Natural Language Visualization with Scattertext

Jason S. Kessler*

Global AI Conference

April 27, 2018

Code for all visualizations is available at:

https://github.com/JasonKessler/GlobalAI2018

*No, not *that* Jason Kessler

@jasonkessler

Lexicon speculation

	Proposed word lists	Accuracy	Ties
Human 1	positive: *dazzling, brilliant, phenomenal, excellent, fantastic* negative: *suck, terrible, awful, unwatchable, hideous*	58%	75%
Human 2	positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting* negative: *bad, cliched, sucks, boring, stupid, slow*	64%	39%

Figure 1: Baseline results for human word lists. Data: 700 positive and 700 negative reviews.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. EMNLP. 2002.

Lexicon mining ≈ lexicon speculation

	Proposed word lists	Accuracy	Ties
Human 1	positive: *dazzling, brilliant, phenomenal, excellent, fantastic* negative: *suck, terrible, awful, unwatchable, hideous*	58%	75%
Human 2	positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting* negative: *bad, cliched, sucks, boring, stupid, slow*	64%	39%

Figure 1: Baseline results for human word lists. Data: 700 positive and 700 negative reviews.

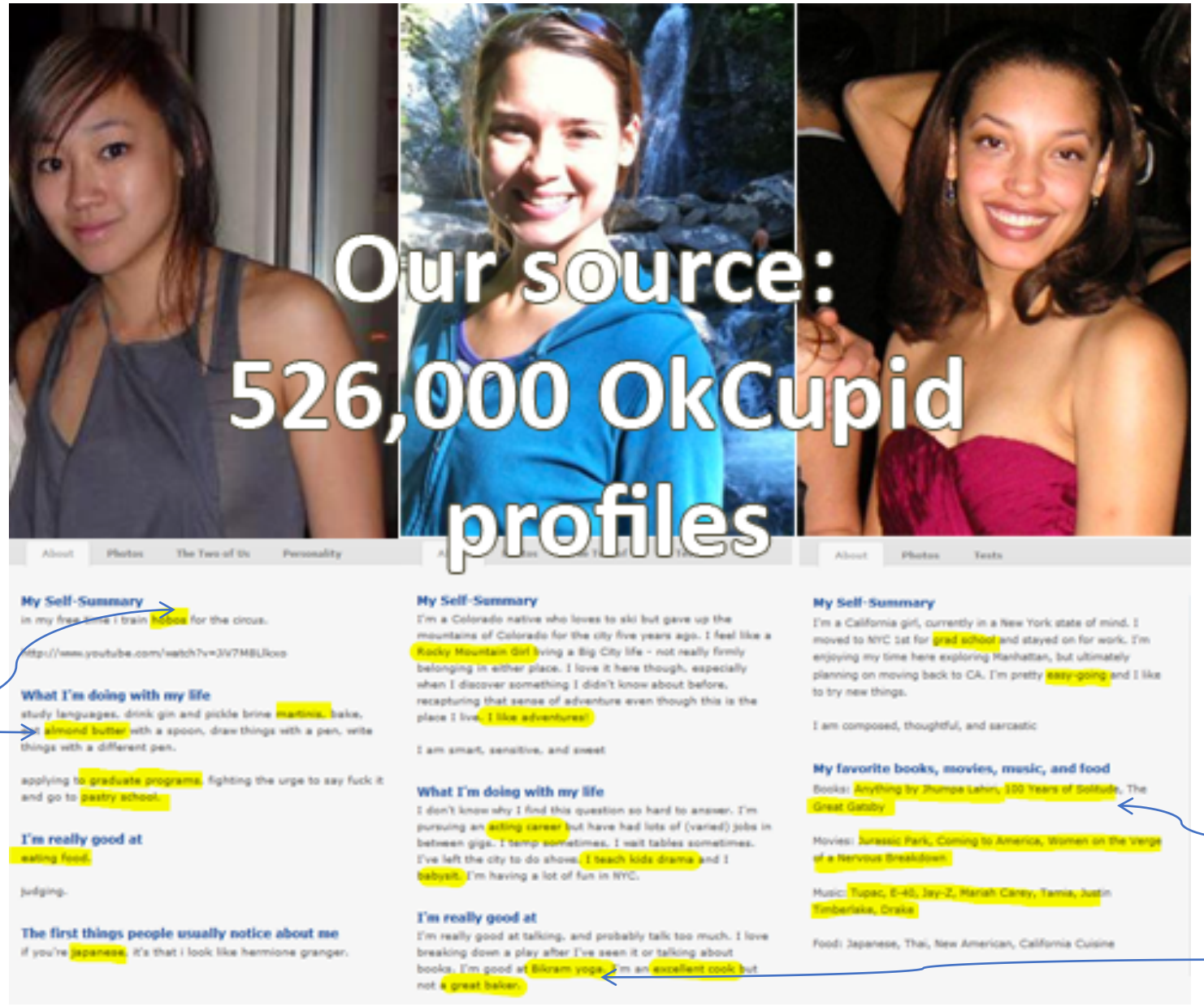
	Proposed word lists	Accuracy	Ties
Human 3 + stats	positive: *love, wonderful, best, great, superb, still, beautiful* negative: *bad, worst, stupid, waste, boring, ?, !*	69%	16%

Figure 2: Results for baseline using introspection and simple statistics of the data (including *test* data).

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. EMNLP. 2002.

@jasonkessler

Language and Demographics



hobos

almond butter

100 Years of Solitude

Bikram yoga

@jasonkessler

Christian Rudder: http://blog.okcupid.com/index.php/page/7/

OKCupid: Words and phrases that distinguish white men.

In general, I won't comment too much on these lists, because the whole point of this piece is to let the groups speak for themselves, but I have to say that the mind of the white man is the world's greatest sausagefest. Unless you're counting **Queens of the Stone Age**, there is not even one vaguely feminine thing on his list, and as far as broad categories go we have: sweaty guitar rock, bro-on-bro comedies, things with engines, and dystopias.

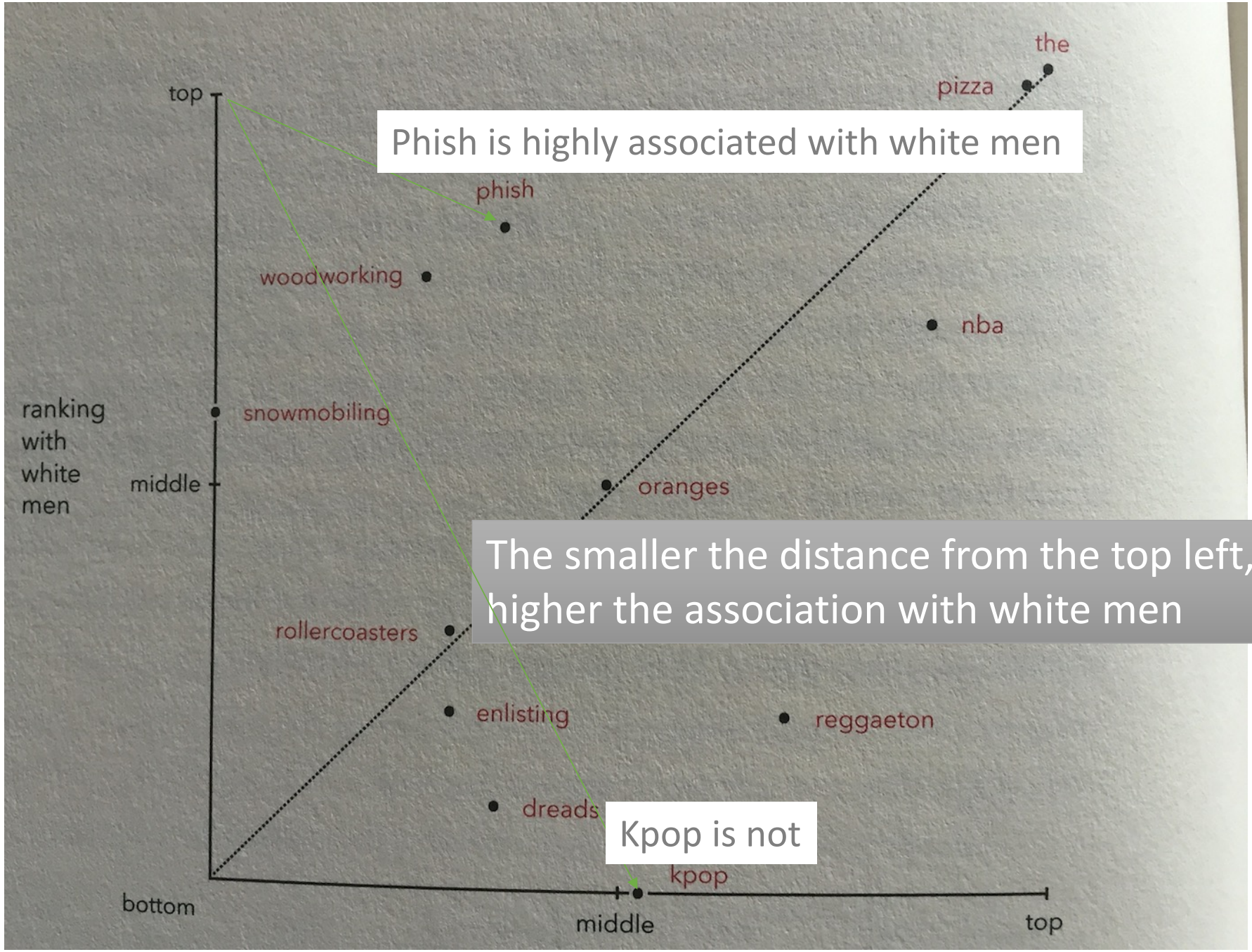
Source: http://blog.okcupid.com/index.php/page/7/ (Rudder 2010)

OKCupid: Words and phrases that distinguish Latin men.

Explanation

Music and dancing—**merengue**, **bachata**, **reggaeton**, **salsa**—are obviously very important to Latinos of both genders. The men have two other fascinating things going on: an interest in telling you about their sense of humor (**i'm a funny guy**, **very funny**, **outgoing and funny**, etc.) and an interest in industrial strength ass-kicking (**mma**, **ufc**, **boxing**, **marines**, etc.) Basically, if a Latin dude tells you a joke, you should laugh.

Source: http://blog.okcupid.com/index.php/page/7/ (Rudder 2010)



Source: Christian Rudder. Dataclysm. 2014.

Ranking with everyone else

@jasonkessler

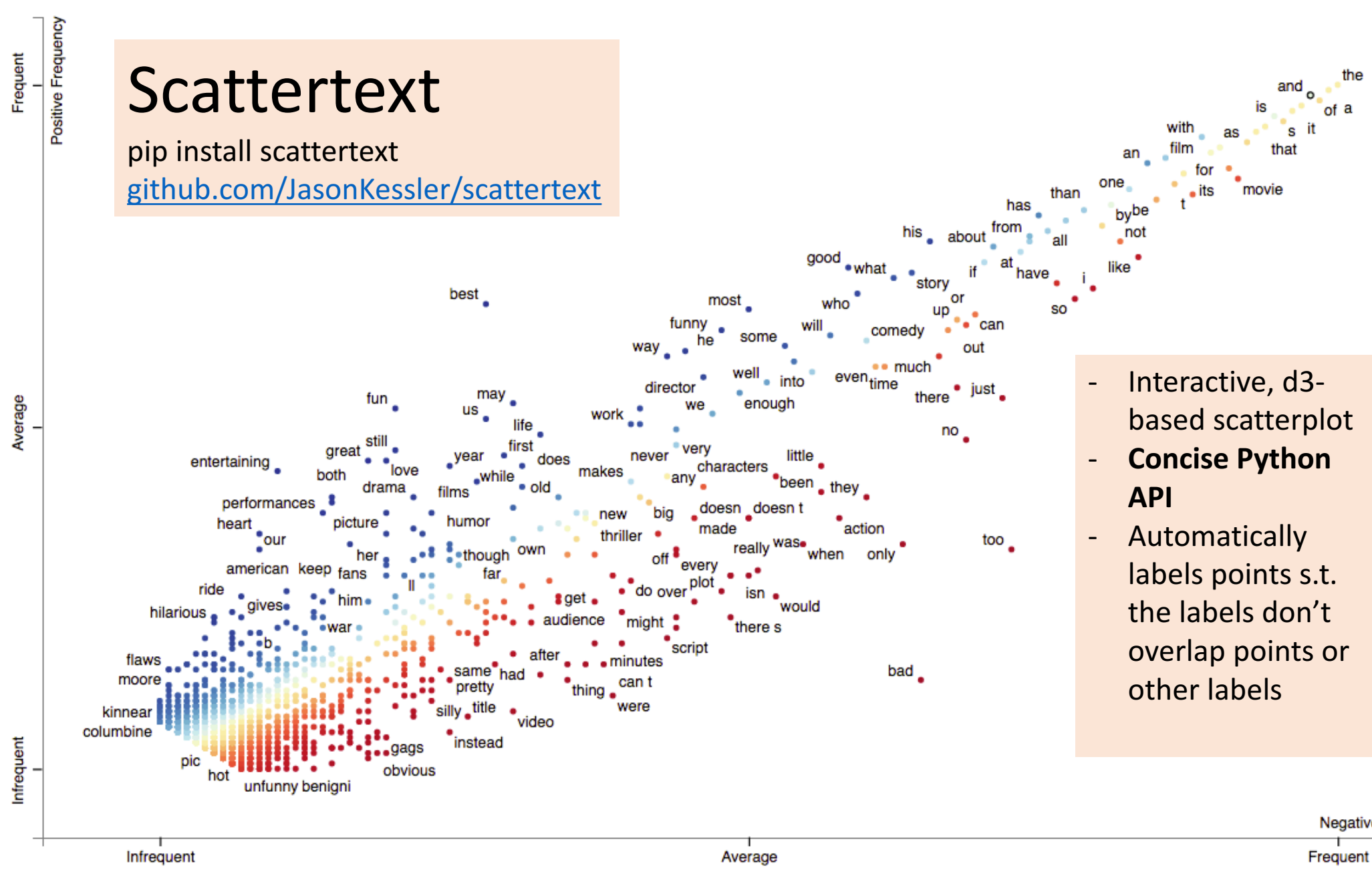
Source: Christian Rudder. Dataclysm. 2014.

@jasonkessler

Scattertext

pip install scattertext

github.com/JasonKessler/scattertext



- Interactive, d3-based scatterplot
- **Concise Python API**
- Automatically labels points s.t. the labels don't overlap points or other labels

Top Positive	Characteristic
best	benigni
entertaining	egoyan
fun	windtalkers
great	soderbergh
still	seagal
heart	drumline
love	forgettable
performance	shyamalan
both	hackneyed
performances	watchable
our	antwone
may	nickleby
us	formulaic
year	retread
	unfunny
	cliches
Top Negative	impostor
bad	inoffensive
too	vapid
only	cliched
would	narc
were	frida
there s	sayles
no	niro
script	cinematic
should	uninspired
really	gags
when	barbershop
action	scorsese
can t	suspenseful
was	

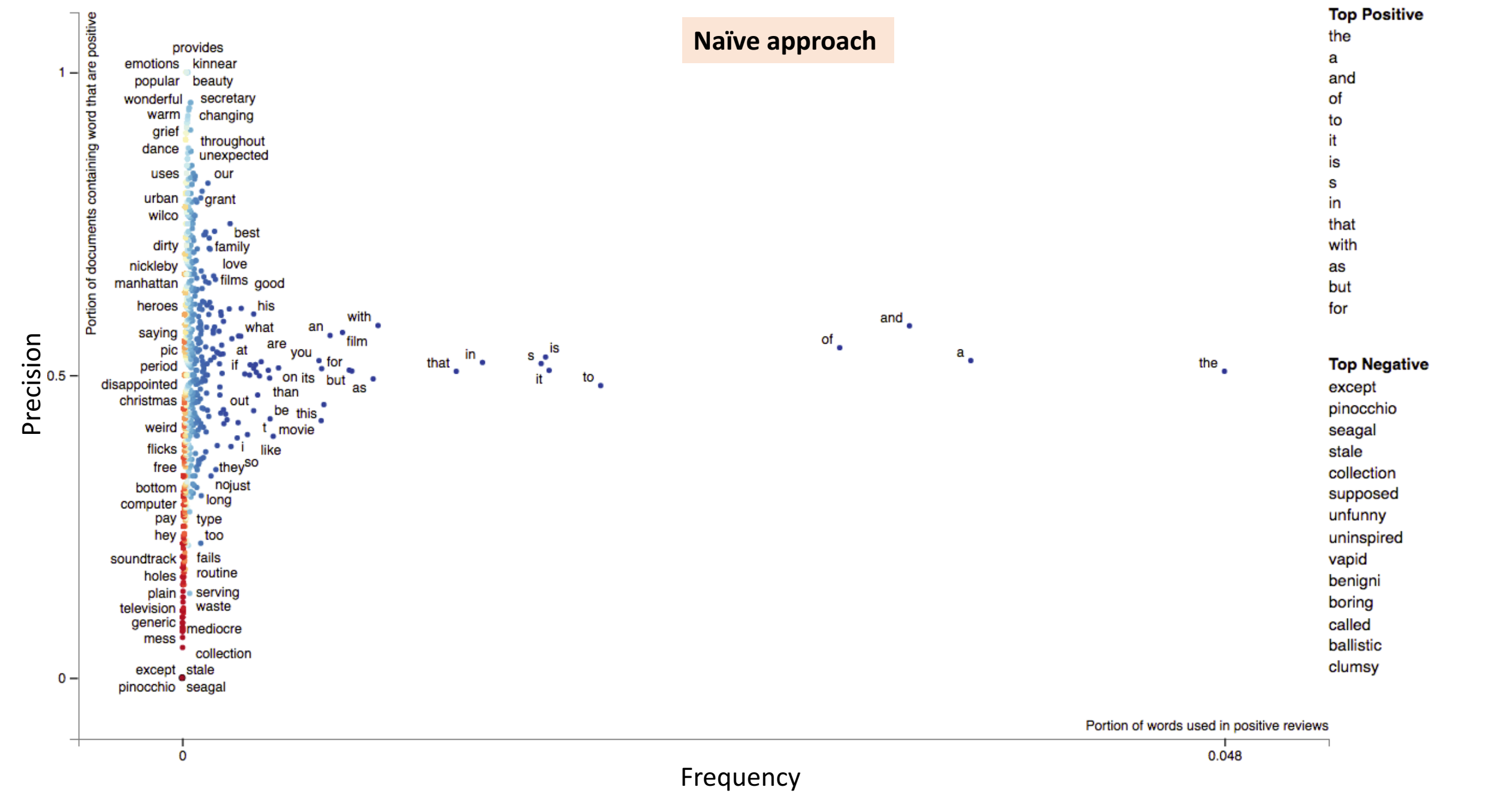
Jason S. Kessler. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. ACL System Demonstrations. 2017.

@jasonkessler

Scaled F-Score

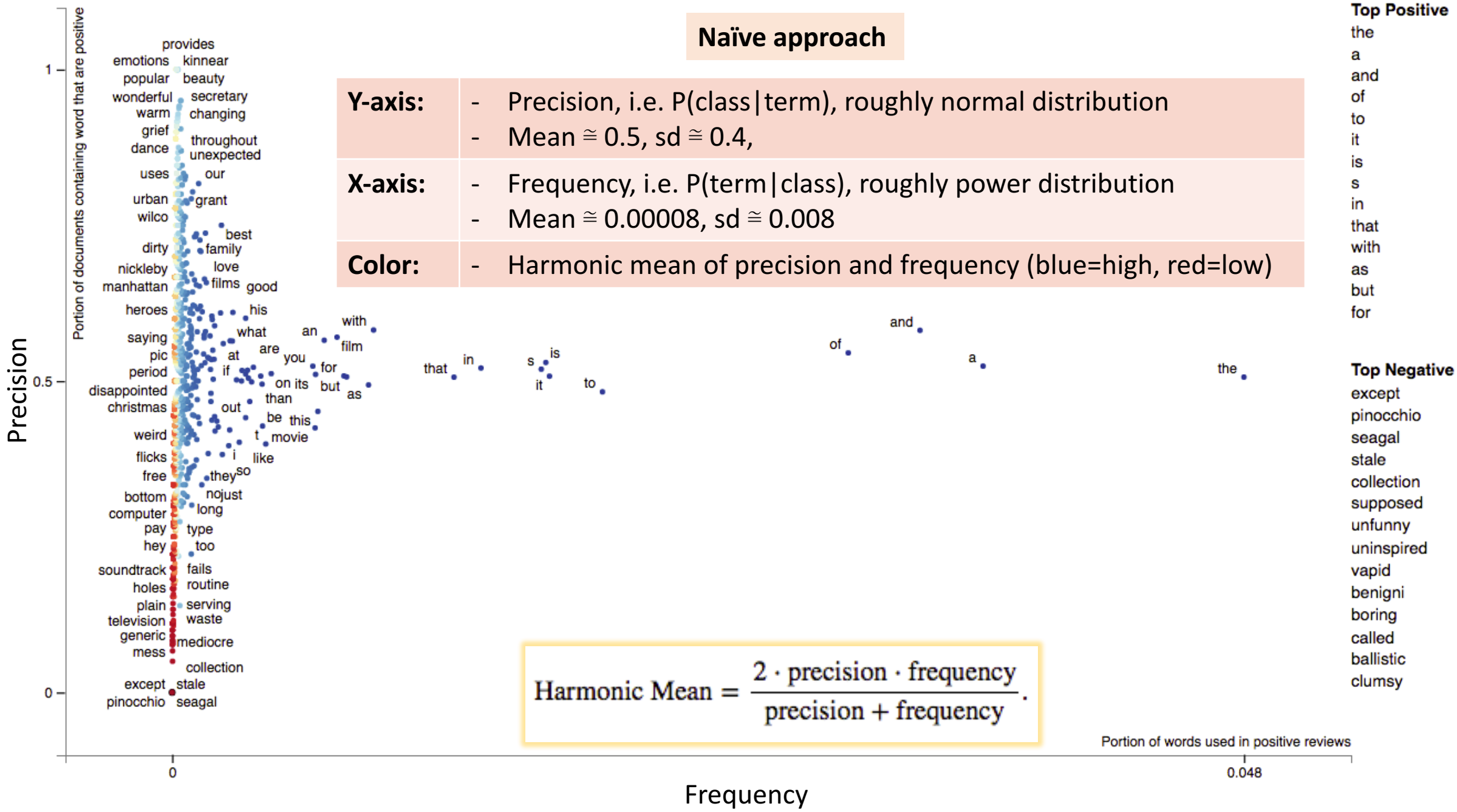
- Term-class* associations:
  - “Good” is associated with the “positive” class
  - “Bad” with the class “negative”
- Core intuition: association relies on two necessary factors
  - **Frequency**: How often a term occurs in a class
  - **Precision:** P(class|document contains term)
- F-Score:
  - Information retrieval evaluation metric
  - Harmonic mean between precision and recall
    - Requires **both** metrics to be high

- *Term: is defined to be a word, phrase or other discrete linguistic element


Naïve approach
Top Positive
the
a
and
of
to
it
is
s
in
that
with
as
but
for
Top Negative
except
pinocchio
seagal
stale
collection
supposed
unfunny
uninspired
vapid
benigni
boring
called
ballistic
clumsy
Portion of documents containing word that are positive
Precision
1
0.5
0
provides
emotions
kinnear
popular
beauty
wonderful
secretary
warm
changing
grief
dance
throughout
unexpected
uses
our
urban
grant
wilco
best
dirty
family
nickleby
love
manhattan
films
good
heroes
his
saying
what
an
with
pic
at
are
you
film
period
if
for
disappointed
on its
but
as
christmas
than
out
be
this
weird
t
movie
flicks
i
like
free
they
so
bottom
nojust
computer
long
pay
type
hey
too
soundtrack
fails
holes
routine
plain
serving
television
waste
generic
mediocre
mess
collection
except
stale
pinocchio
seagal
that
in
s
is
it
to
and
of
a
the
Portion of words used in positive reviews
0
0.048
Frequency
@jasonkessler

Naïve approach

Y-axis:	- Precision, i.e. P(class|term), roughly normal distribution - Mean ≅ 0.5, sd ≅ 0.4,
X-axis:	- Frequency, i.e. P(term|class), roughly power distribution - Mean ≅ 0.00008, sd ≅ 0.008
Color:	- Harmonic mean of precision and frequency (blue=high, red=low)



Top Positive
the
a
and
of
to
it
is
s
in
that
with
as
but
for

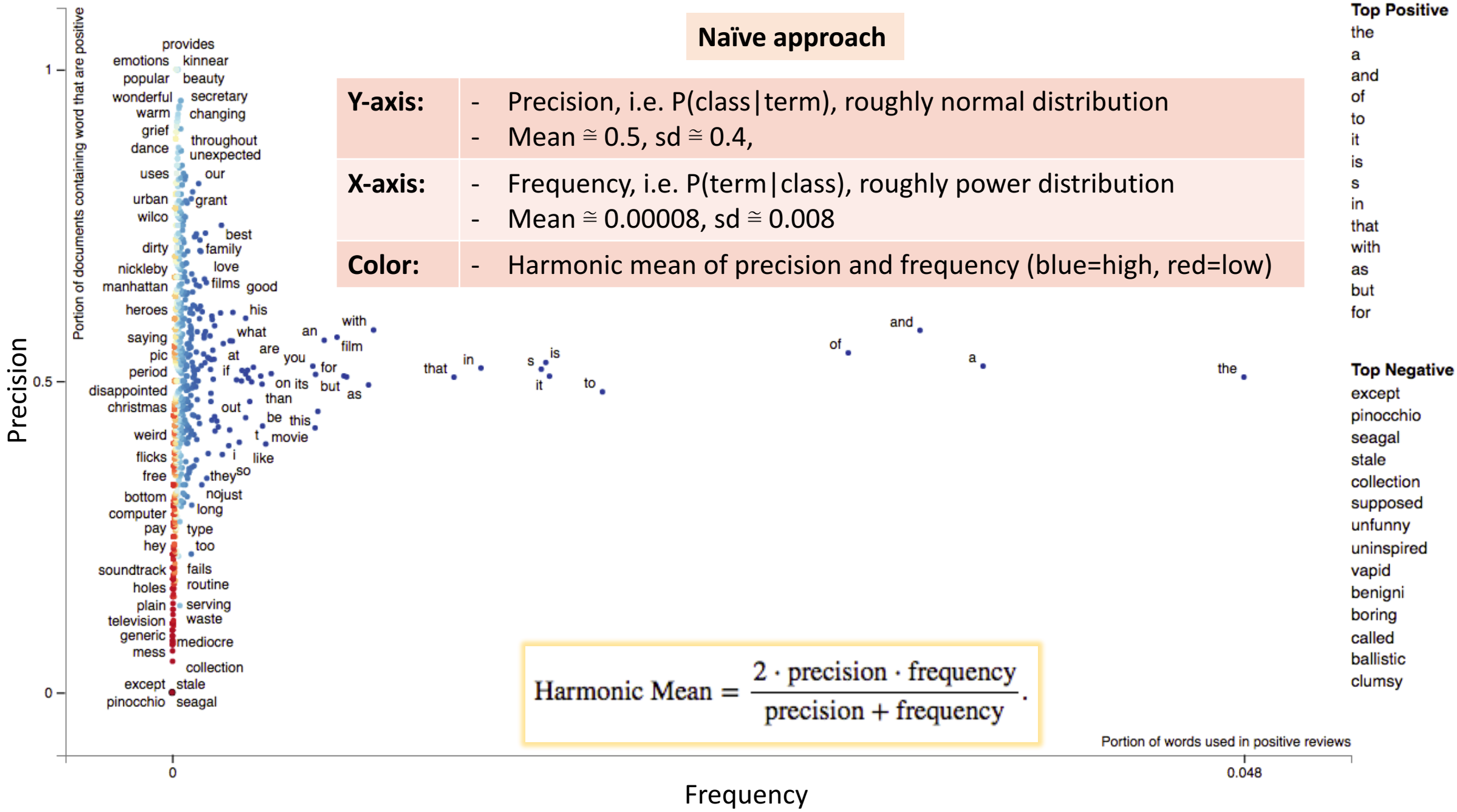
Top Negative
except
pinocchio
seagal
stale
collection
supposed
unfunny
uninspired
vapid
benigni
boring
called
ballistic
clumsy

$$\text{Harmonic Mean} = \frac{2 \cdot \text{precision} \cdot \text{frequency}}{\text{precision} + \text{frequency}}.$$

@jasonkessler

Naïve approach

Y-axis:	- Precision, i.e. P(class|term), roughly normal distribution - Mean ≅ 0.5, sd ≅ 0.4,
X-axis:	- Frequency, i.e. P(term|class), roughly power distribution - Mean ≅ 0.00008, sd ≅ 0.008
Color:	- Harmonic mean of precision and frequency (blue=high, red=low)

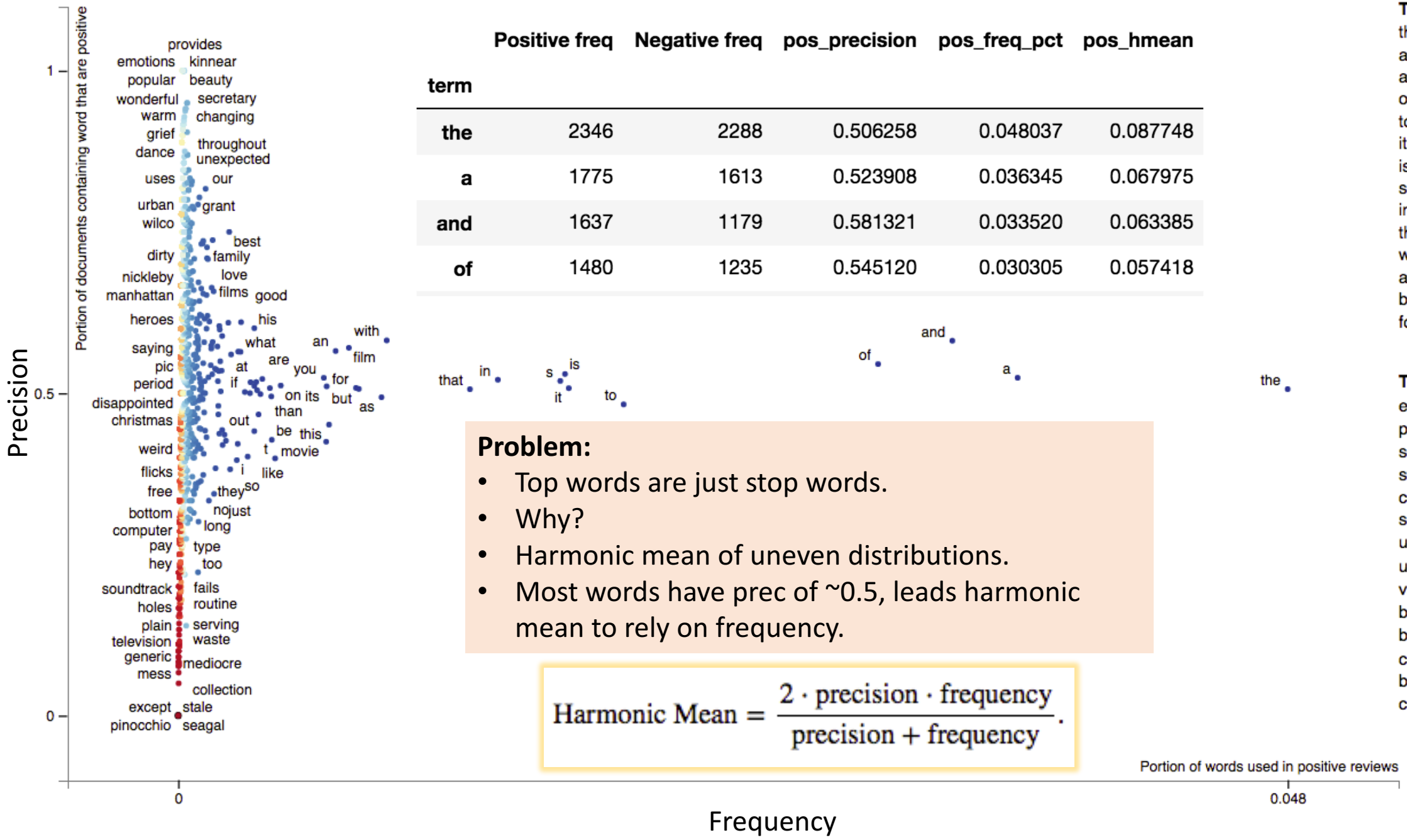


Top Positive
the
a
and
of
to
it
is
s
in
that
with
as
but
for

Top Negative
except
pinocchio
seagal
stale
collection
supposed
unfunny
uninspired
vapid
benigni
boring
called
ballistic
clumsy

$$\text{Harmonic Mean} = \frac{2 \cdot \text{precision} \cdot \text{frequency}}{\text{precision} + \text{frequency}}.$$

@jasonkessler



term	Positive freq	Negative freq	pos_precision	pos_freq_pct	pos_hmean
the	2346	2288	0.506258	0.048037	0.087748
a	1775	1613	0.523908	0.036345	0.067975
and	1637	1179	0.581321	0.033520	0.063385
of	1480	1235	0.545120	0.030305	0.057418

Top Positive

the
a
and
of
to
it
is
s
in
that
with
as
but
for

Top Negative

except
pinocchio
seagal
stale
collection
supposed
unfunny
uninspired
vapid
benigni
boring
called
ballistic
clumsy

Problem:

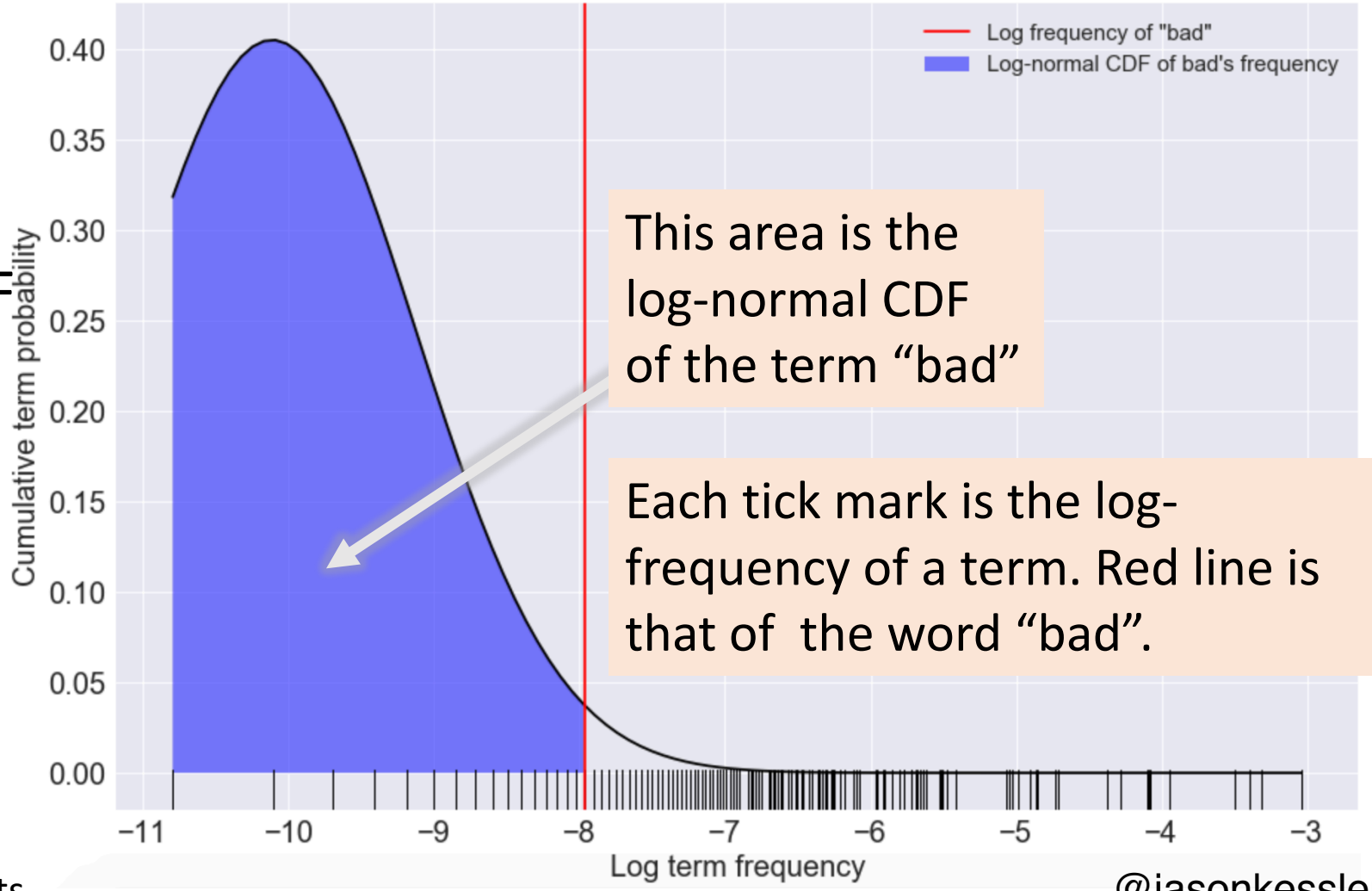
- Top words are just stop words.
- Why?
- Harmonic mean of uneven distributions.
- Most words have prec of ~0.5, leads harmonic mean to rely on frequency.

$$\text{Harmonic Mean} = \frac{2 \cdot \text{precision} \cdot \text{frequency}}{\text{precision} + \text{frequency}}.$$

@jasonkessler

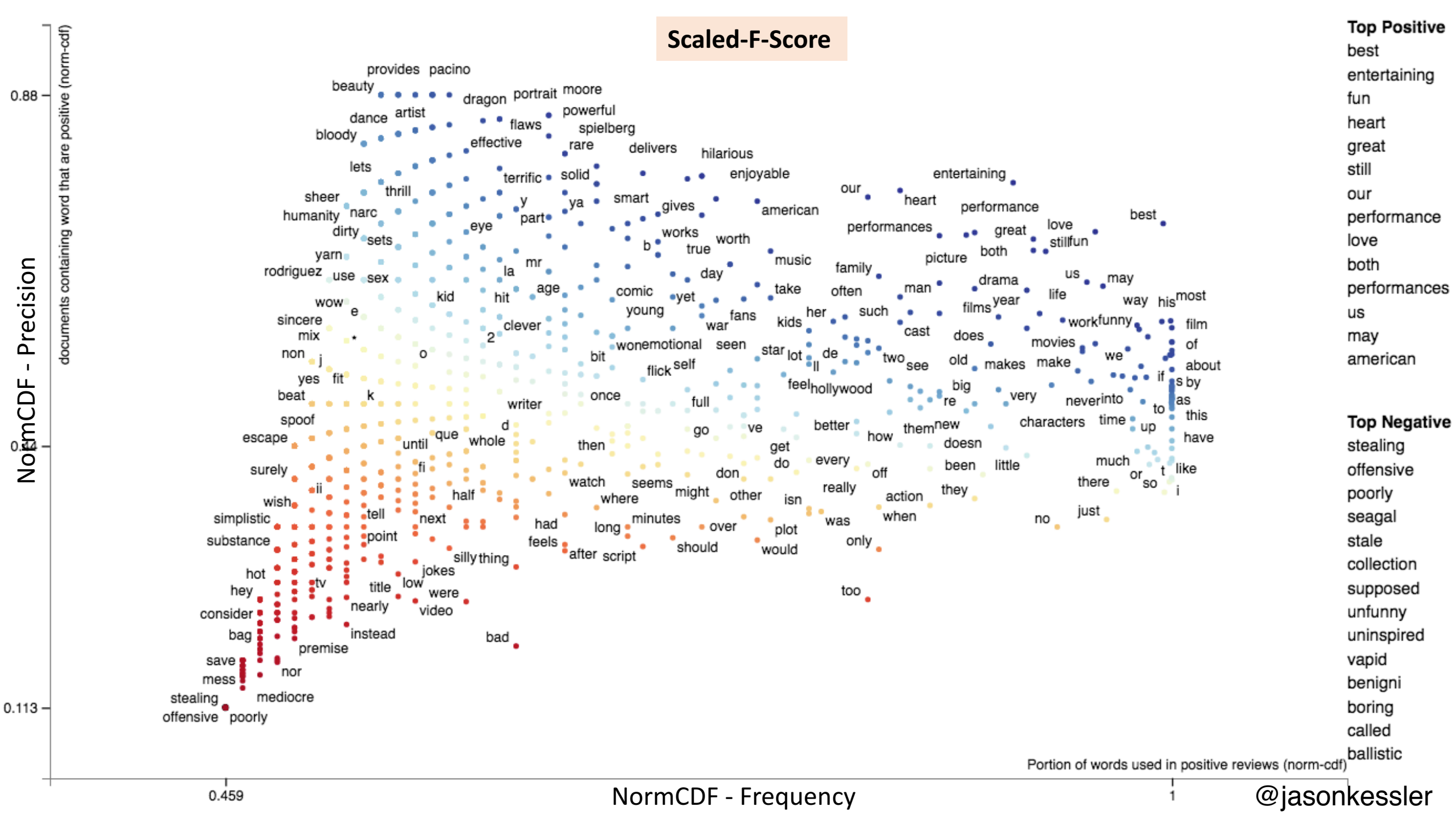
Fix: Normalize Precision and Frequency

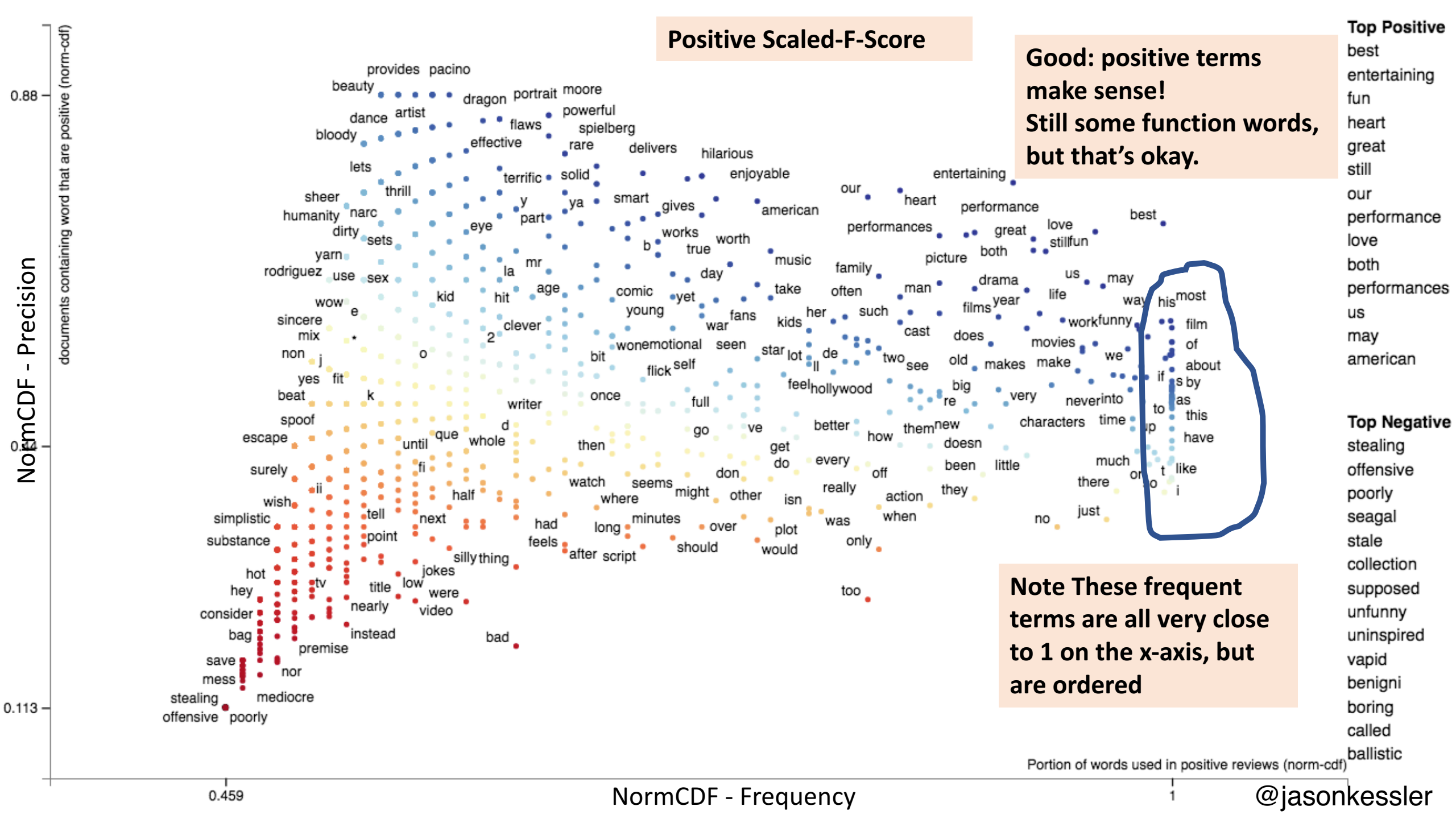
- Task: make precision and frequency similarly distributed
- How: take normal CDF of each term's precision and frequency
- Mean and std. computed from data
- Right: log normal CDF

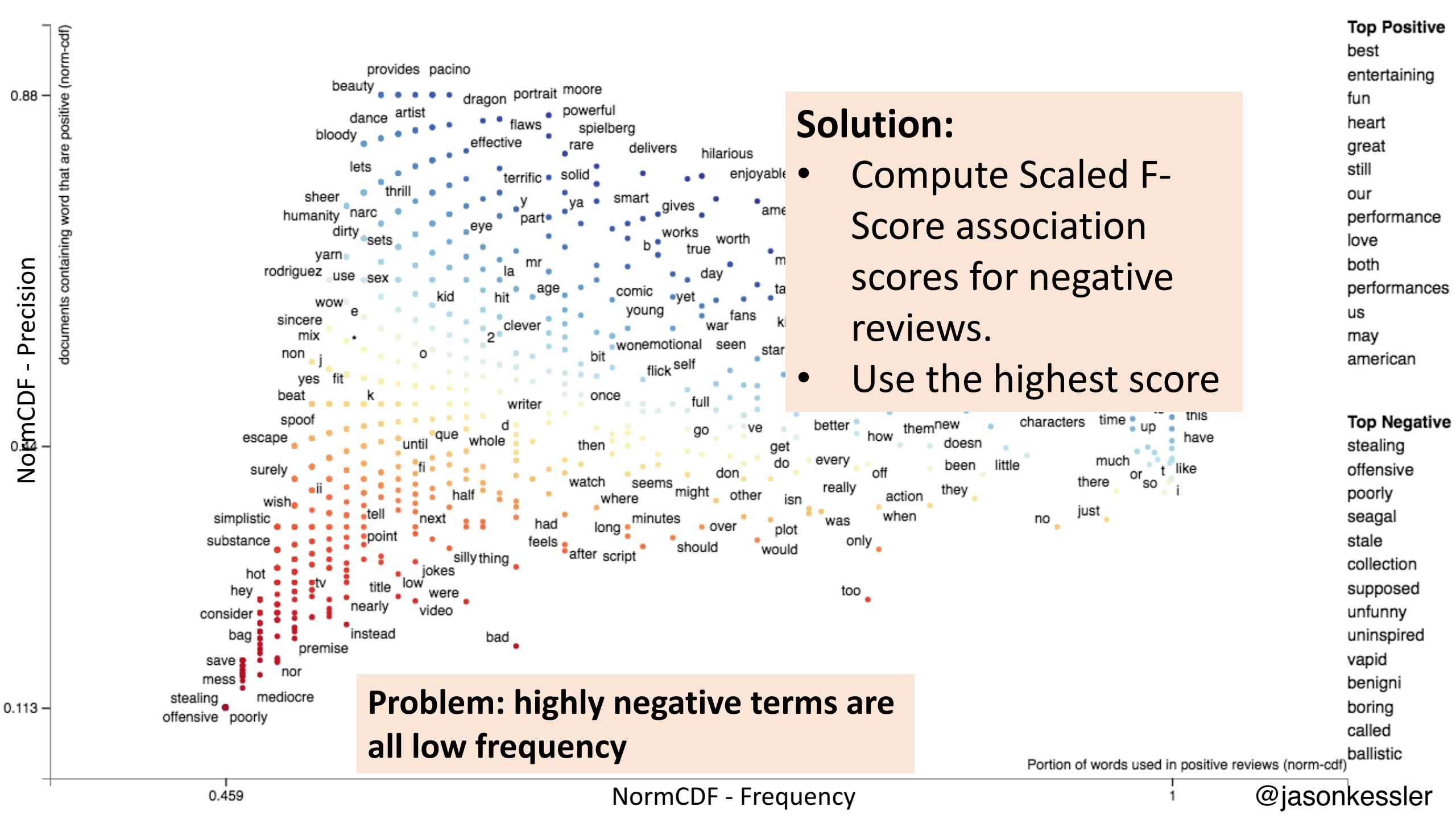


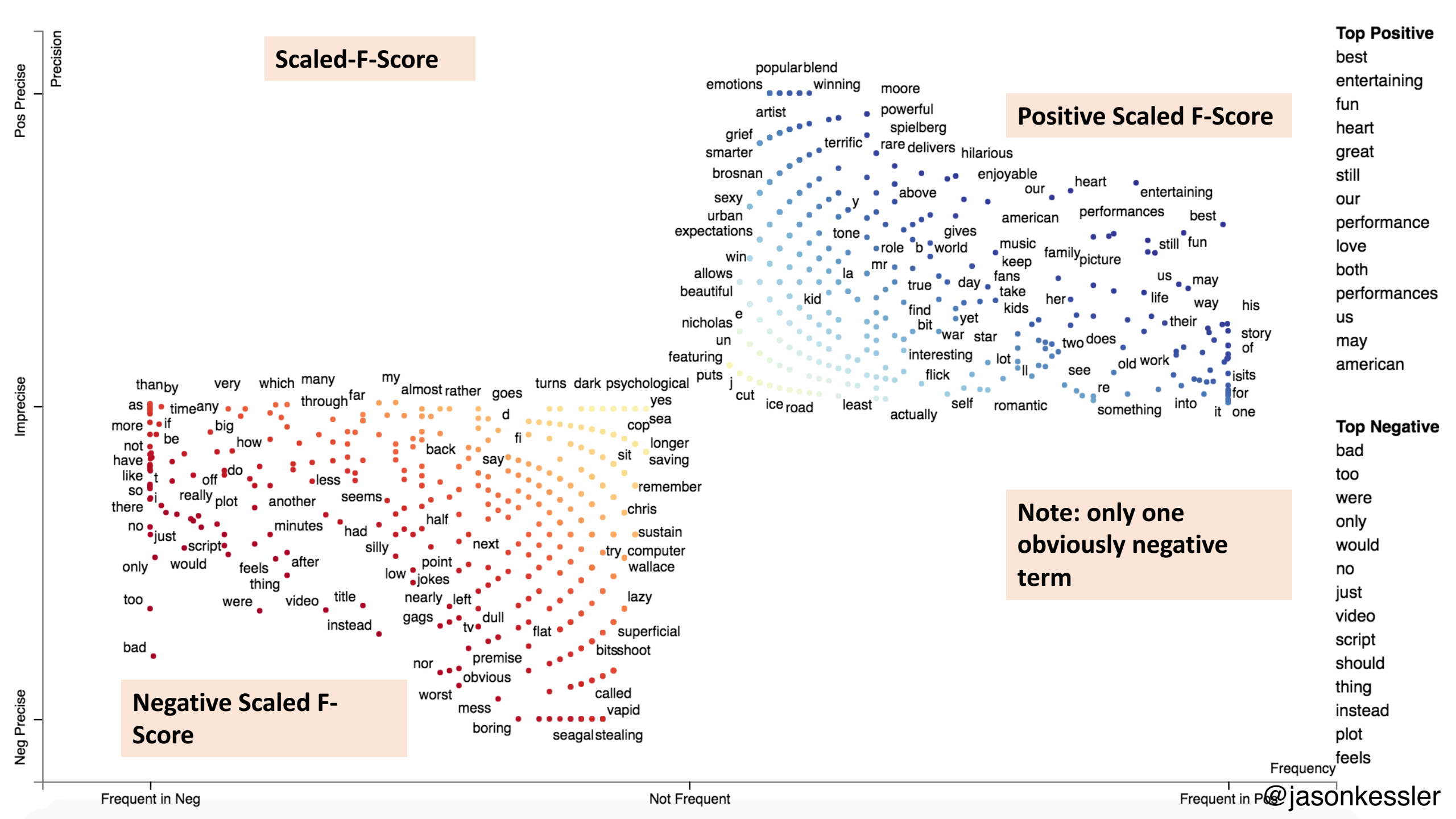
*log-normal CDF isn't used in these charts

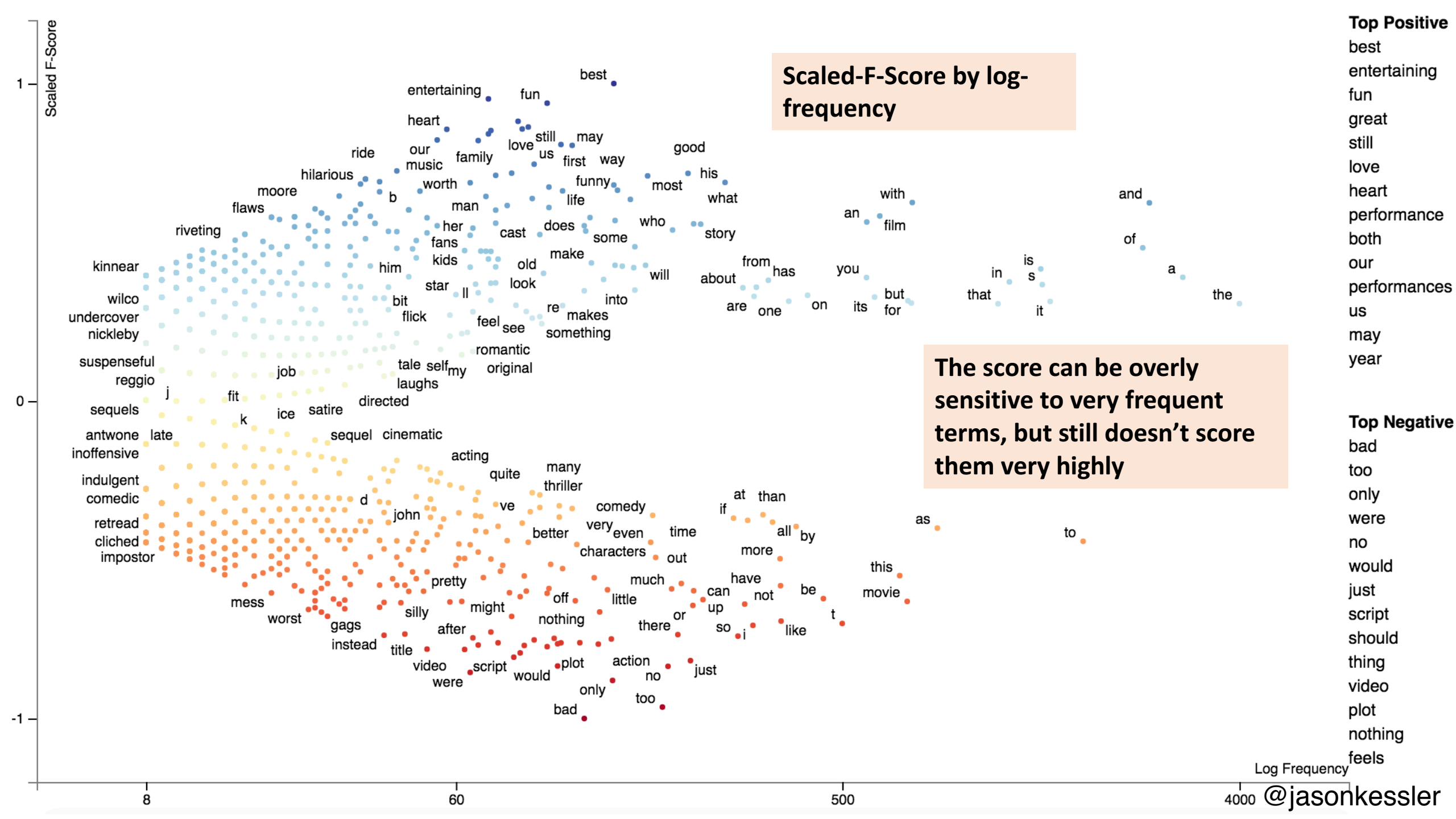
@jasonkessler

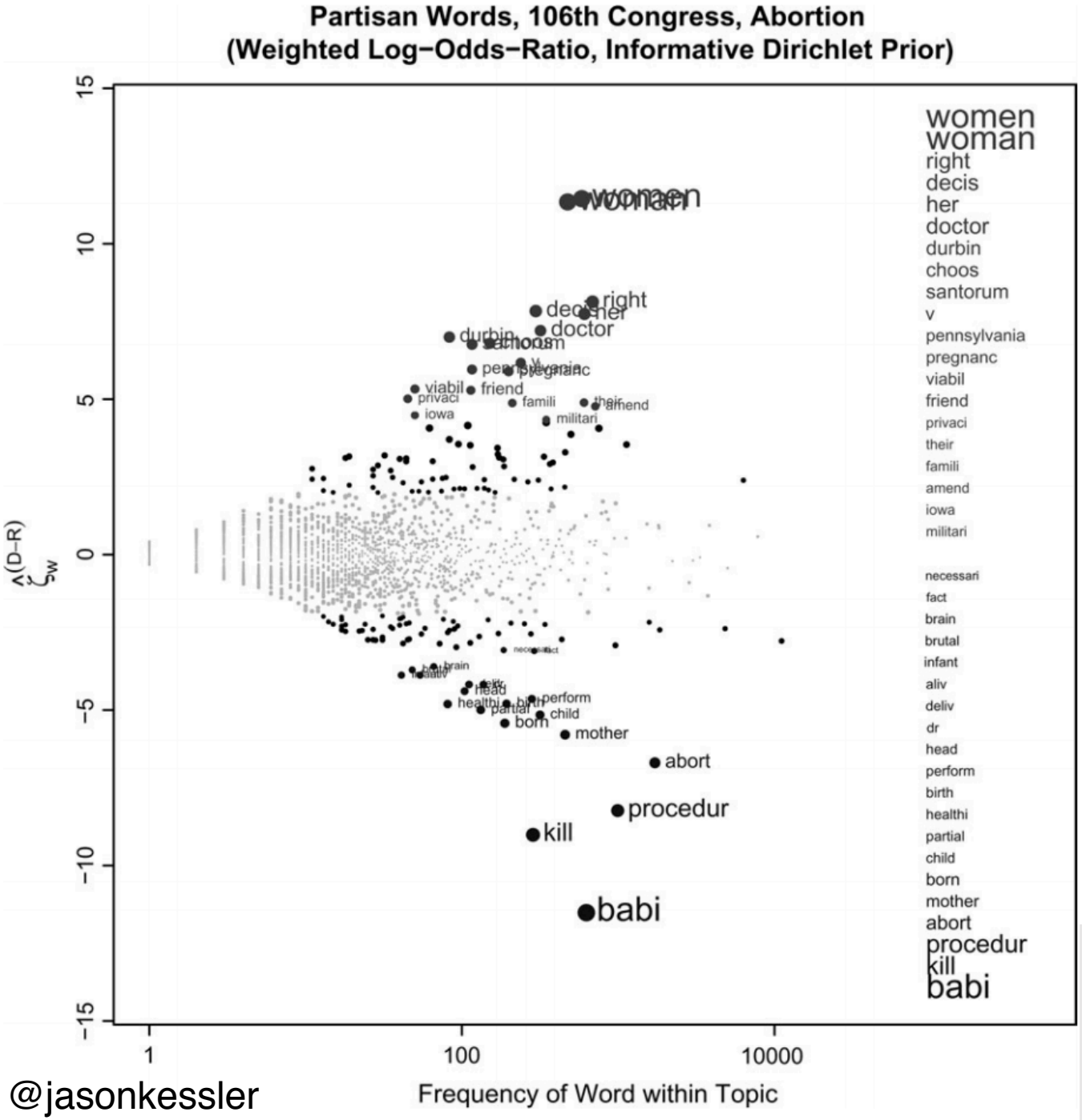

Scaled-F-Score
NormCDF - Precision
documents containing word that are positive (norm-cdf)
0.88
0.113
provides pacino
beauty
dragon
portrait
moore
dance
artist
powerful
bloody
flaws
spielberg
effective
rare
delivers
hilarious
lets
terrific
solid
enjoyable
entertaining
sheer
thrill
y
ya
smart
gives
our
heart
performance
humanity
narc
part
american
best
dirty
eye
works
performances
great
love
sets
b
worth
stillfun
yarn
true
picture
both
mr
music
family
rodriguez
use
sex
la
day
drama
us
may
wow
kid
hit
age
comic
yet
take
often
man
life
way
his
most
e
young
films
year
sincere
clever
fans
her
such
work
funny
film
mix
2
war
kids
cast
does
movies
of
non
wonemotional
seen
star
lot
de
two
see
old
makes
make
we
about
j
o
bit
ll
yes
fit
flick
self
feel
hollywood
big
if
s
by
beat
k
once
re
very
never
into
as
writer
full
to
spoof
d
go
ve
better
how
them
new
characters
time
up
this
escape
que
whole
then
get
doesn
have
until
do
every
surely
fi
off
been
little
much
or
t
like
ii
watch
seems
might
other
isn
really
action
they
there
so
i
wish
half
where
simplistic
tell
next
minutes
over
was
when
no
just
had
long
plot
substance
point
feels
after
script
should
would
only
silly
thing
too
hot
jokes
hey
tv
title
low
were
consider
nearly
video
bag
instead
bad
premise
save
nor
mess
mediocre
stealing
offensive
poorly
0.459
1
Portion of words used in positive reviews (norm-cdf)
NormCDF - Frequency
@jasonkessler
Top Positive
best
entertaining
fun
heart
great
still
our
performance
love
both
performances
us
may
american
Top Negative
stealing
offensive
poorly
seagal
stale
collection
supposed
unfunny
uninspired
vapid
benigni
boring
called
ballistic


Positive Scaled-F-Score
Good: positive terms make sense!
Still some function words, but that's okay.
Note These frequent terms are all very close to 1 on the x-axis, but are ordered
Top Positive
best
entertaining
fun
heart
great
still
our
performance
love
both
performances
us
may
american
Top Negative
stealing
offensive
poorly
seagal
stale
collection
supposed
unfunny
uninspired
vapid
benigni
boring
called
ballistic
NormCDF - Precision
documents containing word that are positive (norm-cdf)
0.88
0.113
NormCDF - Frequency
0.459
1
Portion of words used in positive reviews (norm-cdf)
@jasonkessler


NormCDF - Precision
documents containing word that are positive (norm-cdf)
0.88
0.113
provides pacino
beauty
dragon portrait moore
dance artist
powerful
bloody
flaws
spielberg
effective
rare
delivers
hilarious
lets
terrific
solid
enjoyable
sheer
thrill
y
ya
smart
gives
humanity narc
part
eye
works
worth
dirty
sets
b
true
yarn
mr
rodriguez use sex
la
day
wow
kid
hit
age
comic
yet
e
young
fans
sincere
clever
war
mix
2
wonemotional
seen
non
j
o
bit
flick self
yes fit
beat
k
once
writer
full
spoof
d
go
ve
better
how
them new
characters
time
up
this
escape
que
until
whole
then
get
have
surely
fi
do
every
been
little
much
like
ii
watch
seems
might
other
isn
really
off
or
t
there
wish
half
where
action
they
so
i
simplistic
tell
next
had
minutes
over
was
when
no
just
substance
point
feels
long
plot
hot
silly thing
after script
should
would
only
hey
jokes
tv
title
low
were
too
consider
nearly
video
bag
instead
bad
premise
save
nor
mess
stealing
mediocre
offensive poorly
0.459
NormCDF - Frequency
1
Portion of words used in positive reviews (norm-cdf)
Solution:
• Compute Scaled F-Score association scores for negative reviews.
• Use the highest score
Problem: highly negative terms are all low frequency
Top Positive
best
entertaining
fun
heart
great
still
our
performance
love
both
performances
us
may
american
Top Negative
stealing
offensive
poorly
seagal
stale
collection
supposed
unfunny
uninspired
vapid
benigni
boring
called
ballistic
@jasonkessler


Scaled-F-Score
Precision
Pos Precise
Imprecise
Neg Precise
Positive Scaled F-Score
popular blend
emotions winning
moore
artist
powerful
spielberg
grief
terrific
rare delivers
hilarious
smarter
brosnan
enjoyable
our
heart
entertaining
sexy
above
y
urban
american
performances
best
expectations
tone
gives
win
role
b
world
music
still
fun
family
picture
keep
allows
la
mr
fans
us
may
beautiful
true
day
take
her
life
way
his
kid
kids
nicholas
e
find
yet
their
story
bit
un
war
star
two
does
of
featuring
interesting
lot
ll
see
old
work
isits
puts
j
flick
re
for
cut
ice
road
least
actually
self
romantic
something
into
it
one
Negative Scaled F-Score
than by
very
which
many
my
as
time
any
through
far
almost
rather
goes
turns
dark
psychological
yes
more
if
big
d
cop
sea
be
how
fi
not
back
sit
longer
have
say
saving
like
t
do
off
less
remember
so
i
really
plot
another
seems
there
chris
half
no
minutes
had
sustain
just
silly
next
try
computer
script
only
would
feels
after
point
wallace
low
jokes
thing
too
were
video
title
nearly
left
lazy
instead
gags
tv
dull
flat
superficial
bad
nor
premise
bitsshoot
obvious
worst
called
mess
vapid
boring
seagal
stealing
Note: only one obviously negative term
Frequency
Frequent in Neg
Not Frequent
Frequent in Pos
Top Positive
best
entertaining
fun
heart
great
still
our
performance
love
both
performances
us
may
american
Top Negative
bad
too
were
only
would
no
just
video
script
should
thing
instead
plot
feels
@jasonkessler


Scaled-F-Score by log-frequency
The score can be overly sensitive to very frequent terms, but still doesn't score them very highly
Scaled F-Score
Log Frequency
1
0
-1
8
60
500
4000
Top Positive
best
entertaining
fun
great
still
love
heart
performance
both
our
performances
us
may
year
Top Negative
bad
too
only
were
no
would
just
script
should
thing
video
plot
nothing
feels
@jasonkessler



@jasonkessler

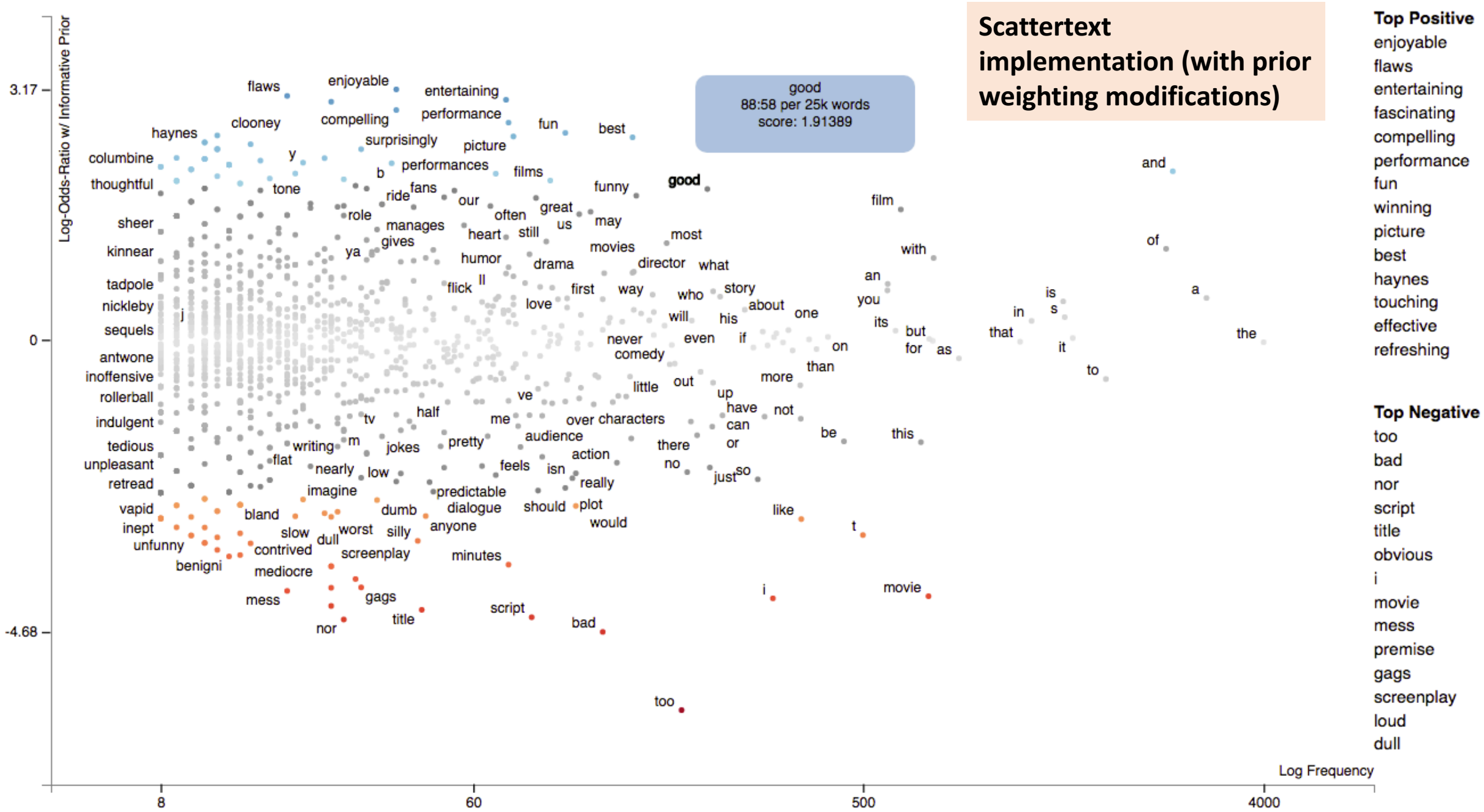
Monroe et. al (2009) approach

- Bayesian approach to term-association
- **Likelihood**: Z-score of log-odds-ratio
- **Prior**: Term frequency in a background corpus
- **Posterior**: Z-score of log-odds-ratio with background counts as smoothing values

Popular, but much more tweaking to get to work than Scaled F Score.

Burt Monroe, Michael Colaresi and Kevin Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis. 2008.

Scattertext implementation (with prior weighting modifications)

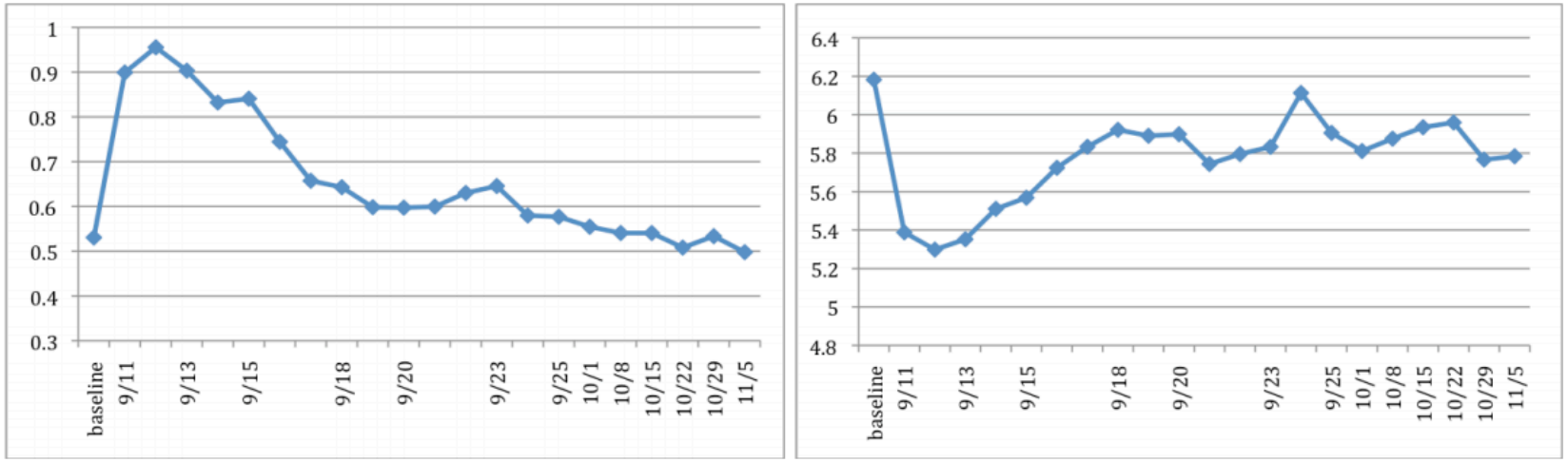


Scattertext reimplementation of Monroe et al. See http://nbviewer.jupyter.org/github/JasonKessler/PuPPyTalk/blob/master/notebooks/Class-Association-Scores.ipynb for code.

@jasonkessler

In defense of stop words

Figure 1. Pronoun Use by Bloggers Before and After September 11, 2001



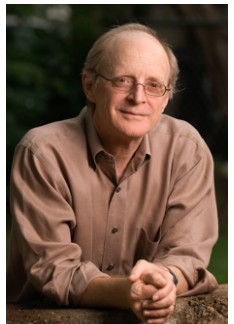
We-words I-words

Note. Graphs reflect percentage of we-words (left) and I-words (right) within daily blog entries of 1,084 bloggers in the two months surrounding September 11, 2001.

In times of shared crisis, "we" use increases, while "I" use decreases.

I/we: age, social integration

I: lying, social rank



Cindy K. Chung and James W. Pennebaker. Counting Little Words in Big Data: The Psychology of Communities, Culture, and History. EASP. 2012

Function words and gender

LIWC Dimension Bold: entirely stop words	Effect Size (Cohen's d) (>0 F, <0 M) MANOVA p<.001
All Pronouns (esp. 3rd person)	0.36
Present tense verbs (walk, **is, be**)	0.18
Feeling (touch, hold, feel)	0.17
Certainty (always, never)	0.14
Word count	NS
Numbers	-0.15
Prepositions	-0.17
Words >6 letters	-0.24
Swear words	-0.22
Articles	-0.24

- Performed on a variety of language categories, including speech.
- Other studies have found that function words are the best predictors of gender.

Newman, ML; Groom, CJ; Handelman LD, Pennebaker, JW. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. 2008.

Function word usage is counter-intuitive

- Pennebaker et al:
  - Testosterone levels (in two therapeutic settings) predict:
    - Modest but significant decreases in:
      - Pronouns referring to others (ex: we, she, they)
      - Communication verbs (ex: hear, say)
    - Modest but significant decreases in:
      - Optimism words (ex: energy, upbeat)
  - Negative (but statistically insignificant) correlation with subject's beliefs about testosterone and category usage!
  - Does not entirely explain gender differences.
- Herring et al:
  - Subjects tasked with impersonating opposite gender in online game
  - Discussed stereotypical topics (cars, shopping) but **didn't change stylistic cues**

- James W. Pennebaker, Carla J. Groom, Daniel Loew, James M. Dabbs. Testosterone as a Social Inhibitor: Two Case Studies of the Effect of Testosterone Treatment on Language. 2004.
- Susan C. Herring, Anna Martinson. Assessing Gender Authenticity in Computer-Mediated Language Use: Evidence From an Identity Game. Journal of Language and Social Psychology. 2004.

Clickbait: what works?

The New York Times

Video by NEETI UPADHYE. Photo by Mark Makela/Getty Images

CRITIC'S NOTEBOOK

Cliff Huxtable Was Bill Cosby's Sickest Joke

By WESLEY MORRIS 8:58 PM ET

Mr. Cosby's signature TV character was a patient, wise father figure. In other words, our critic writes, the ideal cover for terrible behavior.

Cosby Has a Few Words for the Court

By LIAM STACK 10:20 PM ET

Mr. Cosby has not said much about the sexual assault allegations against him, but on Thursday he erupted with a vulgarity aimed at the prosecutor.

· Did the #MeToo Movement Affect the Cosby Verdict?

Opinion

North Korea's Phony Peace Ploy

By NICHOLAS EBERSTADT

If the past is any guide, Pyongyang will offer Seoul unenforceable verbiage at this week's summit meeting.

The Supreme Court and the New Civil War

By LINDA GREENHOUSE

The battle between the White House and blue states raises questions about the limits of federal authority.

· When Misogynists Become Terrorists
· Krugman: Trump's War on the Poor
· Bruni: Ronny Jackson, From Fawning to Fiasco

EDITORIAL

A Reckoning for Cosby — Now for Others?

Cosby's persistent pathology was unusual. The fears of his victims are not.

· The Cosby Jury Finally Believes the Women
· Editorial: Mr. Macron Comes to Washington
· Meet the New Boss. Quite Different From the Old Boss.
· Editorial: El Salvador's Chanc to Save Women's Lives
· The Future of Frats
· Dear Abby: #MeToo
· A Trade War the U.S. Is Actually Winning, for Now
· Blow: The White Rebellion
· Stephens: Bush 41, Trump, ar American Decline
· We Are Republican Teachers. It's Time to Raise Taxes.

BuzzFeed News Videos Quizzes Tasty As/Is

This Test Will Reveal Which Dog Breed Best Suits You

Sniff out the answer with this scientific quiz.

Sarah Aspler • 6 hours ago

Shopping

31 Things That'll Make You Say "Well, There Goes All My Money"

Sloth car decals, a birthday cake candle (complete with sprinkles), gorgeous knit maxi dresses, and more! Much more!

21 Things You Probably Didn't Know You Could Take An Online Class For

Even If You Buy Everything On This List, You Still Won't Spend More Than $50

25 Amazing Products With Over 1,000 Reviews On Amazon

31 One-Piece Bathing Suits That'll Make You Ditch Your Bikini

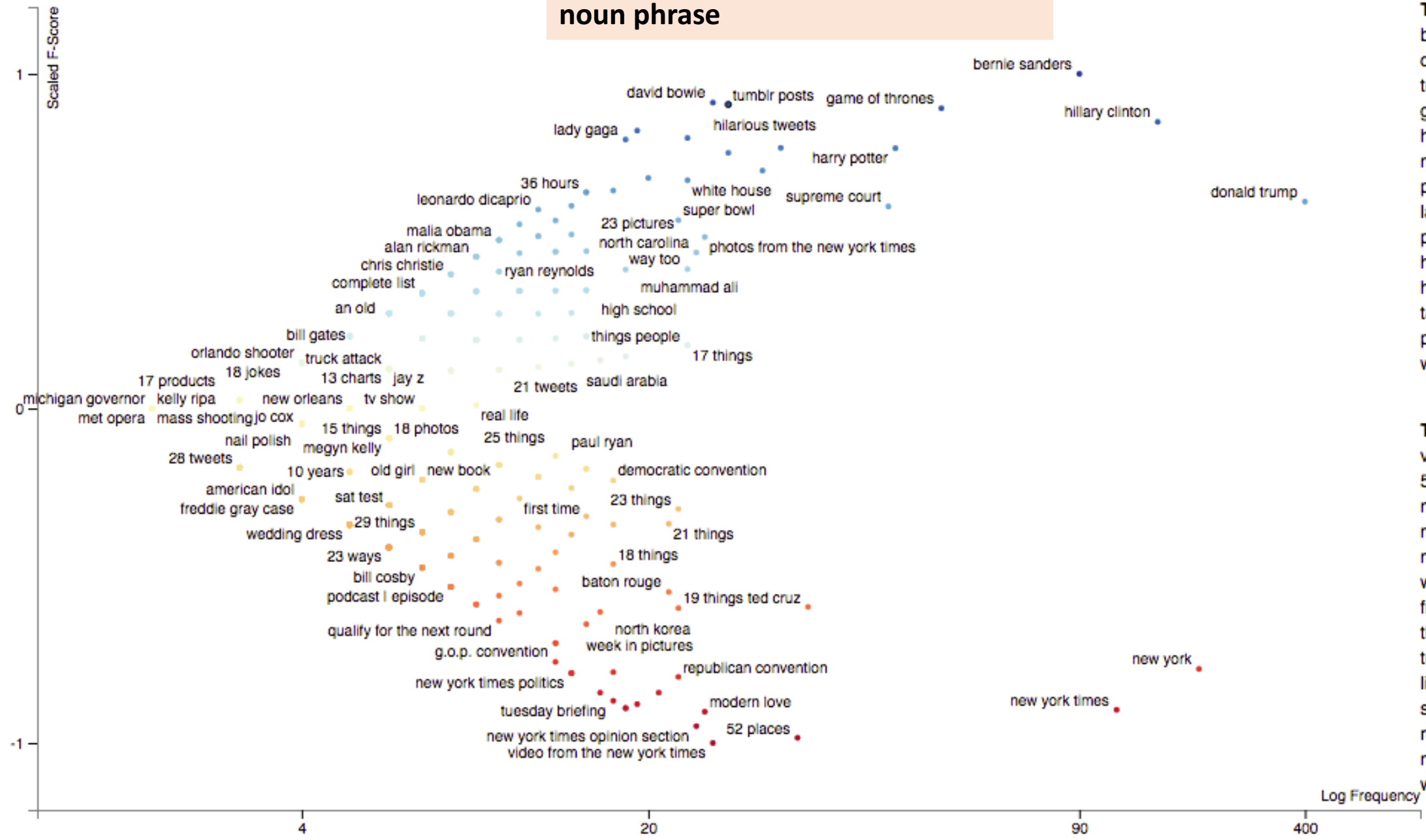
BuzzFeed's 7-Day Spring Cleaning Challenge Will Help You Declutter Your Life!

Clickbait corpus

- Facebook posts from BuzzFeed, NY Times, etc/ from 2010s.
- Includes headline and the number of Facebook likes
- Scraped by researcher Max Woolf at github.com/minimaxir/clickbait-cluster.
- We'll separate articles from 2016 into the upper third and lower third of likes.
- Identify words and phrases that predict likes.
- Begin with noun phrases identified from Phrase Machine (Handler et al. 2016)
- Filter out redundant NPs.

Abram Handler, Matt Denny, Hanna Wallach, and Brendan O'Connor. Bag of what? Simple noun phrase extraction for corpus analysis. NLP+CSS Workshop at EMNLP 2016.

Scaled-F-Score of engagement by noun phrase



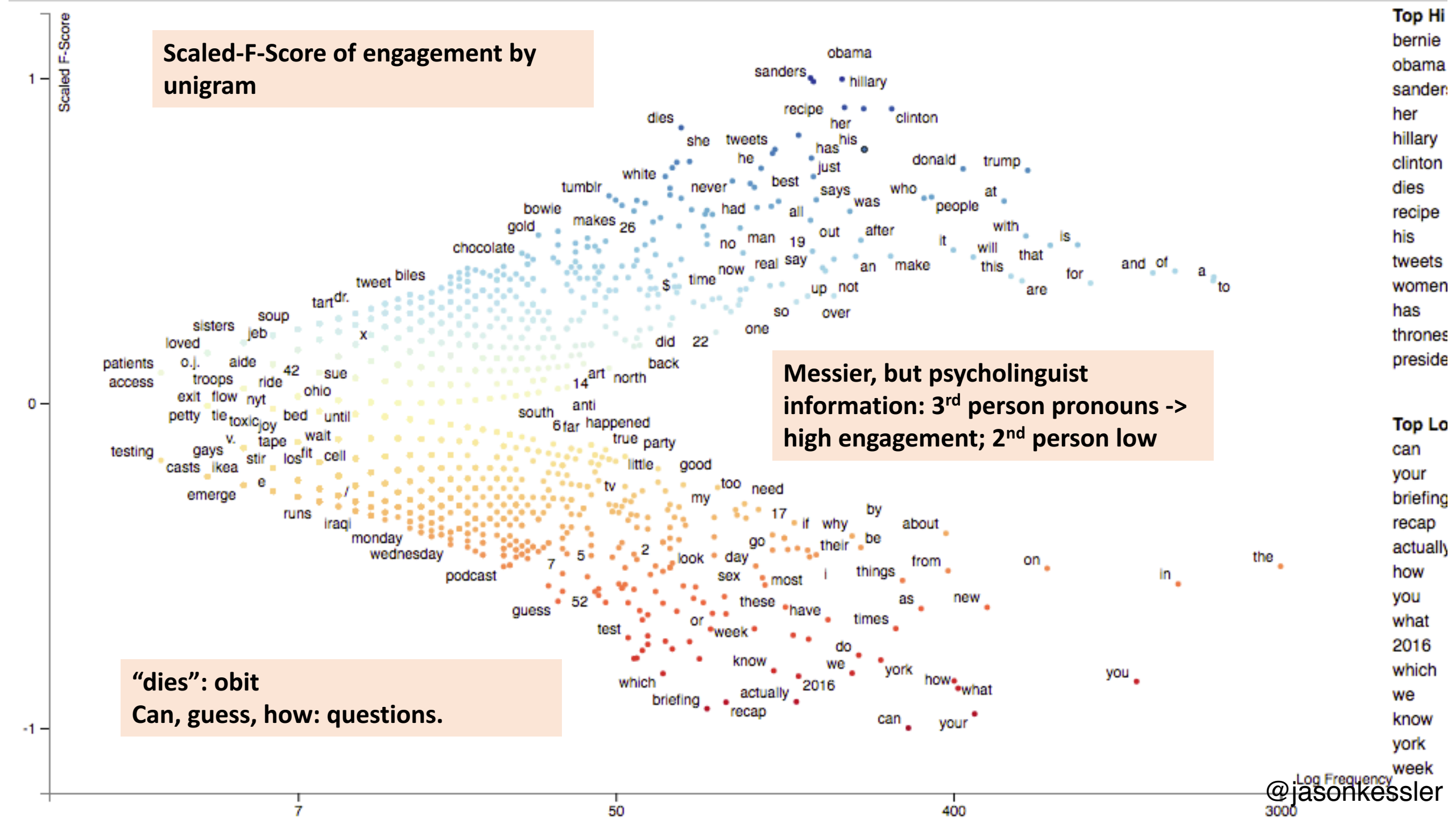
Top Hi

bernie sanders
david bowie
tumblr posts
game of thrones
hillary clinton
michelle obama
president obama
lady gaga
pokémon go
harry potter
hilarious tweets
taylor swift
pope francis
white house

Top Lo

video from the new york times
52 places
new york times opinion section
modern love
new york times
wednesday briefing
friday briefing
thursday briefing
tuesday briefing
live updates
silicon valley
republican convention
new york times politics
week in pictures

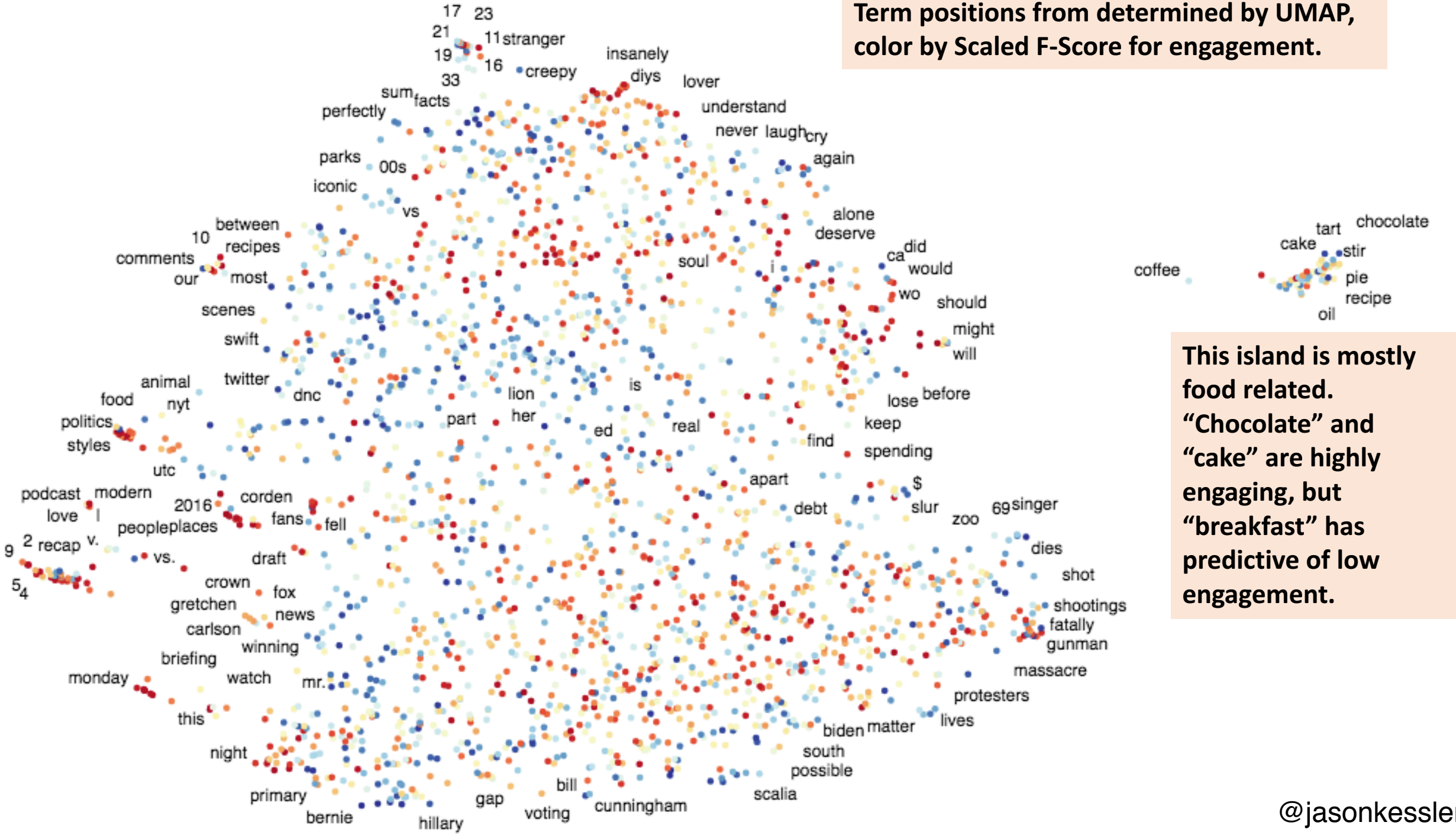
@jasonkessler


Scaled-F-Score of engagement by unigram
Scaled F-Score
1
0
-1
Log Frequency
7
50
400
3000
obama
sanders
hillary
recipe
clinton
her
dies
she
tweets
his
has
he
just
donald
trump
white
never
best
says
who
at
tumblr
had
all
was
people
bowie
makes
26
gold
out
after
with
chocolate
no
man
19
it
is
will
that
now
real
say
an
make
this
for
and
of
a
to
time
up
not
are
so
over
one
tweet
biles
tart
dr.
soup
sisters
jeb
x
loved
did
22
patients
o.j.
aide
back
42
sue
art
north
14
access
troops
ride
ohio
exit
flow
nyt
anti
petty
tie
toxic
joy
bed
until
south
6
far
happened
true
party
v.
tape
wait
testing
gays
stir
los
fit
cell
little
good
casts
ikea
e
tv
too
need
my
emerge
17
if
why
by
about
runs
iraqi
go
be
their
monday
wednesday
2
look
day
from
on
the
7
5
podcast
sex
most
i
things
in
guess
52
these
have
as
new
test
or
week
times
do
know
we
york
how
what
you
which
briefing
actually
2016
recap
can
your
Messier, but psycholing uist information: 3rd person pronouns -> high engagement; 2nd person low
“dies”: obit
Can, guess, how: questions.
Top Hi
bernie
obama
sander
her
hillary
clinton
dies
recipe
his
tweets
women
has
throne
preside
Top Lo
can
your
briefing
recap
actually
how
you
what
2016
which
we
know
york
week
@jasonkessler

Clickbait corpus

- How do terms with similar meanings differ in terms of their engagement rates?
- Use Gensim (https://radimrehurek.com/gensim/) to find word embeddings
- Use UMAP (McInnes and Healy 2018) to project them into two dimensions, and explore them with Scattertext.
  - Locally groups words with similar embeddings together.
  - Better alternative to T-SNE; allows for cosine instead of Euclidean distance criteria

Leland McInnes, John Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Arxiv. 2018.



Term positions from determined by UMAP, color by Scaled F-Score for engagement.

This island is mostly food related. “Chocolate” and “cake” are highly engaging, but “breakfast” has predictive of low engagement.

@jasonkessler

Clickbait corpus

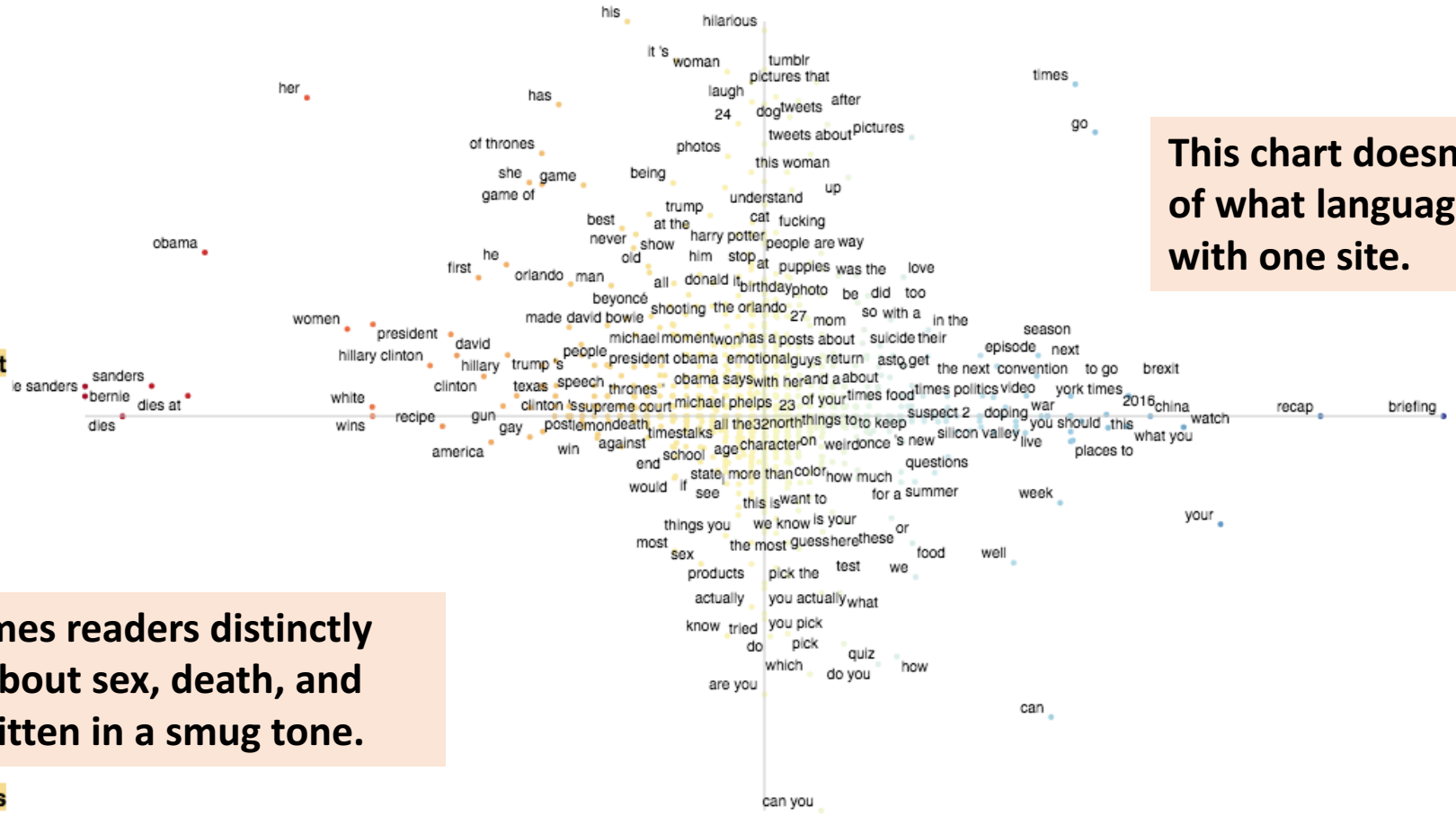
- How do the Times and Buzzfeed differ in what they talk about, and their content engages their readers?
- Scattertext can easily create visualizations to help answer these questions.
- First, we'll look at how what engages for Buzzfeed contrasts with what engages for the Times, and vice versa

Leland McInnes, John Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Arxiv. 2018.

Appeals to all
her, obama, bernie, sanders, women, bernie sanders, president, dies at, has, of thrones

BuzzFeed: ↑ Engagement
hilarious, tumblr, woman, laugh, tweets, after, dog, it 's, pictures that, 26

Ignored by elite, appeals to masses
go, times, brexit, this, season, to go, pictures, next, war, the new



This chart doesn't give a good sense of what language is more associated with one site.

NY Times: ↑ Engagement
bernie sanders, bernie, dies, sanders, dies at, obama, women, white, wins, president

NY Times: ↓ Engagement
briefing, recap, watch, china, brexit, what you, your, 2016, this, watch this

Oddly, NY Times readers distinctly like articles about sex, death, and which are written in a smug tone.

Appeals to elite, ignored by masses
know, actually, sex, most, products, the most, how well, things you, dies, we know

BuzzFeed: ↓ Engagement
can you, are you, which, do you, can, how, quiz, do, know, pick

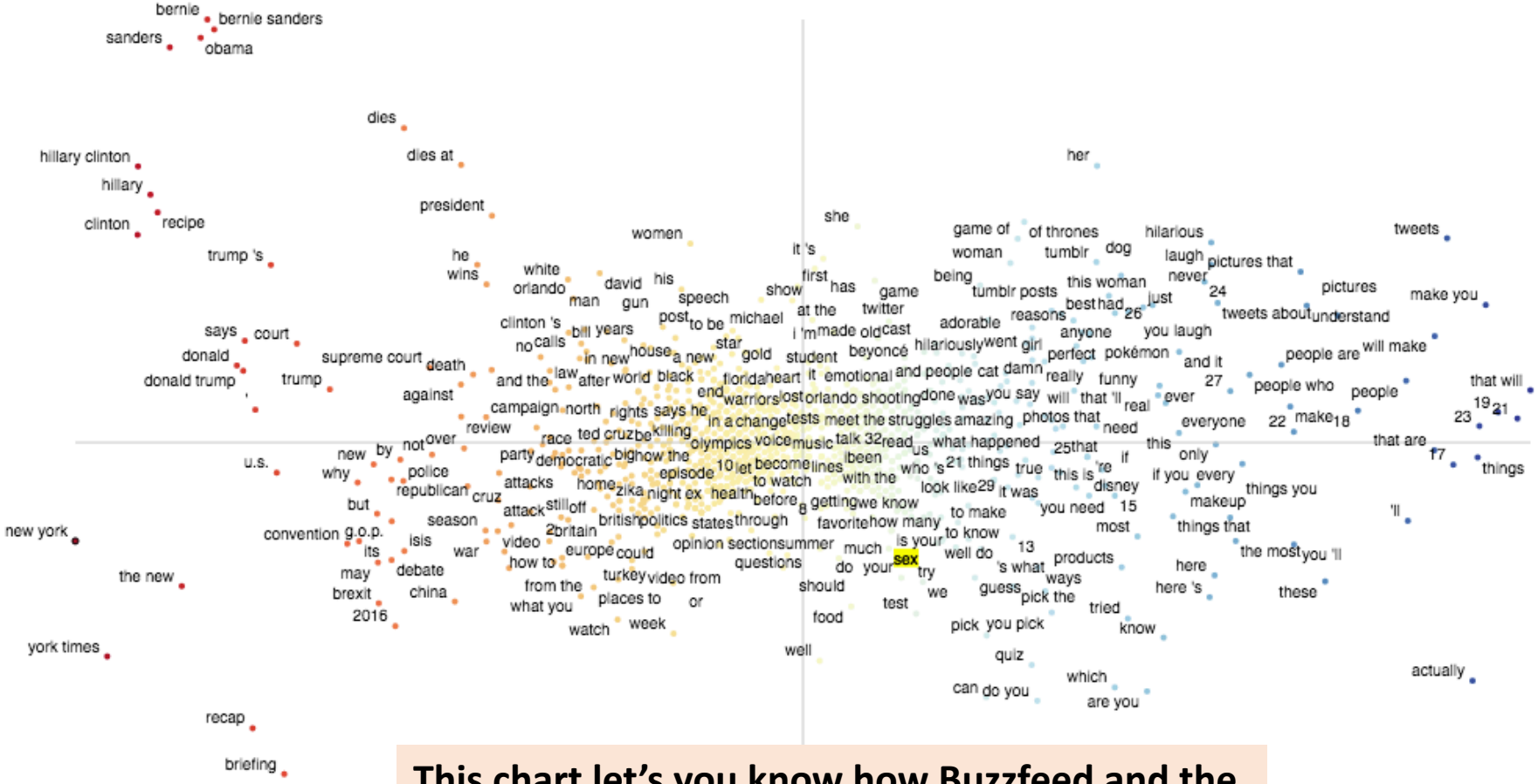
Ignored by all
can, can you, how, do you, your, well, which, quiz, are you, we

@jasonkessler

Highbrow Engagment
sanders, obama, bernie, bernie sanders, hillary clinton, hillary, recipe, clinton, trump 's, dies

Many Facebook Reactions
she, it 's, women, has, first, game, of thrones, game of, her, at the

Lowbrow Engagment
tweets, make you, pictures that, will make, pictures, understand, hilarious, laugh, that will, 24



NYTimes
new york, ', donald, donald trump, u.s., says, the new, court, clinton, trump

BuzzFeed
21, 19, 23, things, that will, 17, that are, people, will make, 'll

This chart let's you know how Buzzfeed and the Times are distinct, while still distinguishing engaging content,

Highbrow Ignored
briefing, recap, york times, the new, 2016, new york, brexit, may, china, convention

Few Facebook Reactions
can, well, do you, week, quiz, pick, food, test, which, watch

Lowbrow Ignored
can you, actually, are you, these, which, know, here 's, here, 'll, do you

Thank you! Questions?

Jason S. Kessler

Global AI Conference

April 27, 2018

https://github.com/JasonKessler/GlobalAI2018