Level Up:
Fancy NLP with Straightforward Tools

Kimberly Fessel, PhD

@kimberlyfessel • kimberlyfessel

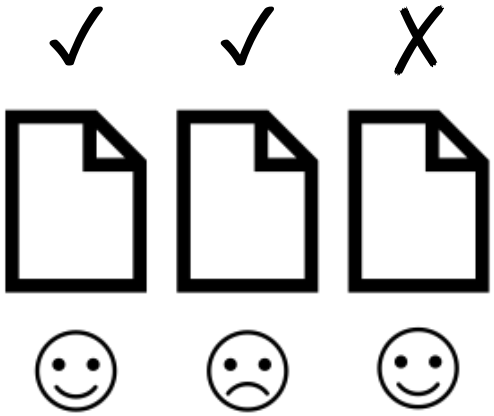
#ODSC

N L P

Natural Language Processing

Pre-Processing

Text Classification

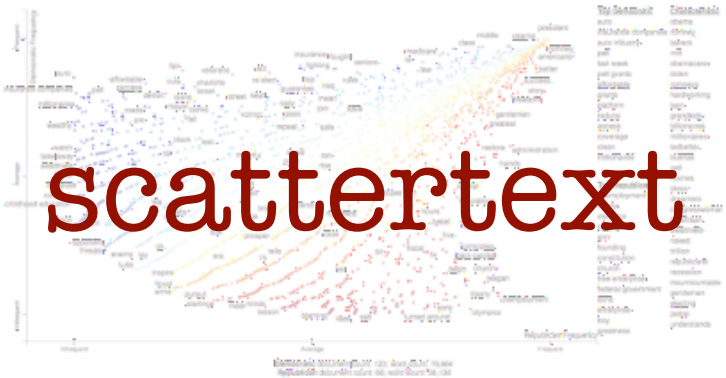


Sentiment Analysis

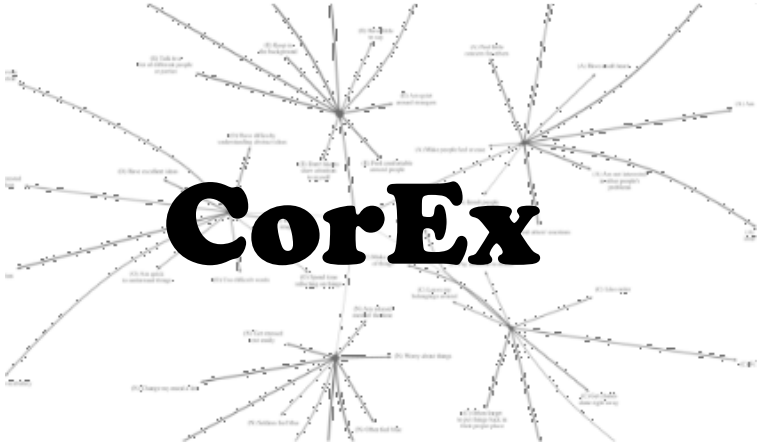
Topic Modeling

Introduction

spaCy

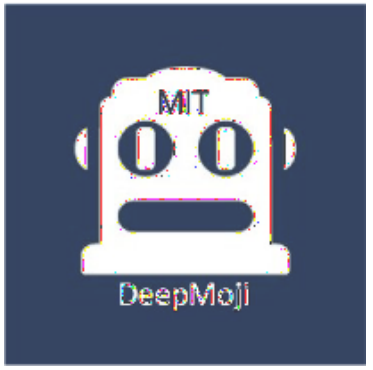


scattertext



CorEx

DeepMoji



Dataset

Fine Food Reviews

350K+ Unique Reviews

1999 – 2012

Review Text	Product ID
Rating	Helpfulness

https://www.kaggle.com/snap/amazon-fine-food-reviews

Dataset Products

amazon

spaCy

What is spaCy?

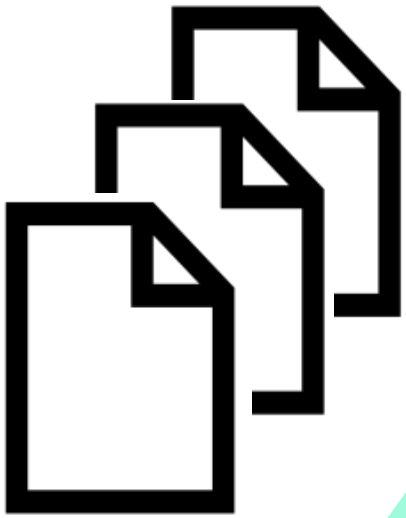
Effective, efficient text processing

Help small businesses leverage NLP

Alternative to NLTK

Highly customizable

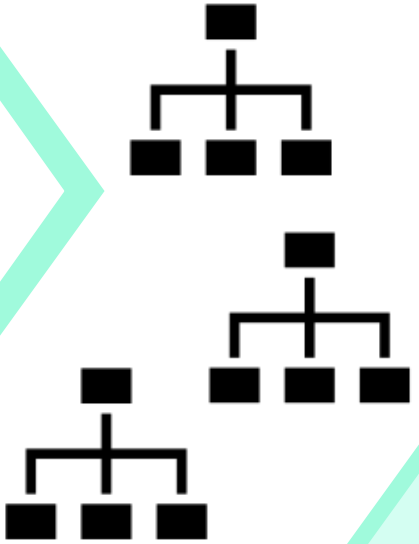
Raw text documents



Pipeline with language model

- ✓ Tokenization
- ✓ Parts of speech
- ✓ Lemmatization
- ✓ Stop words
- ✓ Named entities
- ✓ Custom

Parsed tokens, docs



Language Model

“ I’m so happy I went to this awesome Vegas buffet! ”

Language Model

"I'm so happy I went to this awesome Vegas buffet!"

I
'm
so
happy
I
went
to
this
awesome
Vegas
buffet
!

so
POS - ADV
LEMMA - so

went
POS - VERB
LEMMA - go
IS STOP? - False

Vegas
POS - PROPN
LEMMA - Vegas
IS STOP? - False
NER? - GPE

Syntactic Dependencies

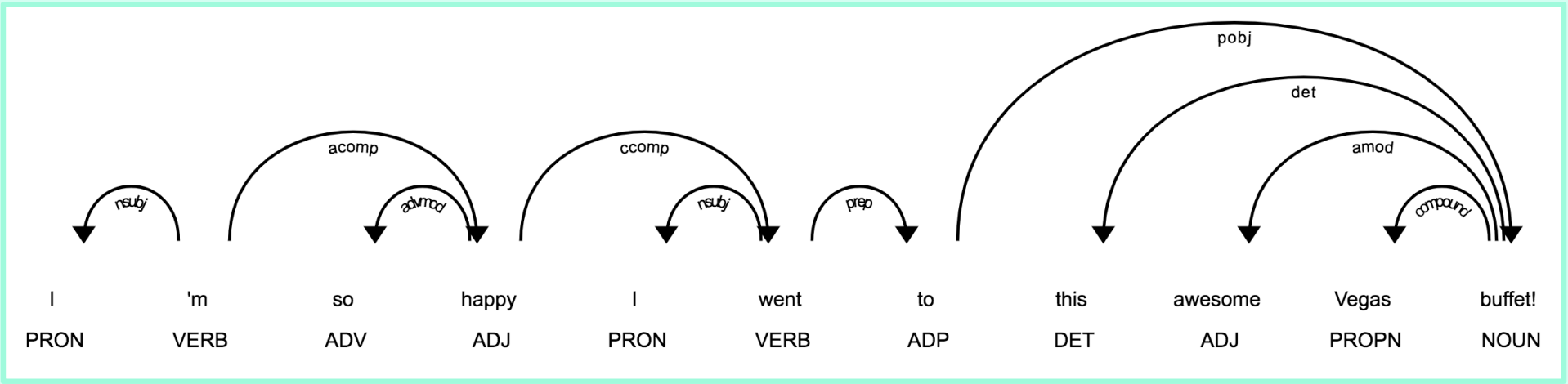
"I'm so happy I went to this awesome Vegas buffet!"

I	nsubj
'm	ROOT
so	advmod
happy	acomp
I	nsubj
went	ccomp
to	prep
this	det
awesome	Vegas
buffet	pobj
!	punc

Syntactic Dependencies

"I'm so happy I went to this awesome Vegas buffet!"

displaCy



Adjectives

spaCy + Data

great

best

free

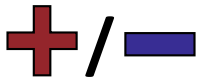
sweet

delicious

"… I also like that it is salt **free**…"

"… totally natural yet calorie **free** sweetness…"

amazon



Adjectives

spaCy + Data

great

best

free

sweet

delicious

"… desperately needed some **more** chocolate chips! …"

"… thought this would have **more** health benefits than…"

"… I love the **old** original yogi tea…"

more

bad

same, first

old, different

disappointed

Adjectives

spaCy + Data

great

best

free

sweet

delicious

BOTH

good

other

little

better

many

"… and is not as **good** …"

ıore

ɔad

ıe, first

old, different

disappointed

Product Adjectives

spaCy + Data

instant

good

great

fresh

best

hot

strong

other

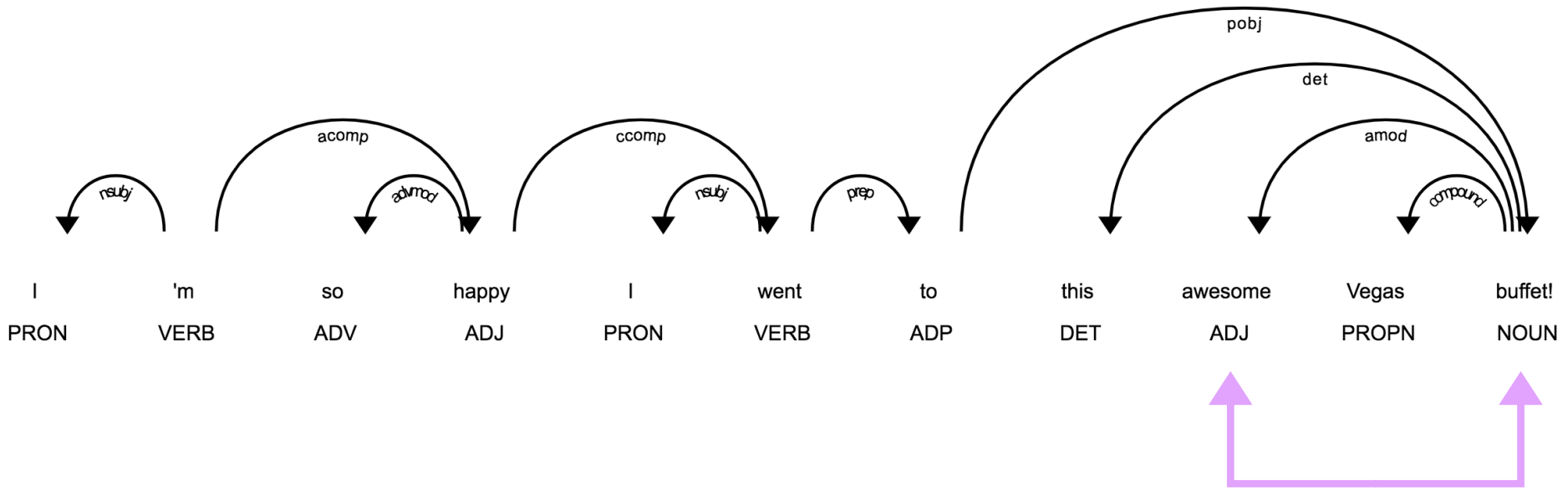
easy

little

convenient

Adjective Modifiers

spaCy + Data



“awesome buffet”

Adjective Modifiers

spaCy + Data

“coffee”

+	BOTH	−
great	good	instant
best	flavored	worst
iced	strong	weak
dark, bold		cheap, bitter

Adjective Modifiers

spaCy + Data

“delivery”

+	−
fast	next
quick	timely
automatic	prompt


scattertext

What is scattertext?

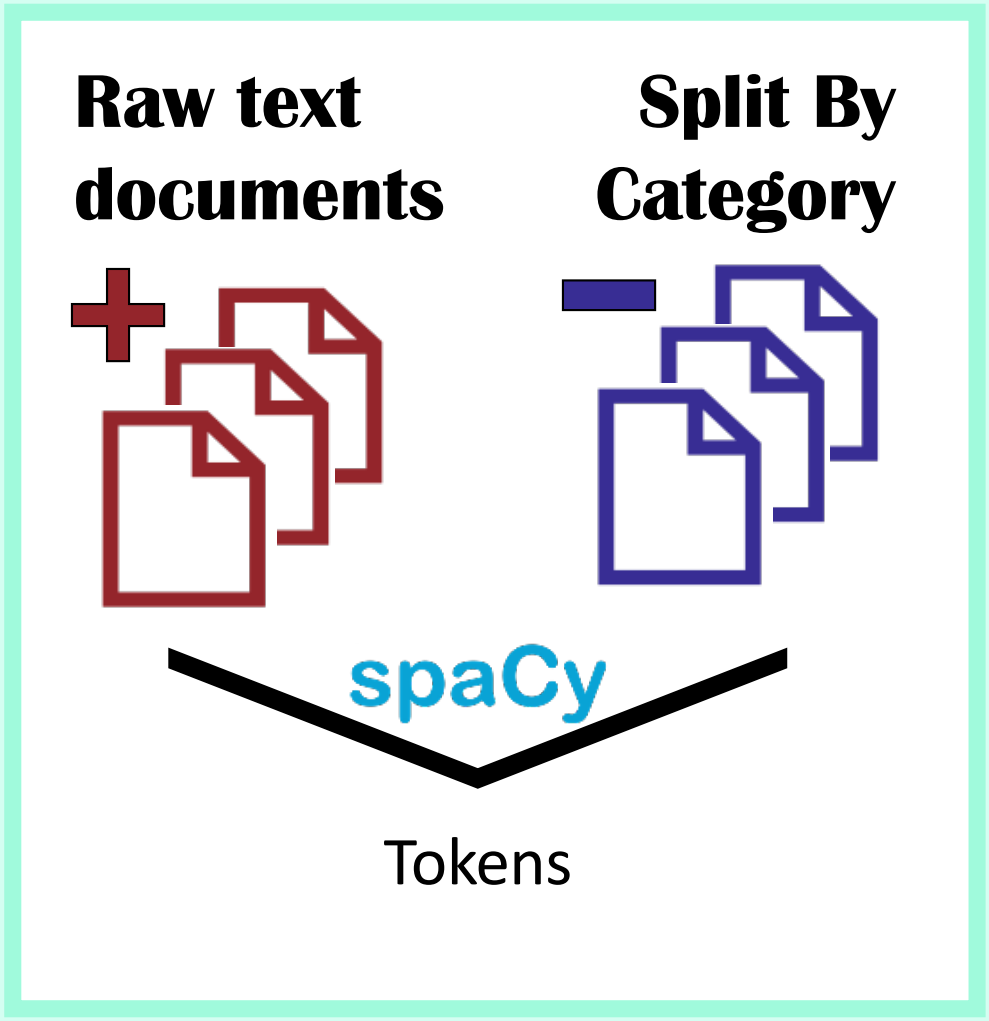
Visualization tool, mid-sized corpora

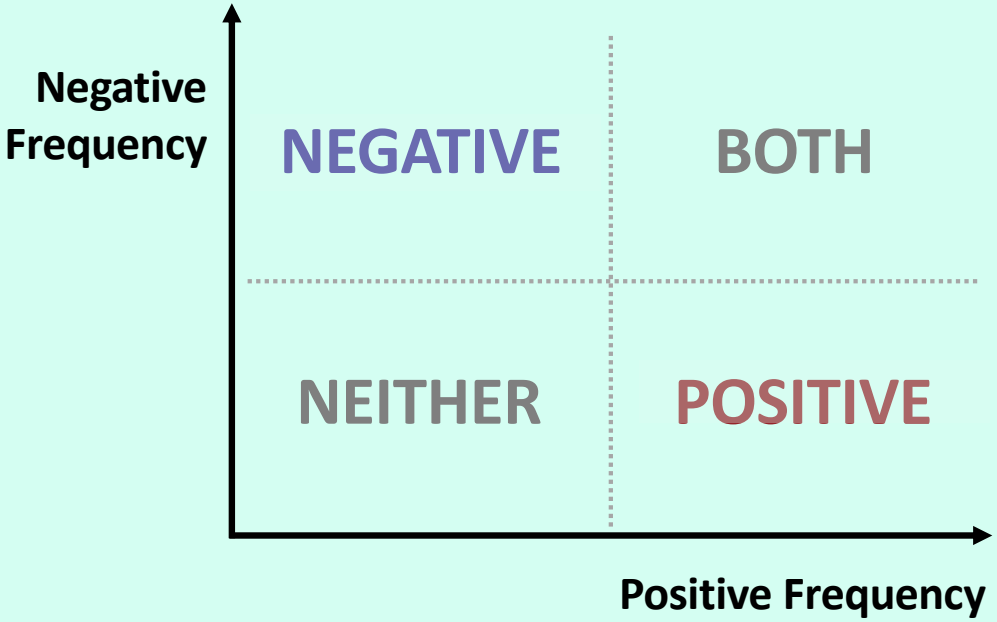
Interactive scatter plots

Terms more characteristic of category

Leverages spaCy language model

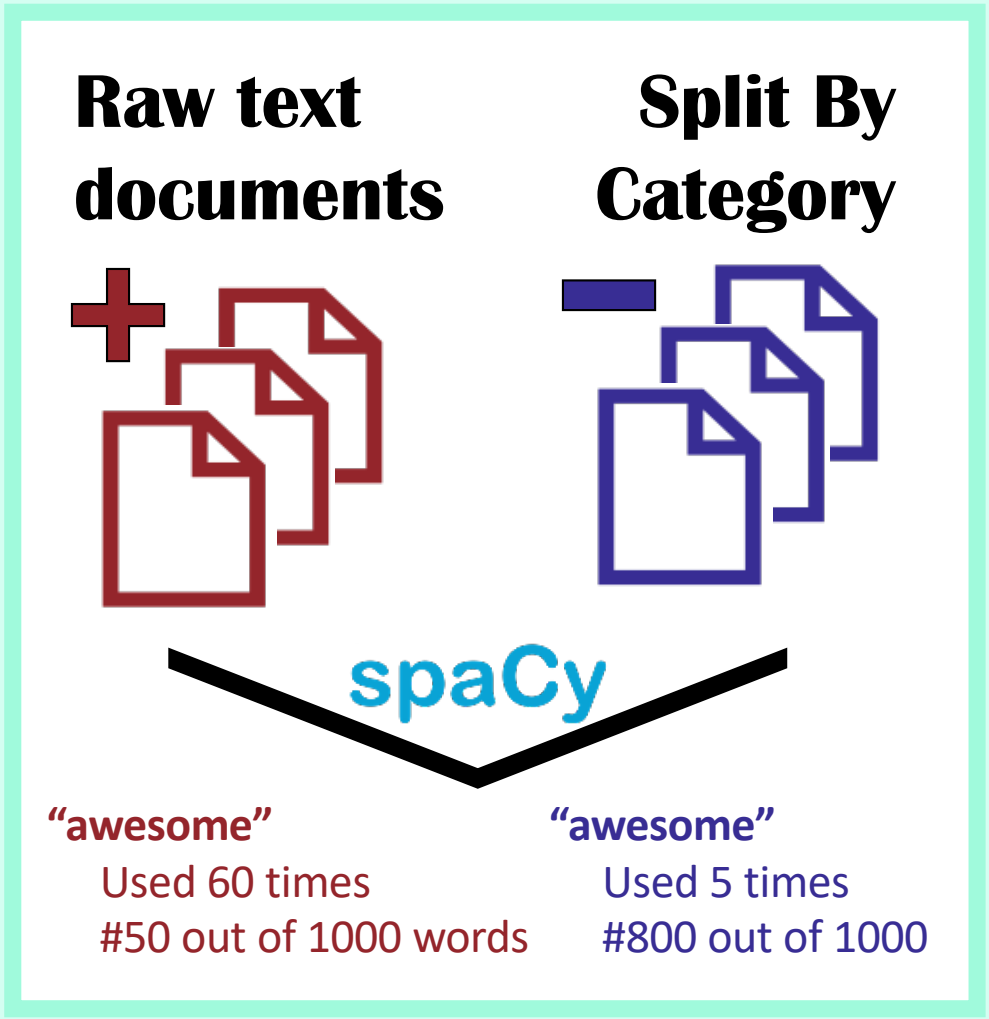
Jason S. Kessler. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ.
ACL System Demonstrations. 2017.

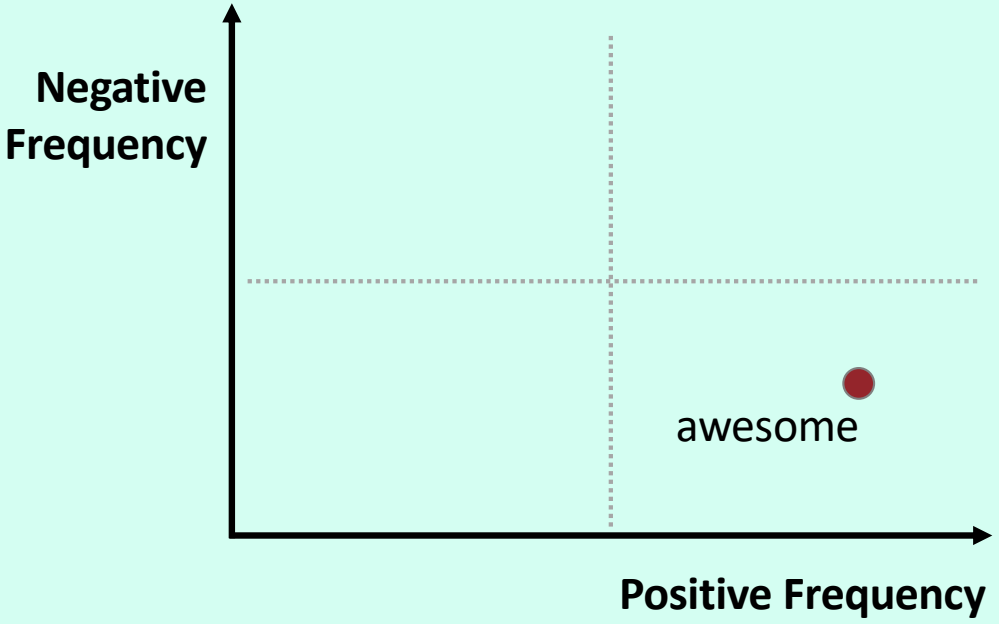




Token scatter point

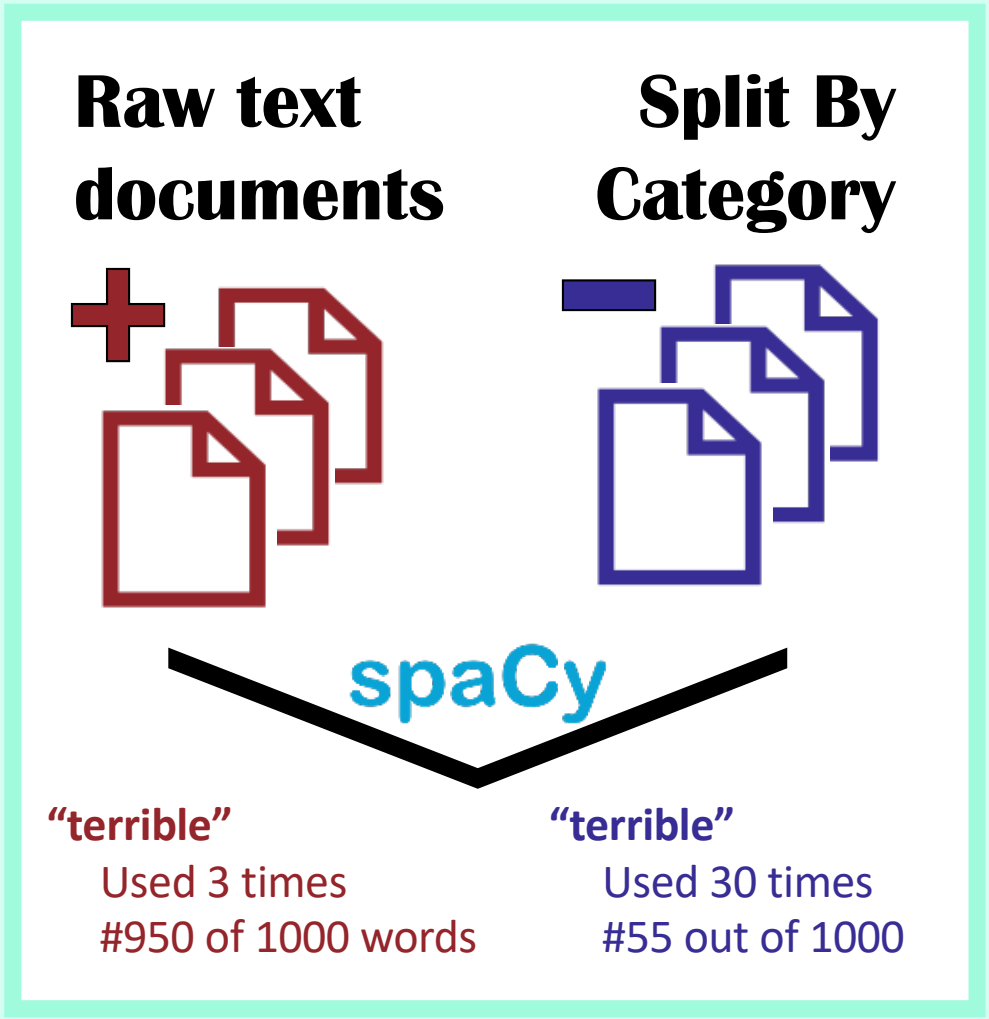
- Position by frequency rank
- Color by scaled f-score

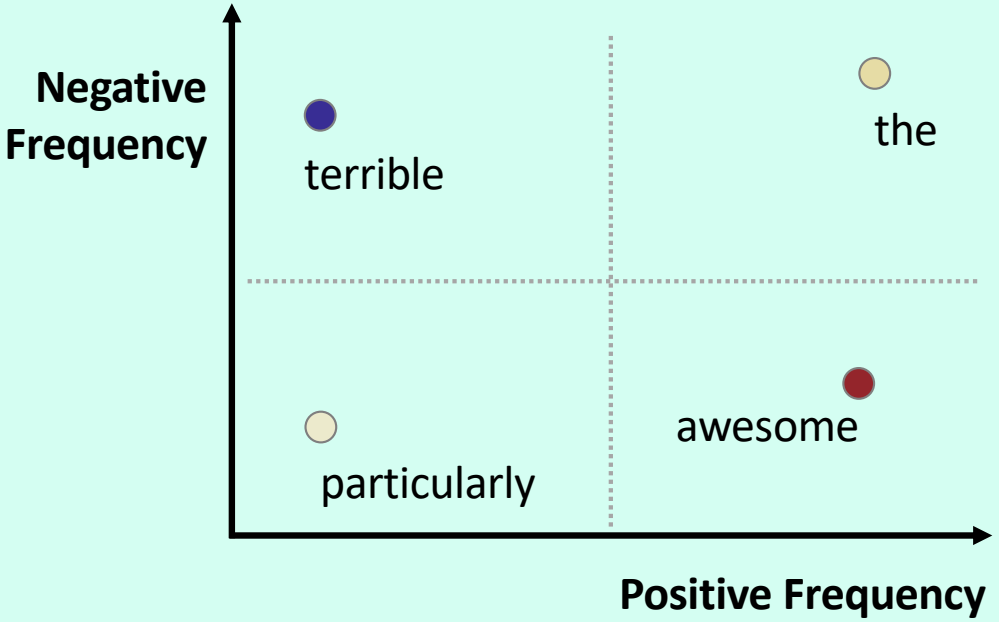




Token scatter point

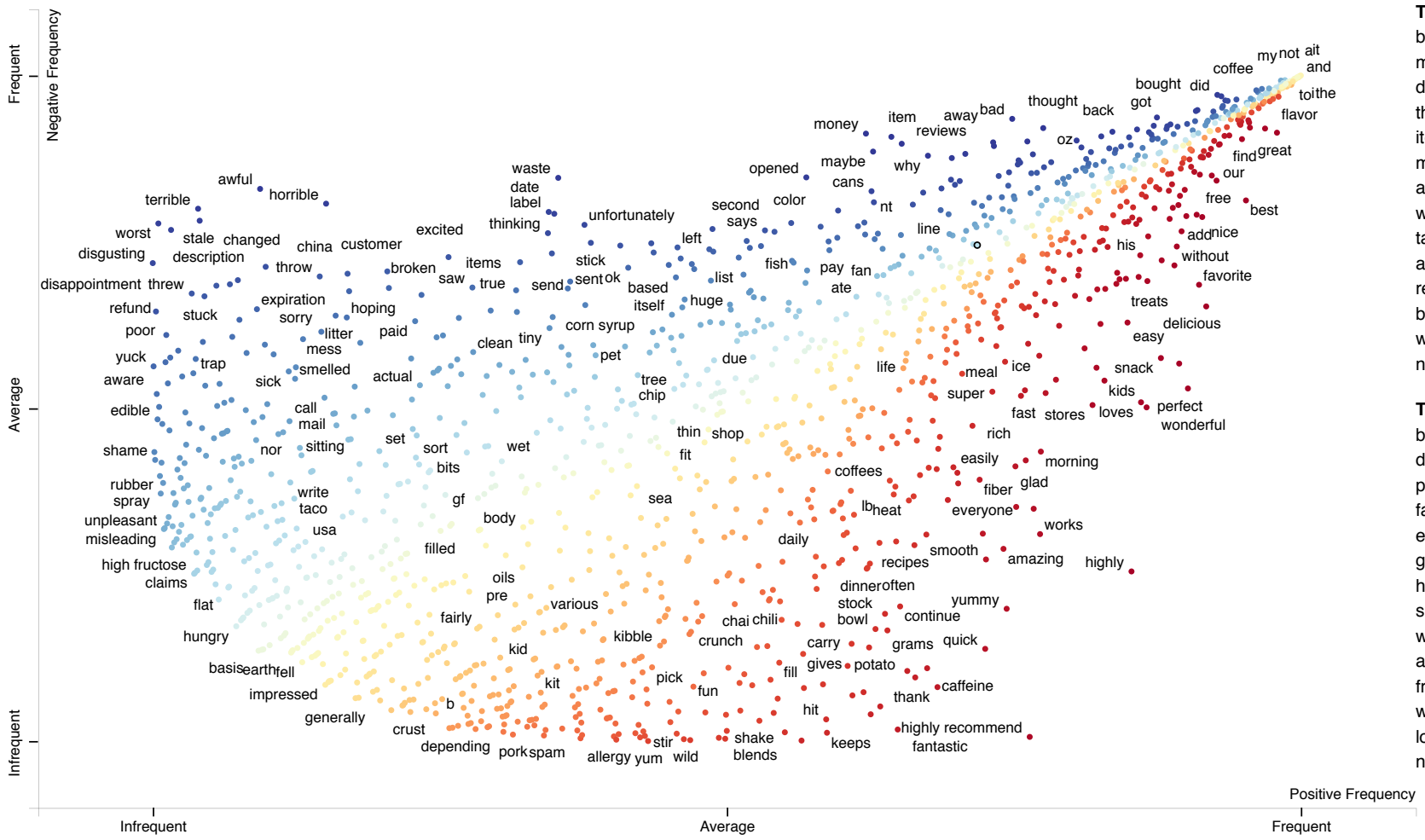
- Position by frequency rank
- Color by scaled f-score





Token scatter point

- Position by frequency rank
- Color by scaled f-score



Top Negative	Characteristic
bad	keurig
money	flavor
disappointed	aftertaste
thought	tasted
item	tastes
maybe	chewy
away	gluten
why	senseo
tasted	taste
awful	kibble
received	granola
box	decaf
waste	sweetener
nothing	stevia
	salty
Top Positive	dented
best	flavors
delicious	flavoring
perfect	flavorful
favorite	stale
easy	jerky
great	coffees
highly	wo
snack	flavored
wonderful	picky
add	oatmeal
free	crunchy
without	splenda
loves	sweetened
nice	cereal

Negative document count: 2,000; word count: 170,758
Positive document count: 2,000; word count: 155,174

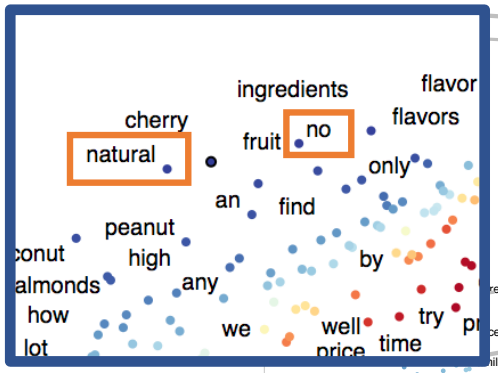
Product Comparisons

scattertext + Data

VS.

Product Comparisons

scattertext + Data

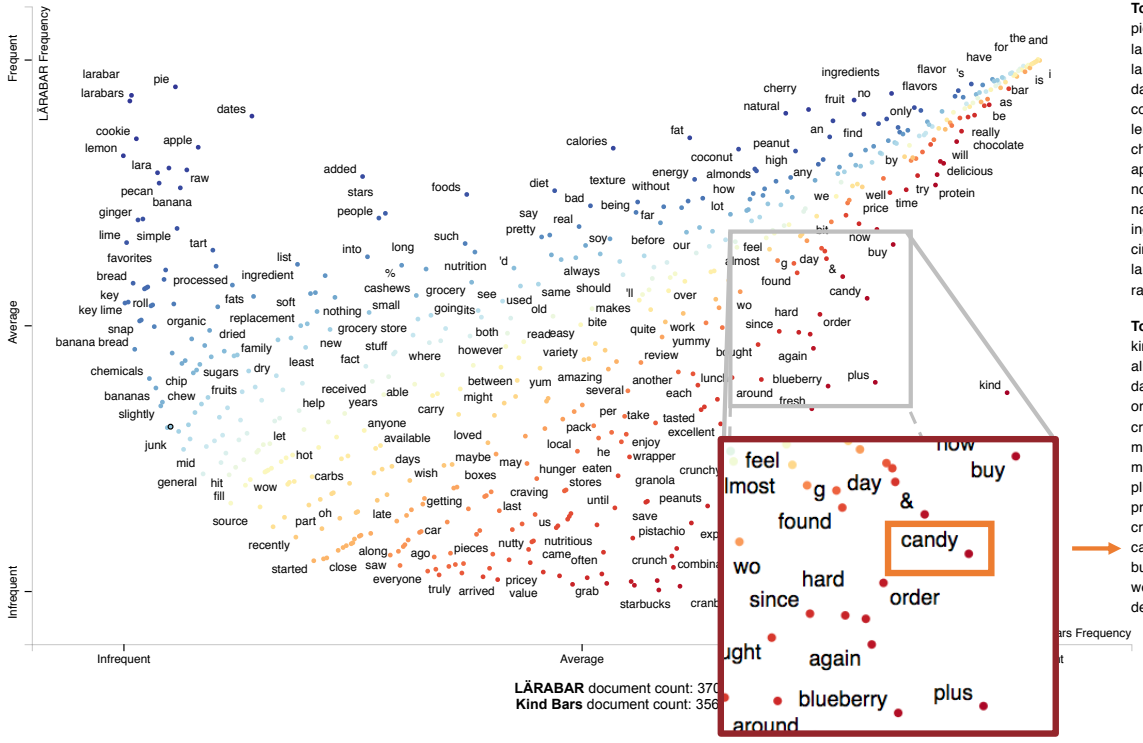


LÄRABAR document count: 370; word count: 29,937
Kind Bars document count: 356; word count: 21,939

Top LÄRABAR	Characteristic
pie	larabars
larabar	larabar
larabars	cashew
dates	cashews
cookie	gluten
lemon	macadamia
cherry	chewy
apple	unsweetened
no	snack
natural	pistachio
ingredients	almonds
cinnamon	flavors
lara	aftertaste
raw	granola
	flavor
Top Kind Bars	pecan
kind	crunchy
almond	pomegranate
dark	paleo
ordered	unprocessed
cranberry	preservatives
macadamia	almond
mango	blueberry
plus	glycemic
protein	bars
cranberry almond	peanut
candy	tasty
buy	nutty
were	tastes
delicious	cranberries

Product Comparisons

scattertext + Data



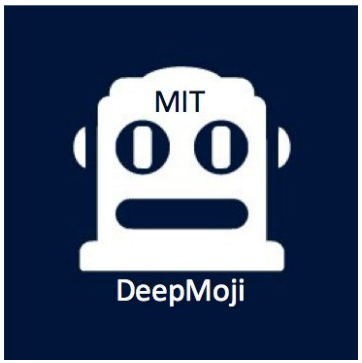
Top LÄRABAR	Characteristic
pie	larabars
larabar	larabar
larabars	cashew
dates	cashews
cookie	gluten
lemon	macadamia
cherry	chewy
apple	unsweetened
no	snack
natural	pistachio
ingredients	almonds
cinnamon	flavors
lara	aftertaste
raw	granola
	flavor
Top Kind Bars	pecan
kind	crunchy
almond	pomegranate
dark	paleo
ordered	unprocessed
cranberry	preservatives
macadamia	almond
mango	blueberry
plus	glycemic
protein	bars
cranberry almond	peanut
candy	tasty
buy	nutty
were	tastes
delicious	cranberries


DeepMoji


DeepMoji
Artificial emotional intelligence
→ Teach AI!
DeepMoji has learned to understand emotions and sarcasm based on millions of emojis. Here's a video explaining a bit more. Type a sentence to see what our AI algorithm thinks.
Type something..
SUBMIT
Examples
Click on one!
You love hurting me, huh?
This is the s***!
My flight is delayed.. amazing.
I know good movies, this ain't one
It was fun, but I'm not going to miss you
What is happening to me??

What is DeepMoji?

Released
2017



Deep learning emotional classifier

Emoji predictions with confidence

Trained on 1.2 billion tweets

Examples

DeepMoji + Data

+

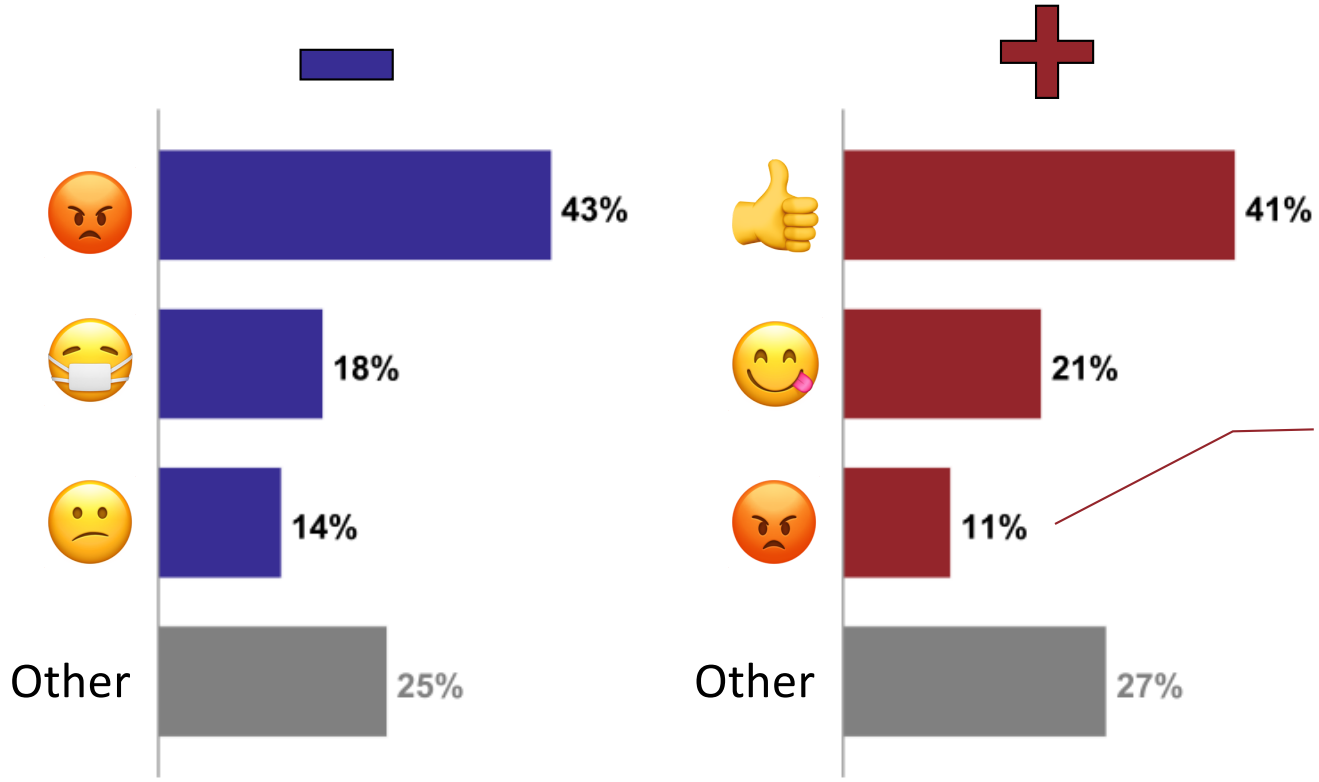
"These are, without serious rival, my favorite chips ever. They are so yummy!"


−

"Old expired product - past the expiry date by 2 years!! Disgusting!!!!!!..."


Dominate Emotions

DeepMoji + Data



Rating: 5/5

"… Unfortunately, they arrived in a solid mass of melted chocolate… I won't order them online again, but if I see them in a store, I would pick them up.

CorEx

What is CorEx?

Correlation **ex**planation, topic modeling

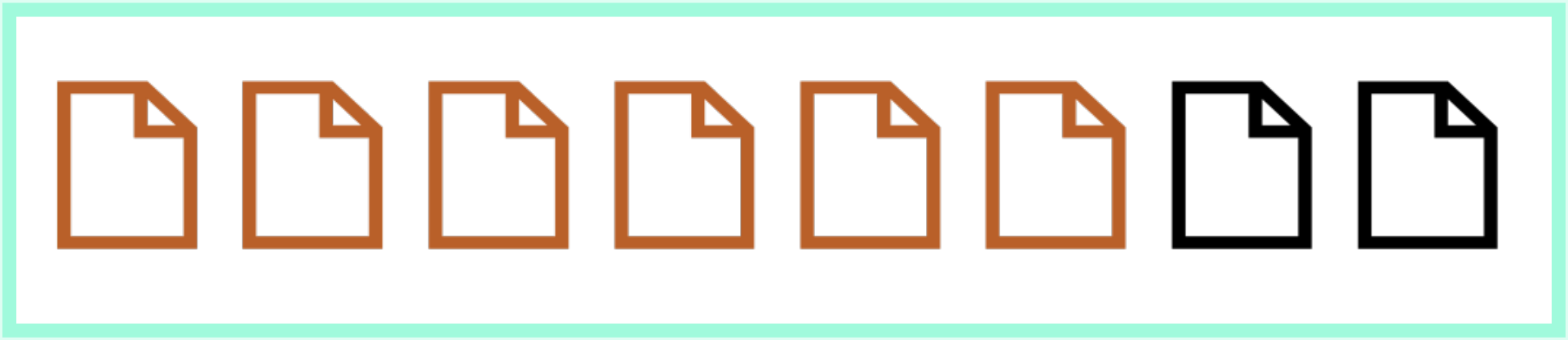
Optional anchor words

Optional hierarchical modeling

Total correlation → number of topics

Released 2017

Gallagher, Ryan J., Kyle Reing, David Kale, and Greg Ver Steeg. "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Transactions of the Association for Computational Linguistics (TACL), 2017.



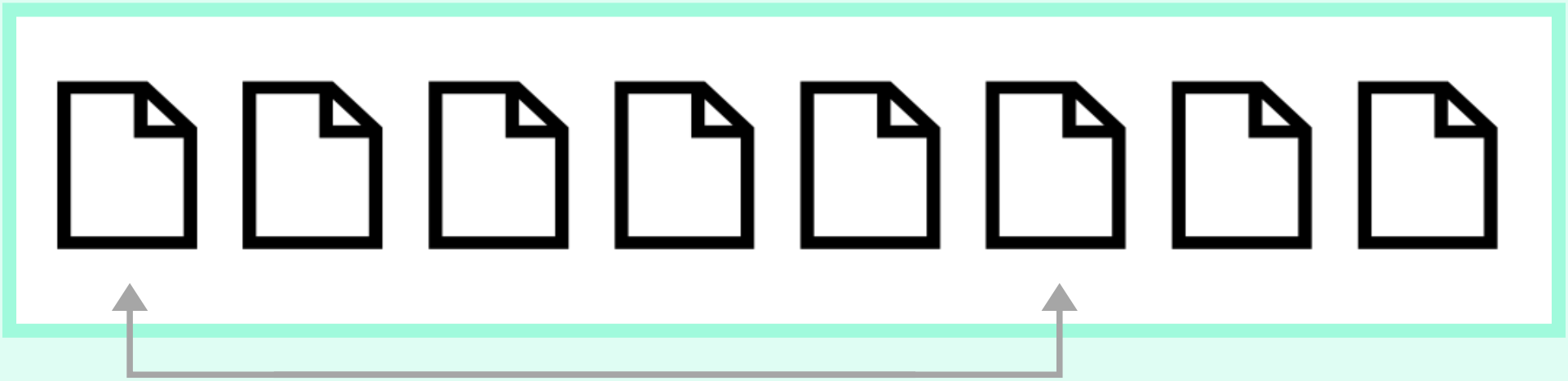
“flavor”

6

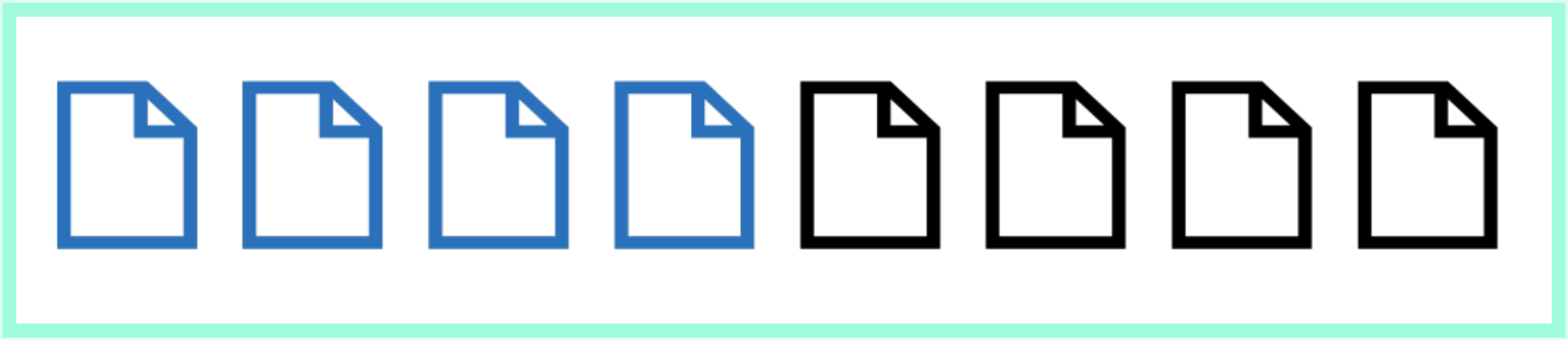
“flavor” 6

“taste” 6

“flavor”	“taste”	“smell”
6	6	4



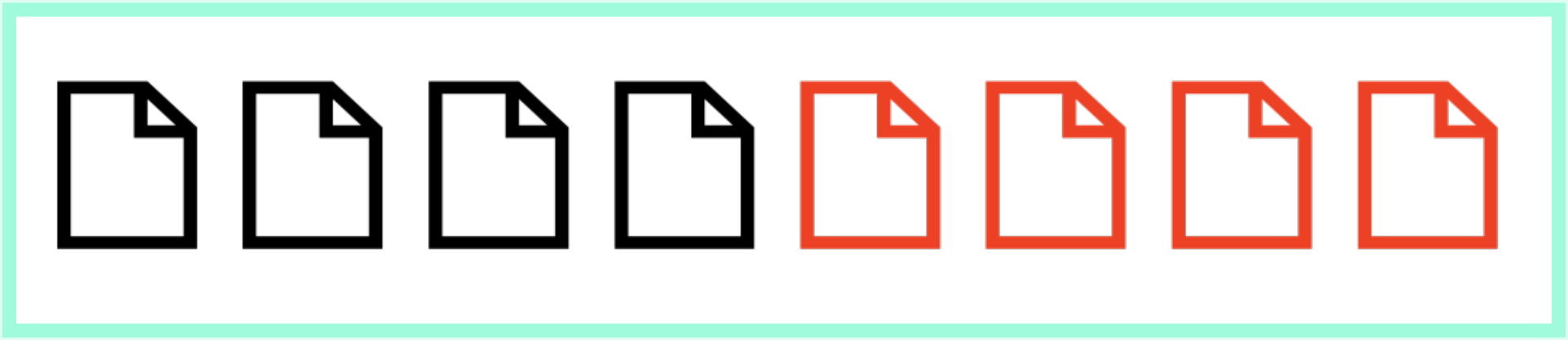
"flavor"		"taste"		"smell"		Group		Score
6	+	6	+	4	-	6	=	10



“flavor”		“taste”		“smell”		Group		Score
6	+	6	+	4	-	6	=	10

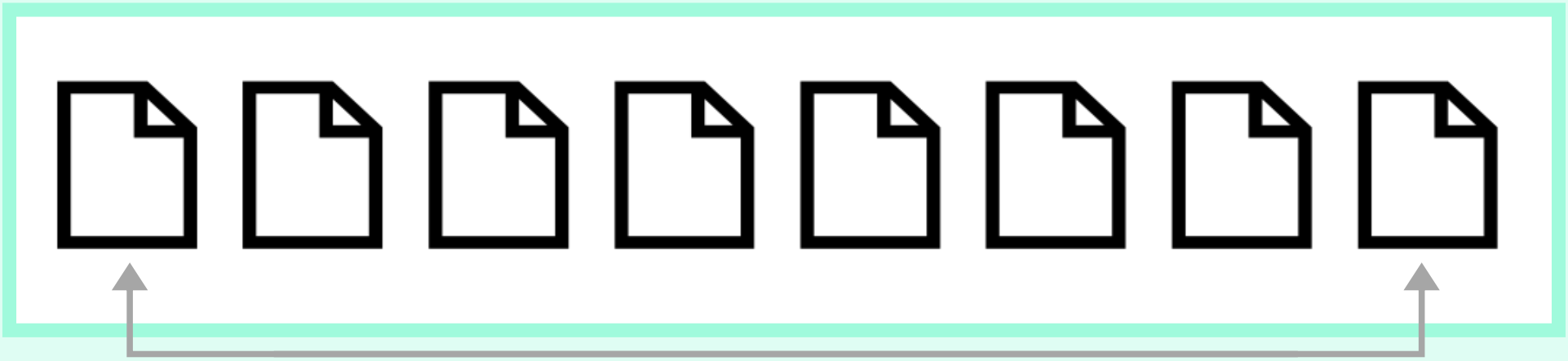
“healthy”

4



“flavor”		“taste”		“smell”		Group		Score
6	+	6	+	4	-	6	=	10

“healthy”	“delivery”
4	4



“flavor”		“taste”		“smell”		Group		Score
6	+	6	+	4	-	6	=	10

“healthy”		“delivery”		Group		
4	+	4	-	8	=	0

Topics

CorEx + Data

Coffee & Tea

coffee, tea, cup, drink, teas, strong

Pet Supplies

dog, food, dogs, treats, cat, vet

Macronutrients

calories, snack, fat, sugar, protein

Entrées

sauce, chicken, soup, meat, pasta

Self Improvement Products

day, energy, hair, help, started, weight

Other Reviews

reviews, did, product, review, read

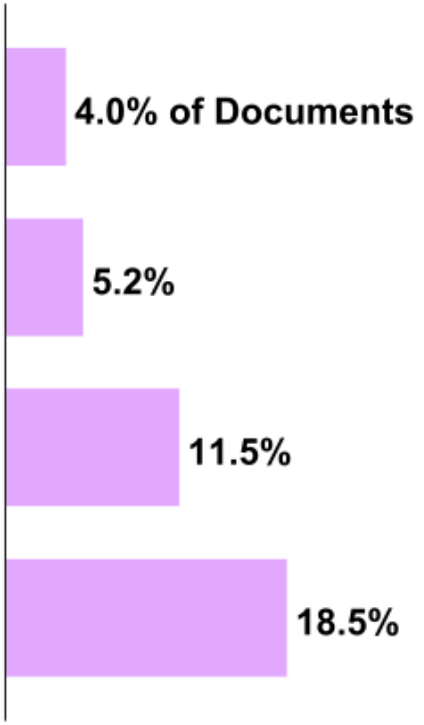
Baking

oil, add, use, mix, make, bread, recipe

Anchored Topics

CorEx + Data

shipping		prime, fast, cost, charges, saver
package, delivery		arrived, prompt, automatic, speedy
price		reasonable, pay, beat, great, competitive
amazon, store		grocery, local, buy, cheaper, available

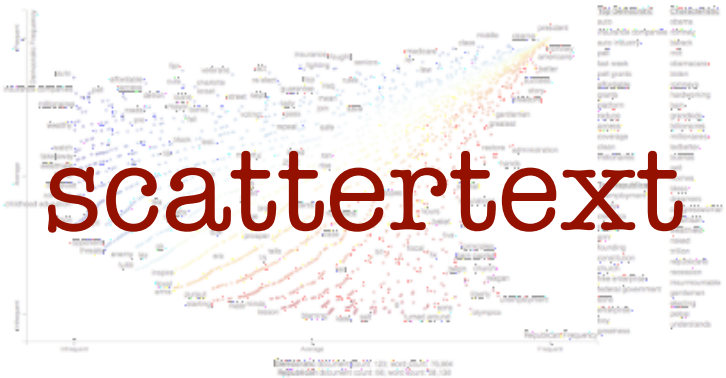


4 additional “free” topics also used

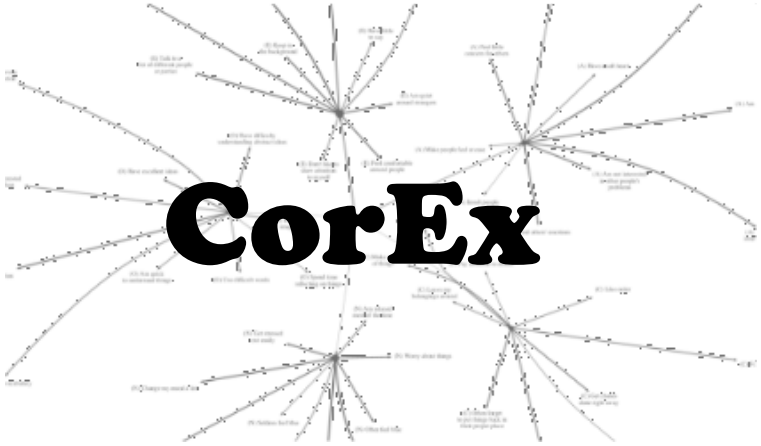
Conclusion

Conclusion

spaCy

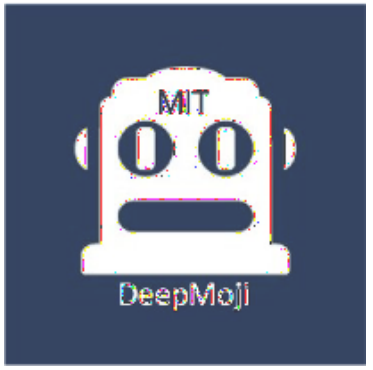


scattertext



CorEx

DeepMoji




QUESTIONS?
Kimberly Fessel, PhD • @kimberlyfessel • kimberlyfessel
#ODSC

APPENDIX

Findings

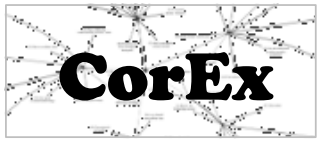
spaCy

- Manage customer expectations
- Product updates risky
- Frequent product comparisons

scattertext

- Flavor top priority
- Direct product comparisons

CorEx



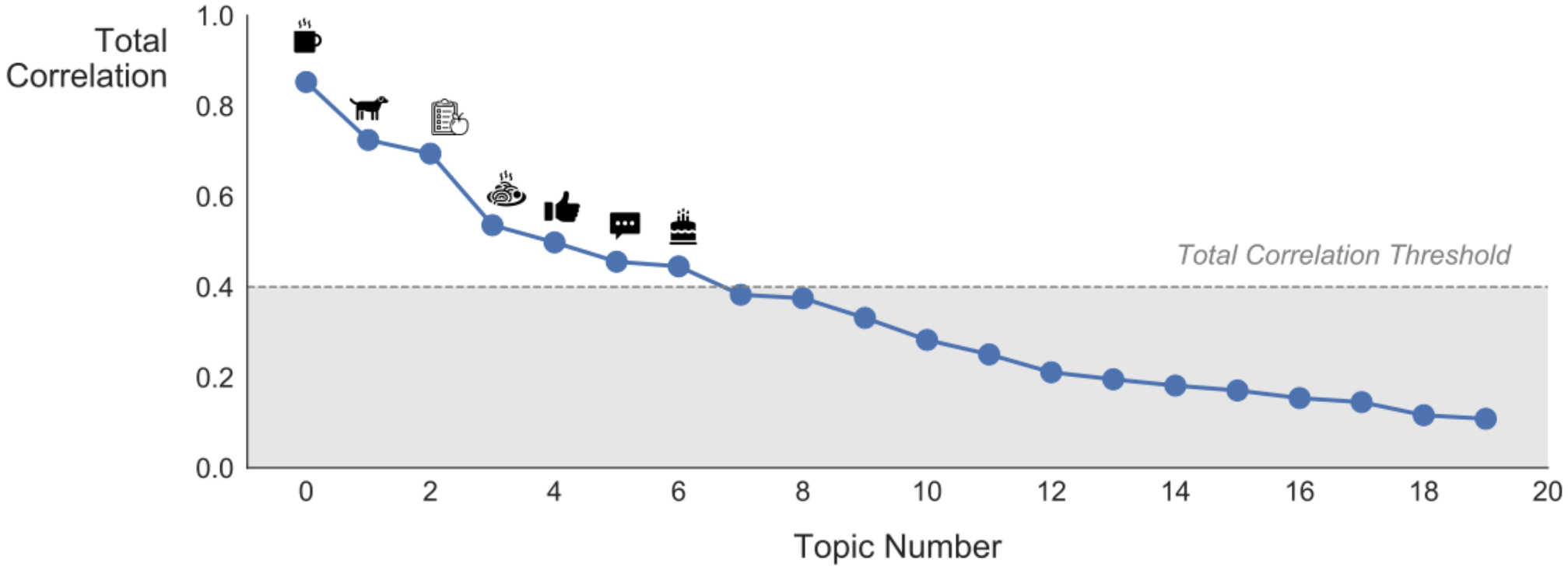
- Product types dominate topics
- Anchored approach for others

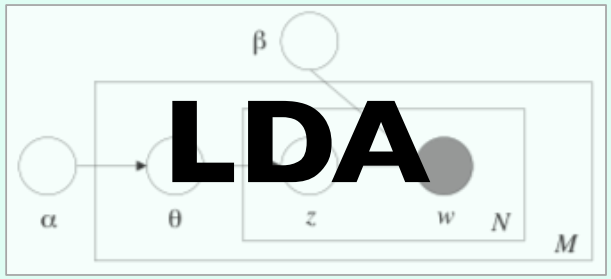
DeepMoji

- Predominate emotions
  - Anger or approval
  - Driven by taste

Topics

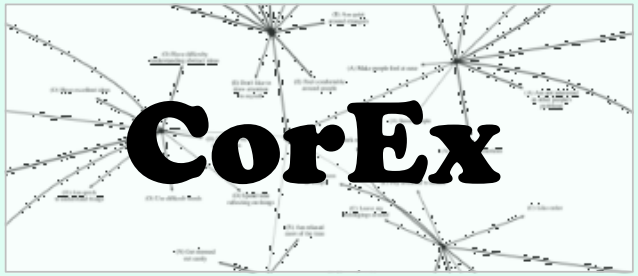
CorEx + Data





Generative model

- From Y (docs/words) → Find X (topics)
- Dirichlet distributions
- P(word|topic), P(topic|doc)



Discriminative model

- From X (docs/words) → Form Y (topics)
- Groups of words, high total correlation
- Topic word set: broad and overlapping

spaCy

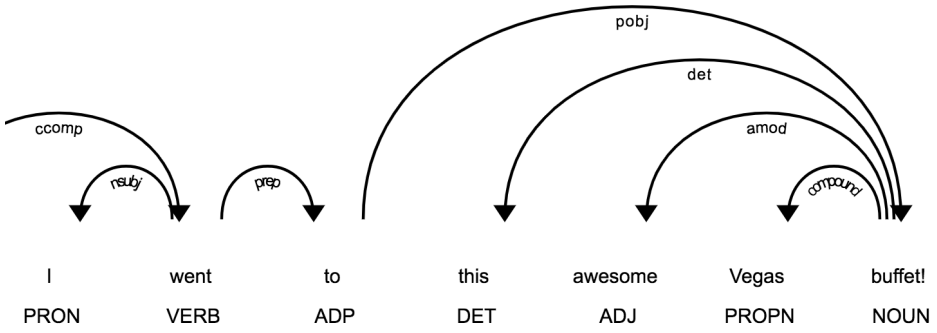
I
‘m
so
happy
I
went
to
this
awesome
Vegas
buffet
!

so

POS - ADV
LEMMA - so
IS STOP? - True
NER? - False

Text pre-processing

Syntactic dependencies



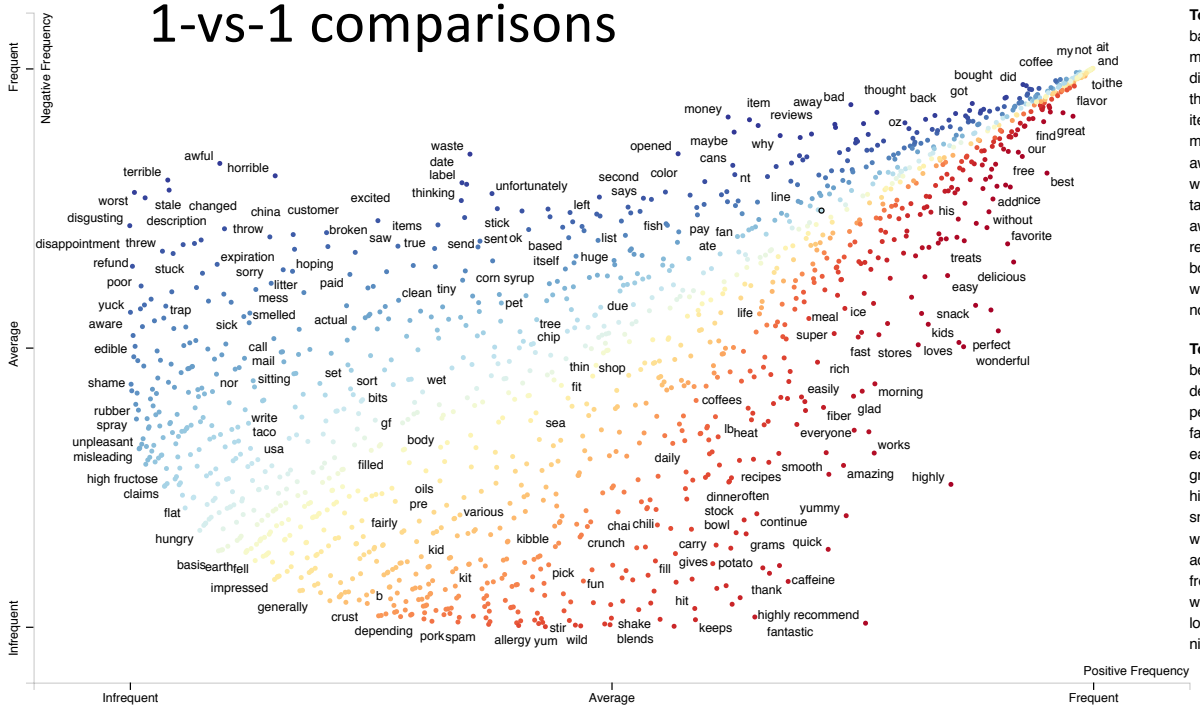
Extensions

- Token/document similarity
- Token/phrase matcher
- Text classification

scattertext

Interactive, visualization tool for text EDA

1-vs-1 comparisons



Top Negative	Characteristic
bad	keurig
money	flavor
disappointed	aftertaste
thought	tasted
item	tastes
maybe	chewy
away	gluten
why	senseo
tasted	taste
awful	kibble
received	granola
box	decaf
waste	sweetener
nothing	stevia
	salty
Top Positive	dented
best	flavors
delicious	flavoring
perfect	flavorful
favorite	stale
easy	jerky
great	coffees
highly	wo
snack	flavored
wonderful	picky
add	oatmeal
free	crunchy
without	splenda
loves	sweetened
nice	cereal

Negative document count: 2,000; word count: 170,758
Positive document count: 2,000; word count: 155,174

Extensions

- Two categories per document
- Empath topics
- Word similarities, word2vec
- Emoji EDA

DeepMoji

Deep learning emoji classifier

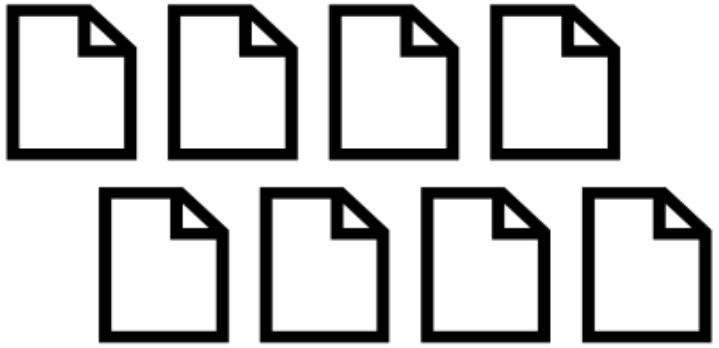
> *"We really enjoyed this product – contains no sugar (contains dates and such) and it's healthy while giving you the chocolate fix you need!"*

Extensions

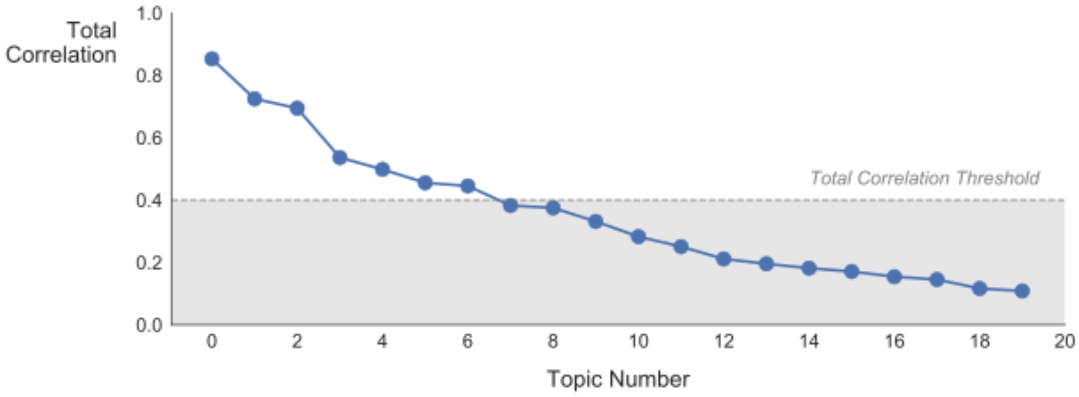
- Label texts with granular emotion
- Encoding vectors
  - → Document emotional similarity
- Transfer learning

CorEx

Topic modeling via total correlation



Choose number of topics



Extensions

- Anchor words
- Hierarchical topic modeling

References

- J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013. (Retrieved from https://www.kaggle.com/snap/amazon-fine-food-reviews, 2020.)

- Kessler, Jason S. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. ACL System Demonstrations, 2017. (preprint: https://arxiv.org/abs/1703.00565)

- Felbo, Bjarke, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017. (https://arxiv.org/pdf/1708.00524.pdf)

- Gallagher, Ryan J., Kyle Reing, David Kale, and Greg Ver Steeg. "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Transactions of the Association for Computational Linguistics (TACL), 2017. (https://transacl.org/ojs/index.php/tacl/article/view/1244)