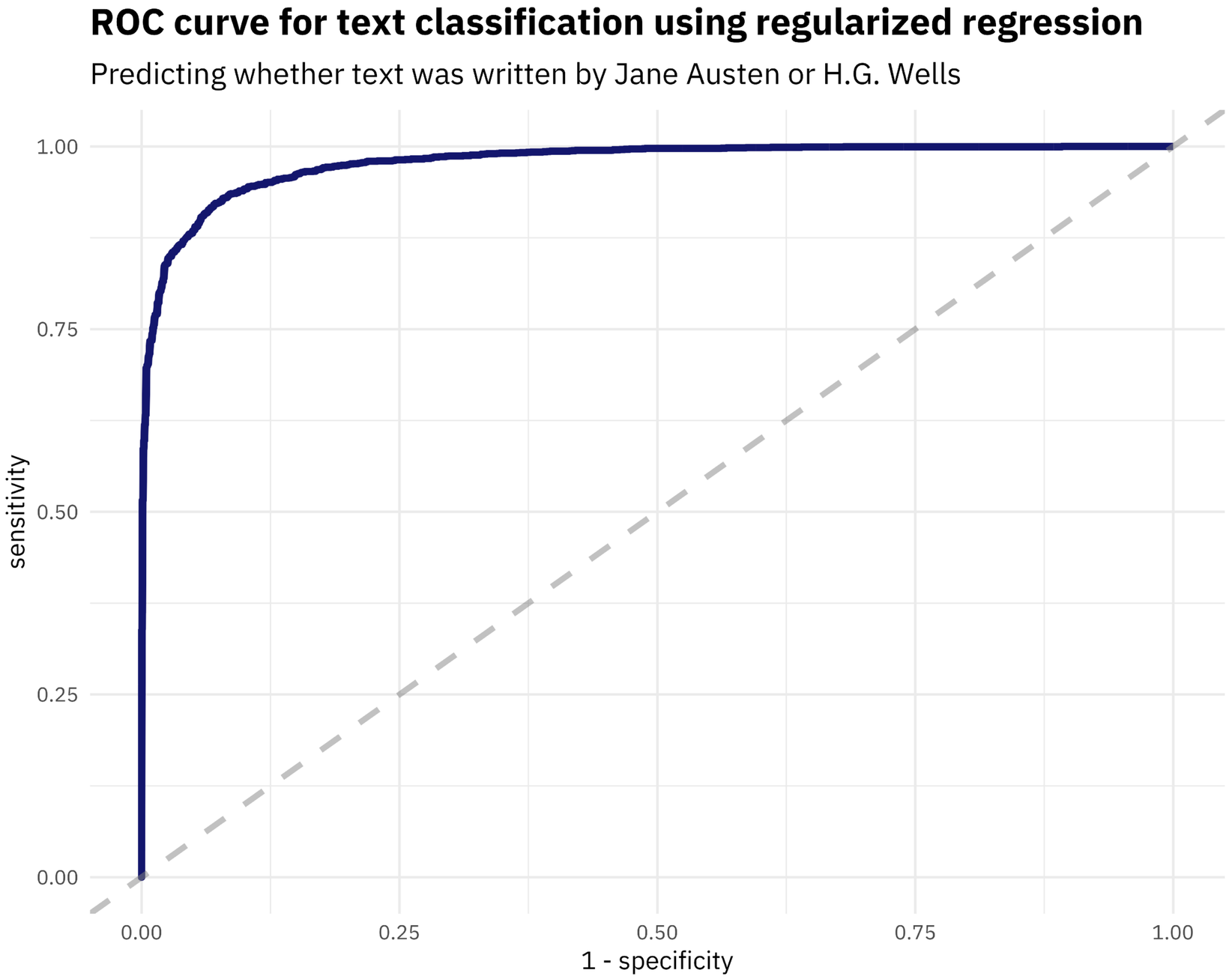
TEXT CLASSIFICATION

```
> library(glmnet)
> library(doMC)
> registerDoMC(cores = 8)
>
> is_jane <- books_joined$title == "Pride and Prejudice"
>
> model <- cv.glmnet(sparse_words, is_jane, family = "binomial",
+                    parallel = TRUE, keep = TRUE)
```

Coefficients that increase/decrease probability the most

A document mentioning Martians is unlikely to be written by Jane Austen




ROC curve for text classification using regularized regression
Predicting whether text was written by Jane Austen or H.G. Wells
1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity

Julia Silge

Data Scientist at Stack Overflow

Salt Lake City, UT, United States http://juliasilge.com/ juliasilge
juliasilge

I enjoy making beautiful charts, the statistical programming language R, black coffee, red wine, and the mountains of my adopted home here in Utah. I have a PhD in astrophysics and an abiding love for Jane Austen. My work involves analyzing and modeling complex data sets while communicating about technical topics with diverse audiences.

Favorite editor: RStudio

I want to work with

r ggplot2 dplyr data-visualization shiny text-mining knitr machine-learning

TOP 5% r rstudio

TOP 10% dplyr tidyr tidyverse

TOP 30% ggplot2

Position • Dec 2016 → Current (2 years, 10 months)

Data Scientist at Stack Overflow

r tidyverse ggplot2 shiny

- Analyze large datasets and build models to understand developers and

Read more

Open source • Apr 2016 → Current (3 years, 6 months)

tidytext: Text mining using dplyr, ggplot2, and other tidy tools
Last commit on Sep 01, 19

351 Commits / 383,231 ++ / 314,272 --

Using tidy data principles can make many text mining tasks easier, more effective, and

search jobs | search companies

Search all jobs | Located anywhere | Save Search

Remote | Tech | Compensation | Perks | Background | More | Debug

Experience Level

Select min. experience
Student
Junior
Mid-Level
Senior
Lead
Manager

to Select max. experien

Full-time
Contract
Internship

Apply filters Cancel

9,735 results

FEATURED

R Developer
YouGov - No office location
Remote

r shiny python etl

Developer)

Data Science Curriculum Writer
Thinkful Inc. - No office location
$108k - 600k Remote

python machine-learning hadoop spark

TIDYTEXT

THANK YOU

JULIA SILGE

@juliasilge

https://juliasilge.com

TIDYTEXT

THANK YOU

JULIA SILGE

@juliasilge

https://juliasilge.com

Author portraits from Wikimedia

Photos by Glen Noble and Kimberly Farmer on Unsplash