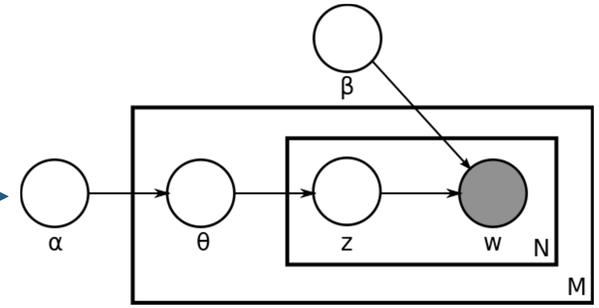
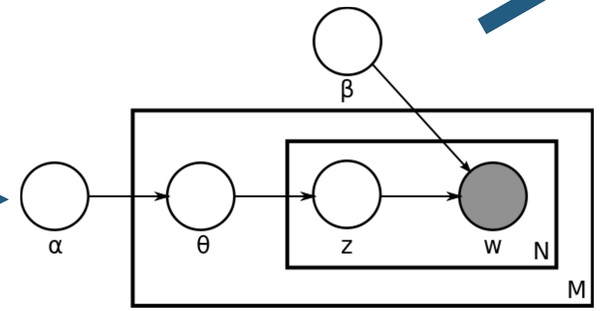
Visualizing Topic Models

Ben Mabey
@bmabey

DATA SCIENCE SUMMIT &
DATO CONFERENCE 2015



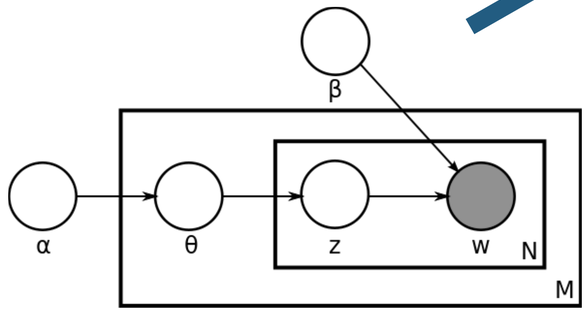
Latent Dirichlet Allocation
(LDA)



Latent Dirichlet Allocation (LDA)

Document-Topic Distributions

	0	1	…	k
doc a	0.25	0.14	…	0.02
doc b	0.01	0.30	…	0.09
…	…	…	…	0.31
doc D	0.13	0.07	…	0.01



Latent Dirichlet Allocation (LDA)

Document-Topic Distributions

	0	1	…	k
doc a	0.25	0.14	…	0.02
doc b	0.01	0.30	…	0.09
…	…	…	…	0.31
doc D	0.13	0.07	…	0.01

Term-Topic Distributions

	0	1	…	k
bird	0.002	0.01	…	0.004
coffee	0.001	0.003	…	0.009
…	…	…	…	0.031
work	0.002	0.006	…	0.021


Y
Hacker News

250k+ stories
July 2007 - May 2014

250k+ stories
July 2007 - May 2014

POS tagging w/spaCy

250k+ stories
July 2007 - May 2014

POS tagging w/spaCy

Phrase detection w/Gensim

250k+ stories
July 2007 - May 2014

POS tagging w/spaCy

Phrase detection w/Gensim

Stopword removal &
only kept nouns or phrases with nouns

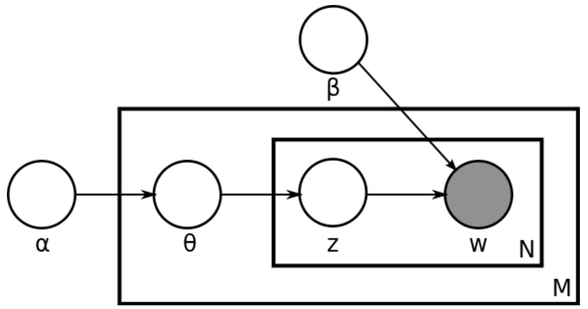
Hacker News

250k+ stories
July 2007 - May 2014

POS tagging w/spaCy

Phrase detection w/Gensim

Stopword removal &
only kept nouns or phrases with nouns



Fit LDA models varying the number of topics

Game written by 14 year old passes Angry Birds as the top free iphone app

Game written by 14 year old passes Angry Birds as the top free iphone app

Document-Topic Distribution

Topic	*P(T | D)*
58	*0.19*
38	*0.14*
16	*0.06*
…	…

Game written by 14 year old passes Angry Birds as the top free iphone app

Document-Topic Distribution

Topic	*P(T|D)*
58	*0.19*
38	*0.14*
16	*0.06*
…	…

Sorted Topic-Term Distributions

58	*38*	*16*
app	*game*	*language*
developer	*player*	*code*
mobile	*video game*	*programming*
user	*gaming*	*java*
app store	*developer*	*programmer*

Hacker News

Game written by 14 year old passes Angry Birds as the top free iphone app

Document-Topic Distribution

Topic	P(T|D)
mobile apps	0.19
38	0.14
16	0.06
…	…

Sorted Topic-Term Distributions

mobile apps	38	16
app	game	language
developer	player	code
mobile	video game	programming
user	gaming	java
app store	developer	programmer

Hacker News

Game written by 14 year old passes Angry Birds as the top free iphone app

Document-Topic Distribution

Topic	P(T|D)
mobile apps	0.19
video games	0.14
16	0.06
…	…

Sorted Topic-Term Distributions

mobile apps	video games	16
app	game	language
developer	player	code
mobile	video game	programming
user	gaming	java
app store	developer	programmer

Hacker News

Game written by 14 year old passes Angry Birds as the top free iphone app

Document-Topic Distribution

Topic	P(T|D)
mobile apps	0.19
video games	0.14
programming	0.06
…	…

Sorted Topic-Term Distributions

mobile apps	video games	programming
app	game	language
developer	player	code
mobile	video game	programming
user	gaming	java
app store	developer	programmer

Interpreting Topic Models

What is the meaning of each topic?

Interpreting Topic Models

What is the meaning of each topic?

How prevalent is each topic?

Interpreting Topic Models

What is the meaning of each topic?

How prevalent is each topic?

How do the topics relate to each other?

Interpreting Topic Models

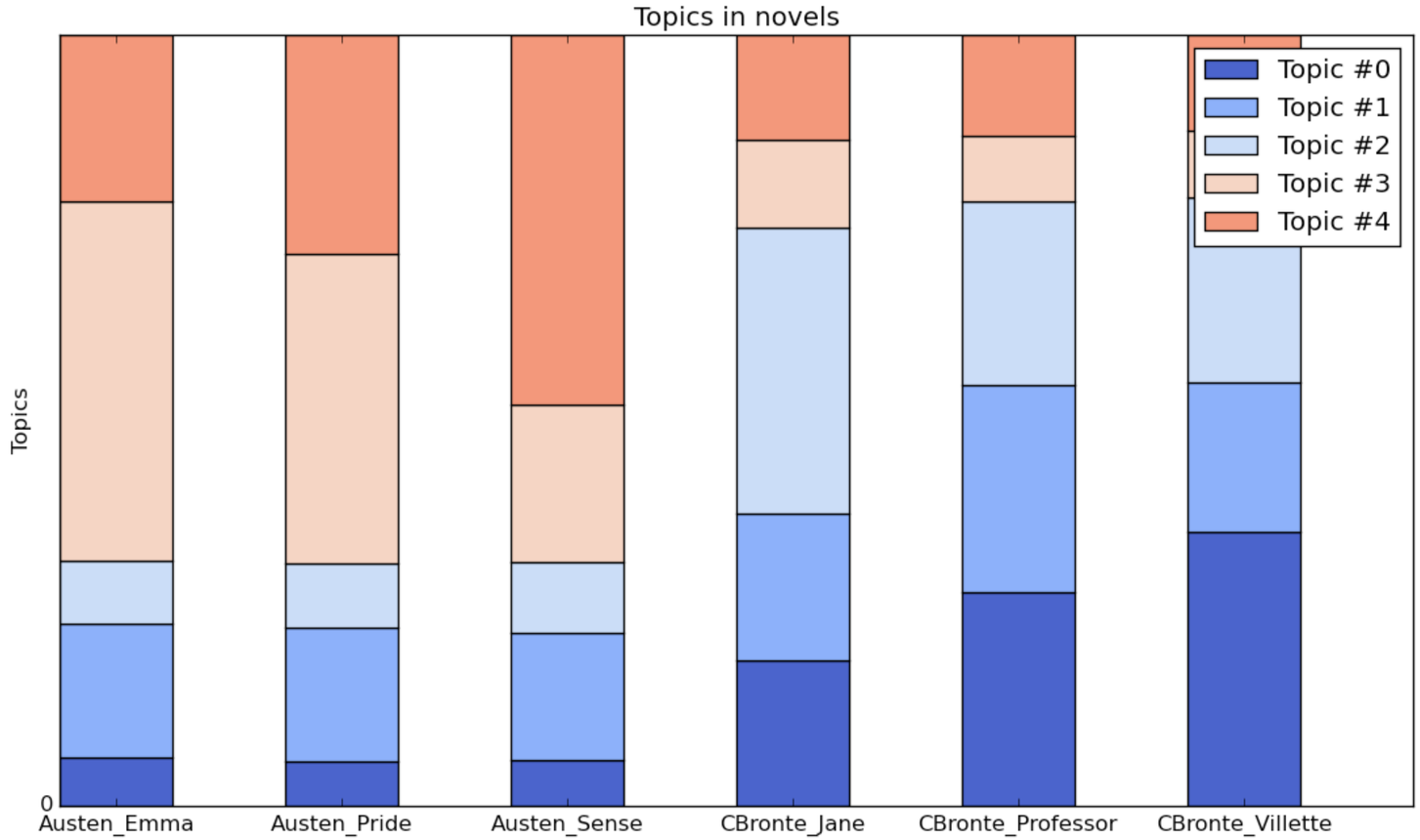
What is the meaning of each topic?

How prevalent is each topic?

How do the topics relate to each other?

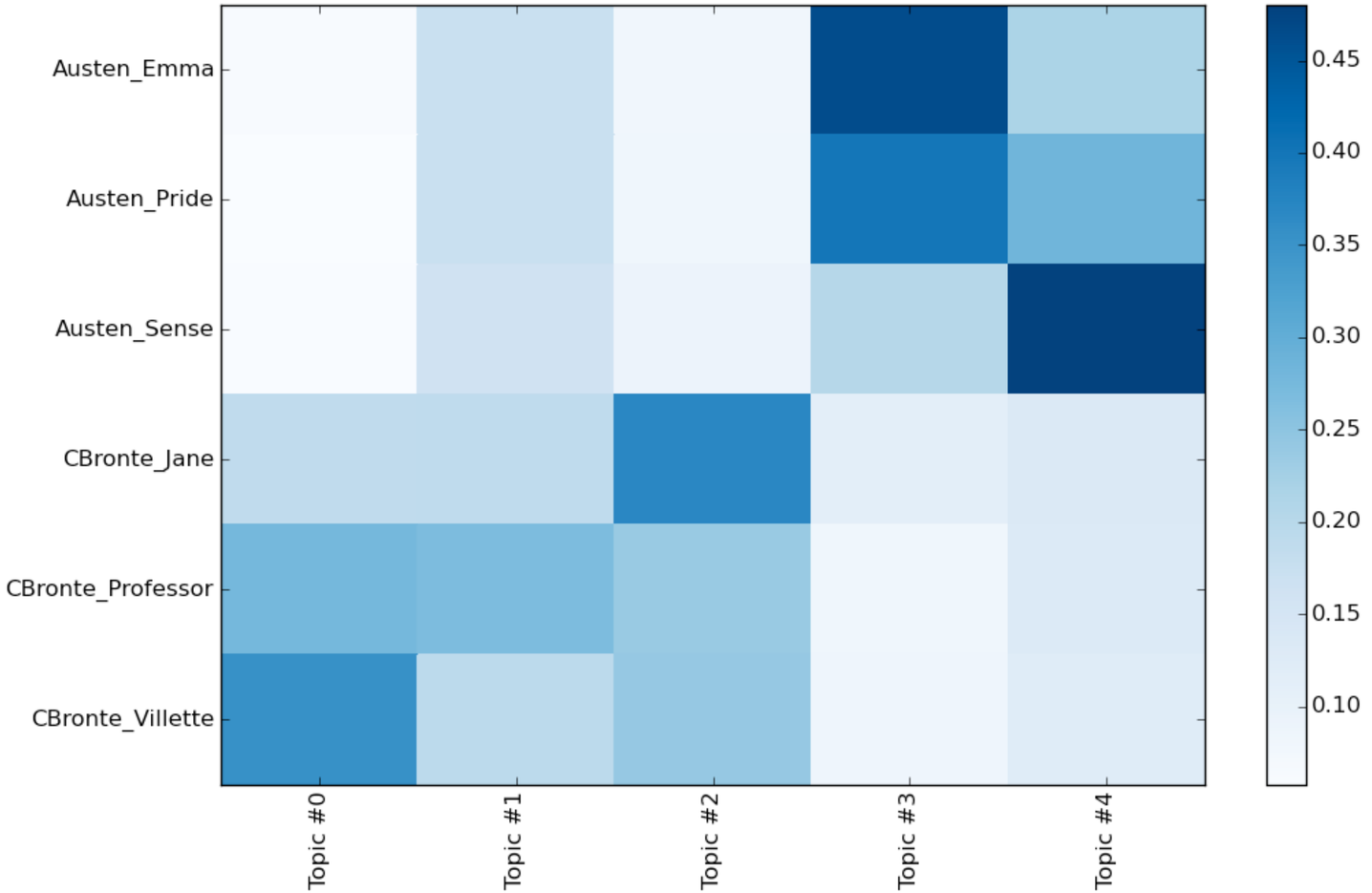
How do the documents relate to each other?

Visualizing Topic Models



https://de.dariah.eu/tatom/topic_model_visualization.html

Visualizing Topic Models

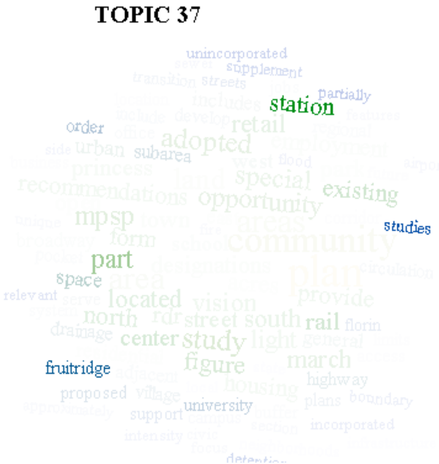


https://de.dariah.eu/tatom/topic_model_visualization.html

Visualizing Topic Models

https://de.dariah.eu/tatom/topic_model_visualization.html

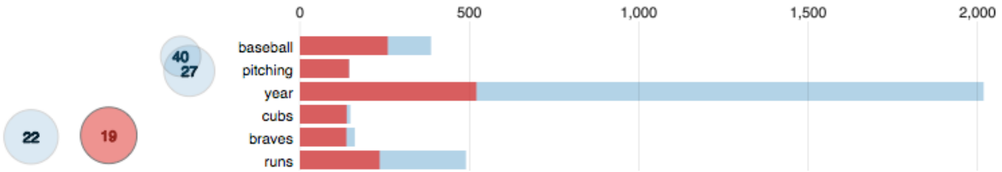
Visualizing Topic Models



Please don't…

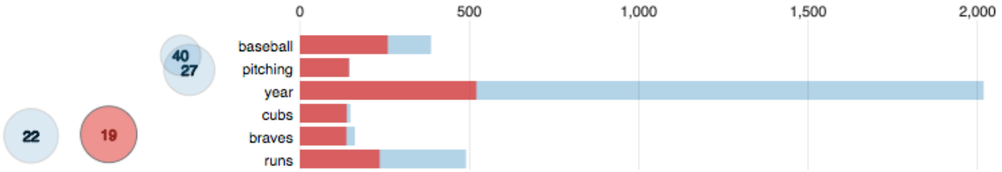
https://dhs.stanford.edu/algorithmic-literacy/using-word-clouds-for-topic-modeling-results/

LDAvis



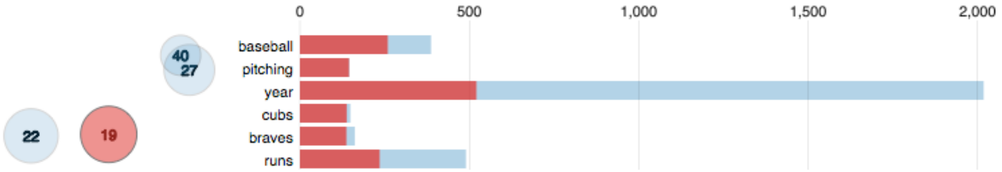
https://github.com/cpsievert/LDAvis

pyLDAvis



https://github.com/bmabey/pyLDAvis

pyLDAvis



https://github.com/bmabey/pyLDAvis

IP[y]

gensim

Demo Time!

Distinctiveness & Saliency

Termite: Visualization Techniques for Assessing Textual Topic Models
Jason Chuang, Christopher D. Manning and Jeffrey Heer. 2012

measure how much information a term conveys about topics

Distinctiveness & Saliency

	coding	tech news	video games
game	10	10	50
apple	20	40	20
angry birds	1	1	30
python	50	5	10

Distinctiveness & Saliency

	coding	tech news	video games
game	10	10	50
apple	20	40	20
angry birds	1	1	30
python	50	5	10

$$distinctiveness(w) = \sum_{T} P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

	coding	tech news	video games
game	10	10	50
apple	20	40	20
angry birds	1	1	30
python	50	5	10

P(T | game)	0.14	0.14	0.71
P(T | apple)	0.25	0.50	0.25
P(T | angry birds)	0.03	0.03	0.94
P(T | pyhton)	0.77	0.08	0.15
P(T)	0.33	0.23	0.45

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

	coding	tech news	video games
game	10	10	50
apple	20	40	20
angry birds	1	1	30
python	50	5	10
P(T | game)	0.14	0.14	0.71
P(T | apple)	0.25	0.50	0.25
P(T | angry birds)	0.03	0.03	0.94
P(T | pyhton)	0.77	0.08	0.15
P(T)	0.33	0.23	0.45

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

	coding	tech news	video games	distinctiveness
game	10	10	50	
apple	20	40	20	
angry birds	1	1	30	0.56
python	50	5	10	

P(T | game)	0.14	0.14	0.71
P(T | apple)	0.25	0.50	0.25
P(T | angry birds)	0.03	0.03	0.94
P(T | pyhton)	0.77	0.08	0.15
P(T)	0.33	0.23	0.45

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

	coding	tech news	video games	distinctiveness
game	10	10	50	0.15
apple	20	40	20	0.18
angry birds	1	1	30	0.56
python	50	5	10	0.41

P(T | game)	0.14	0.14	0.71
P(T | apple)	0.25	0.50	0.25
P(T | angry birds)	0.03	0.03	0.94
P(T | pyhton)	0.77	0.08	0.15
P(T)	0.33	0.23	0.45

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

	coding	tech news	video games	distinctiveness
game	10	10	50	0.15
apple	20	40	20	0.18
angry birds	1	1	30	0.56
python	50	5	10	0.41

$$saliency(w) = P(w) \times distinctiveness(w)$$

distinctiveness weighted by the term's overall frequency

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

	coding	tech news	video games	distinctiveness	P(w)
game	10	10	50	0.15	0.28
apple	20	40	20	0.18	0.32
angry birds	1	1	30	0.56	0.13
python	50	5	10	0.41	0.26

$$saliency(w) = P(w) \times distinctiveness(w)$$

distinctiveness weighted by the term's overall frequency

$$distinctiveness(w) = \sum_{T} P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

	coding	tech news	video games	distinctiveness	P(w)	saliency
game	10	10	50	0.15	0.28	0.04
apple	20	40	20	0.18	0.32	0.06
angry birds	1	1	30	0.56	0.13	0.07
python	50	5	10	0.41	0.26	0.11

$$saliency(w) = P(w) \times distinctiveness(w)$$

distinctiveness weighted by the term's overall frequency

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

computes the KL divergence between the distribution of topics given a term and the marginal distribution of topics

Distinctiveness & Saliency

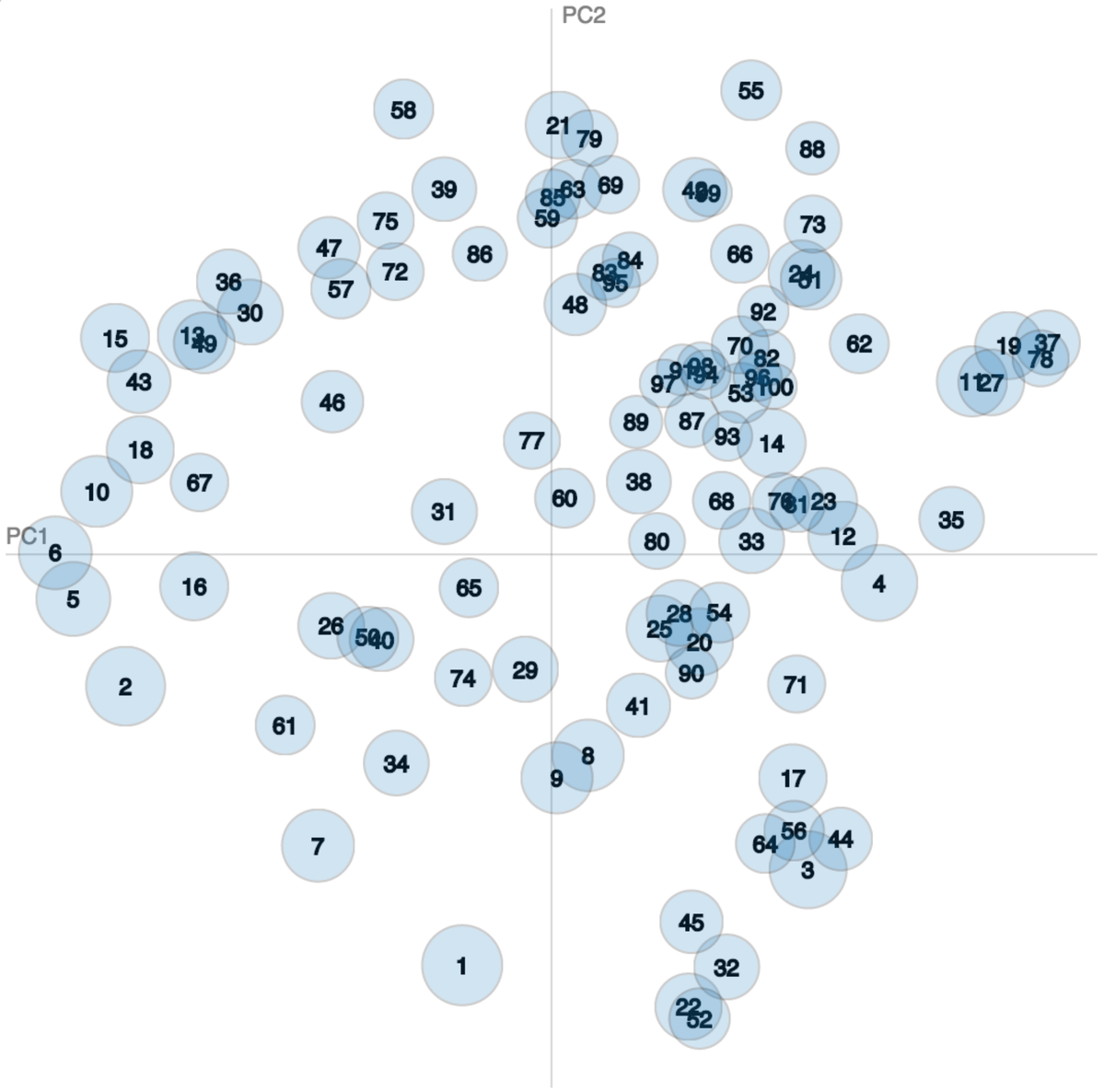
measure how much information a term conveys about topics…

Distinctiveness & Saliency

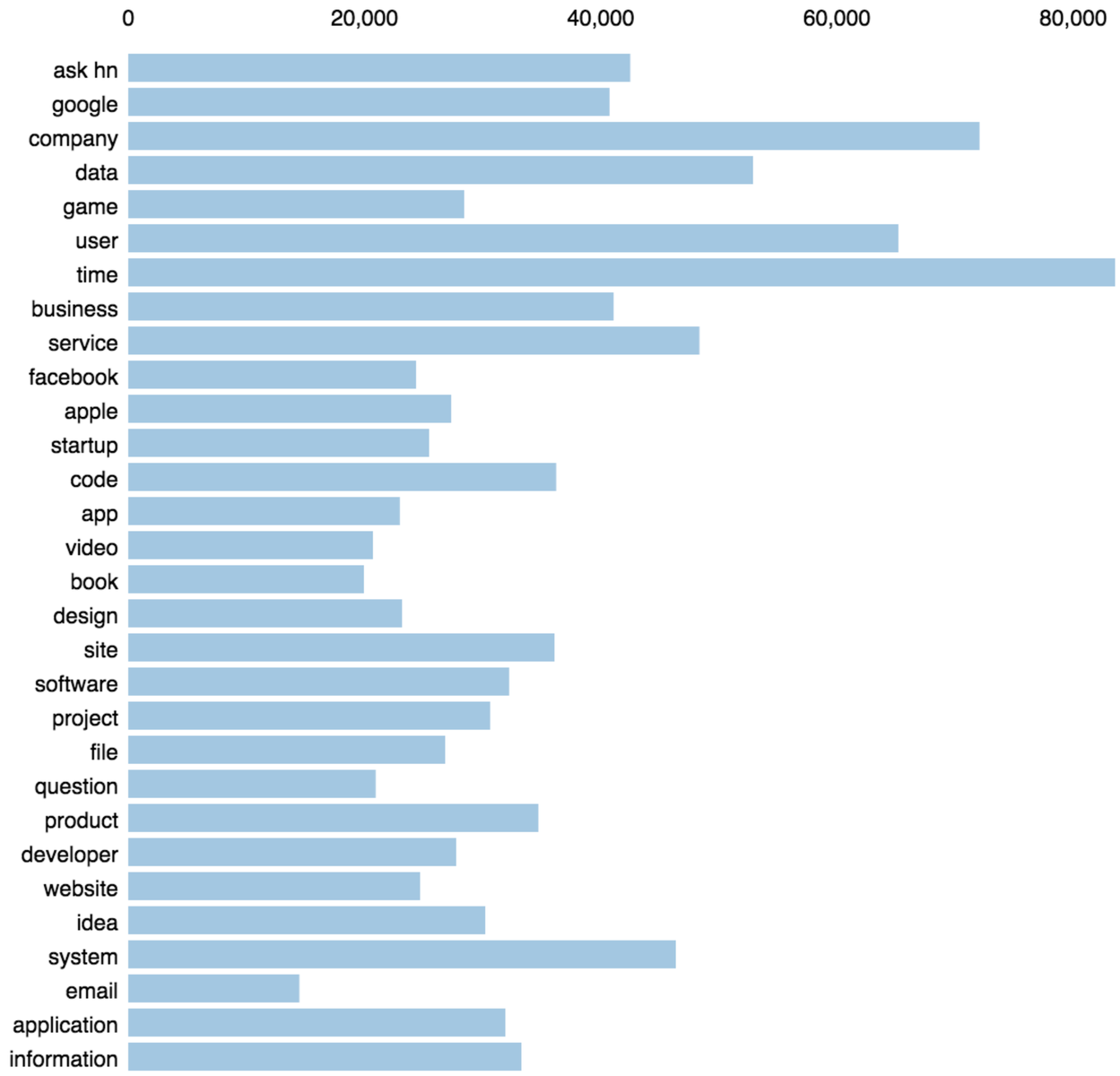
measure how much information a term conveys about topics…

globally

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms[1]



Thank you!

Learn more at http://github.com/bmabey/pyLDAvis

http://nbviewer.ipython.org/github/bmabey/hacker_news_topic_modelling/

Ben Mabey
@bmabey