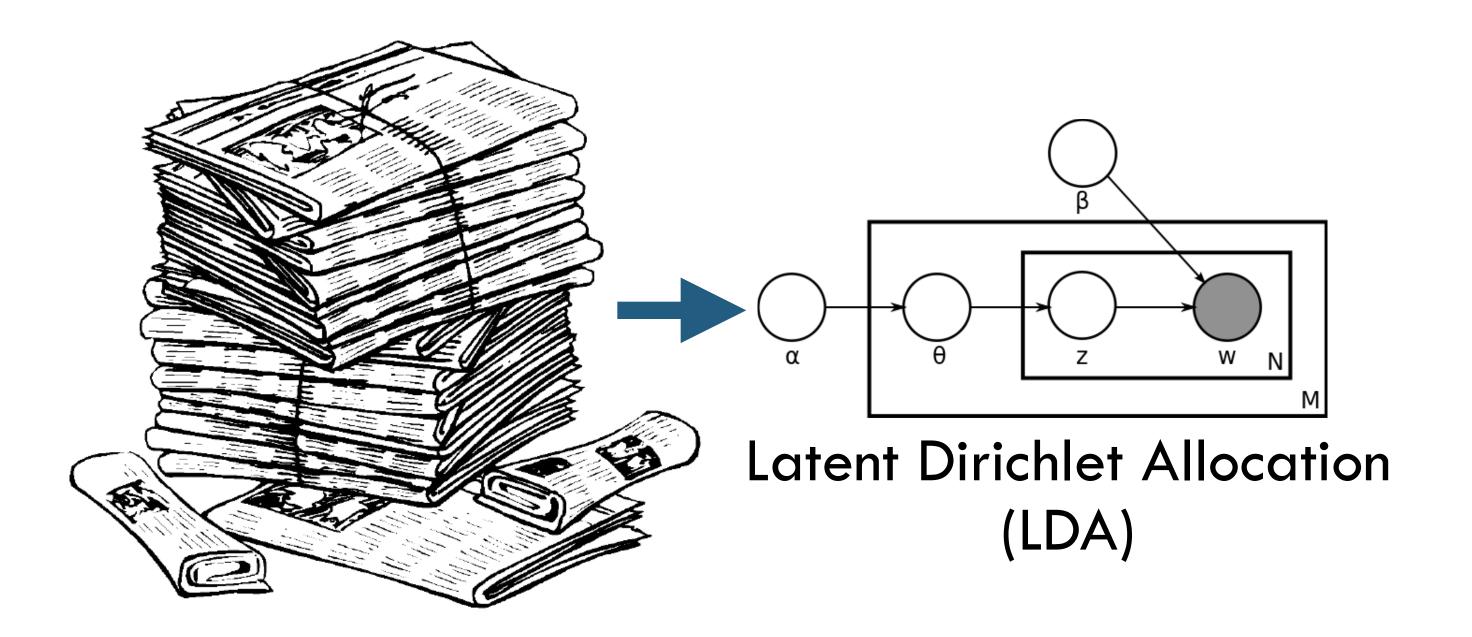
Ben Mabey abmabey (S)



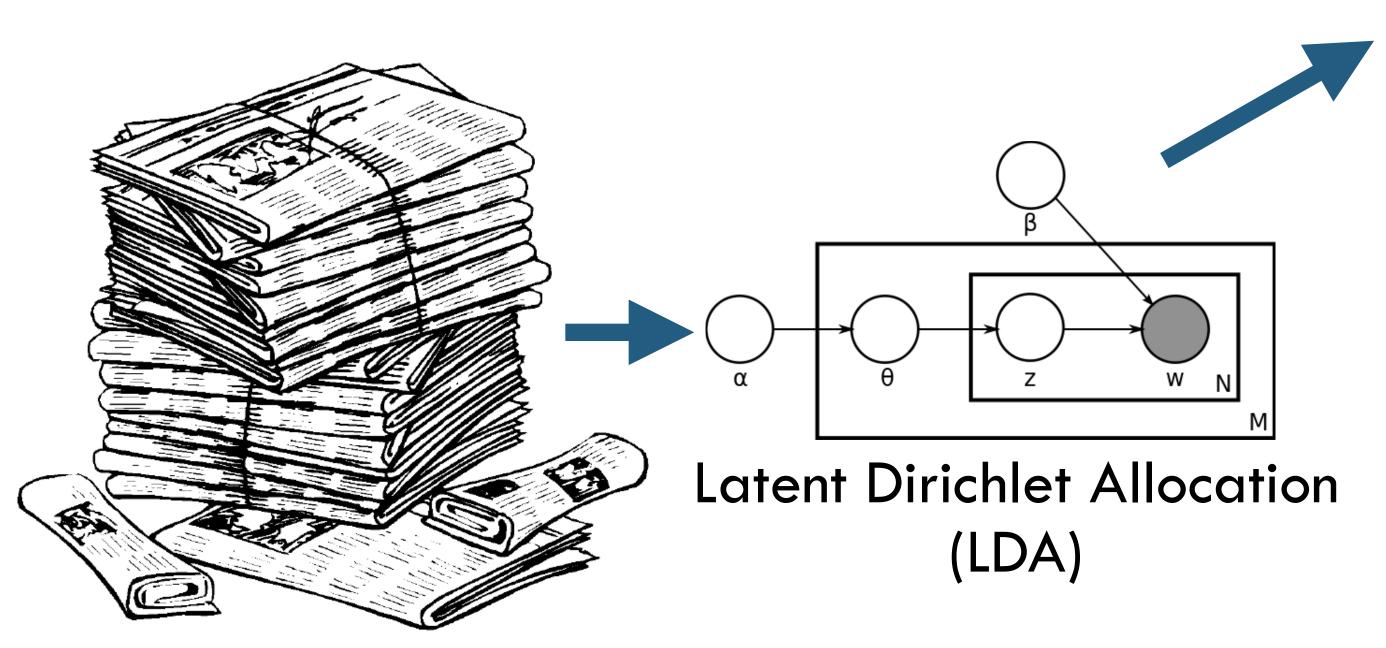






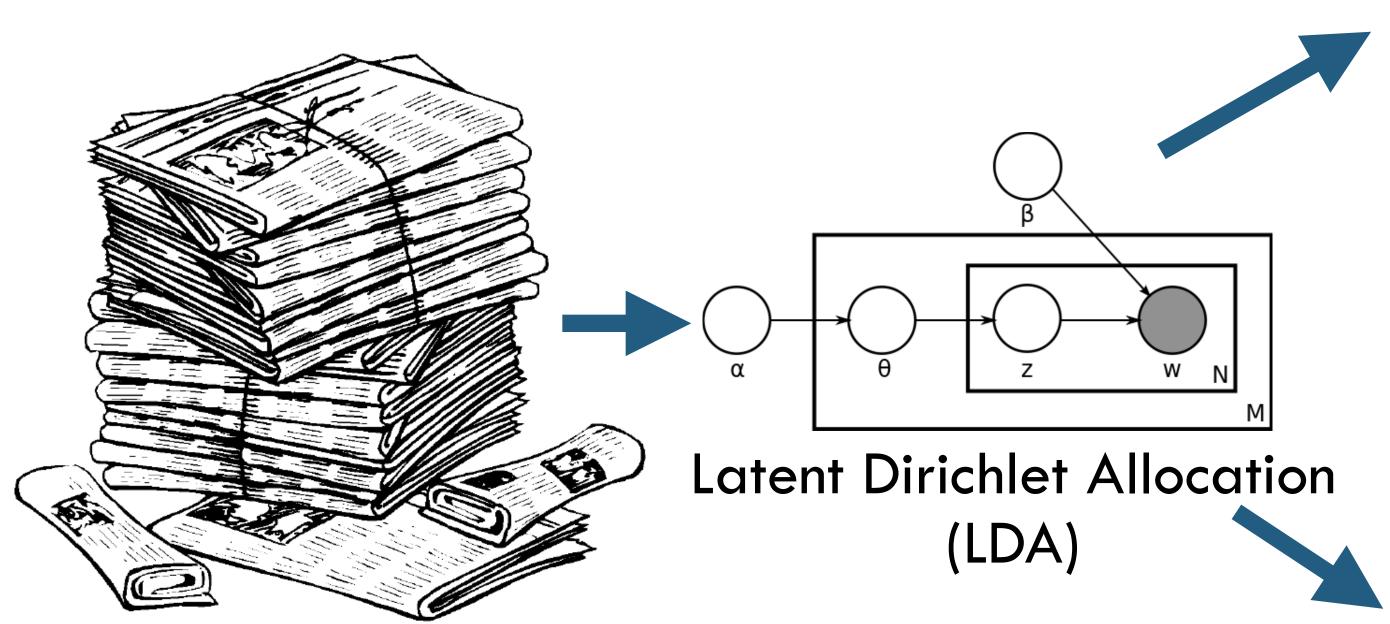


#### Document-Topic Distributions



	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

#### 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













POS tagging w/spaCy



Phrase detection w/Gensim







POS tagging w/spaCy



Phrase detection w/Gensim



Stopword removal & only kept nouns or phrases with nouns







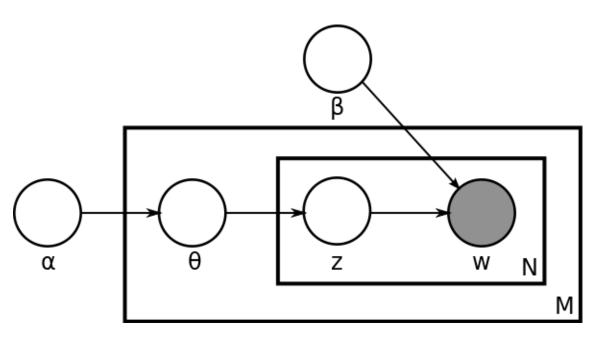
POS tagging w/spaCy



Phrase detection w/Gensim



Stopword removal & only kept nouns or phrases with nouns



Fit LDA models varying the number of topics













#### Document-Topic Distribution

Topic	P(T   D)
58	0.19
38	0.14
16	0.06
• • •	• • •







#### Document-Topic Distribution

Topic	P(T   D)
58	0.19
38	0.14
16	0.06
• • •	• • •

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







#### Document-Topic Distribution

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

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







#### Document-Topic Distribution

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

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







#### Document-Topic Distribution

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

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



What is the meaning of each topic?



What is the meaning of each topic?

How prevalent is each topic?



What is the meaning of each topic?

How prevalent is each topic?

How do the topics relate to each other?



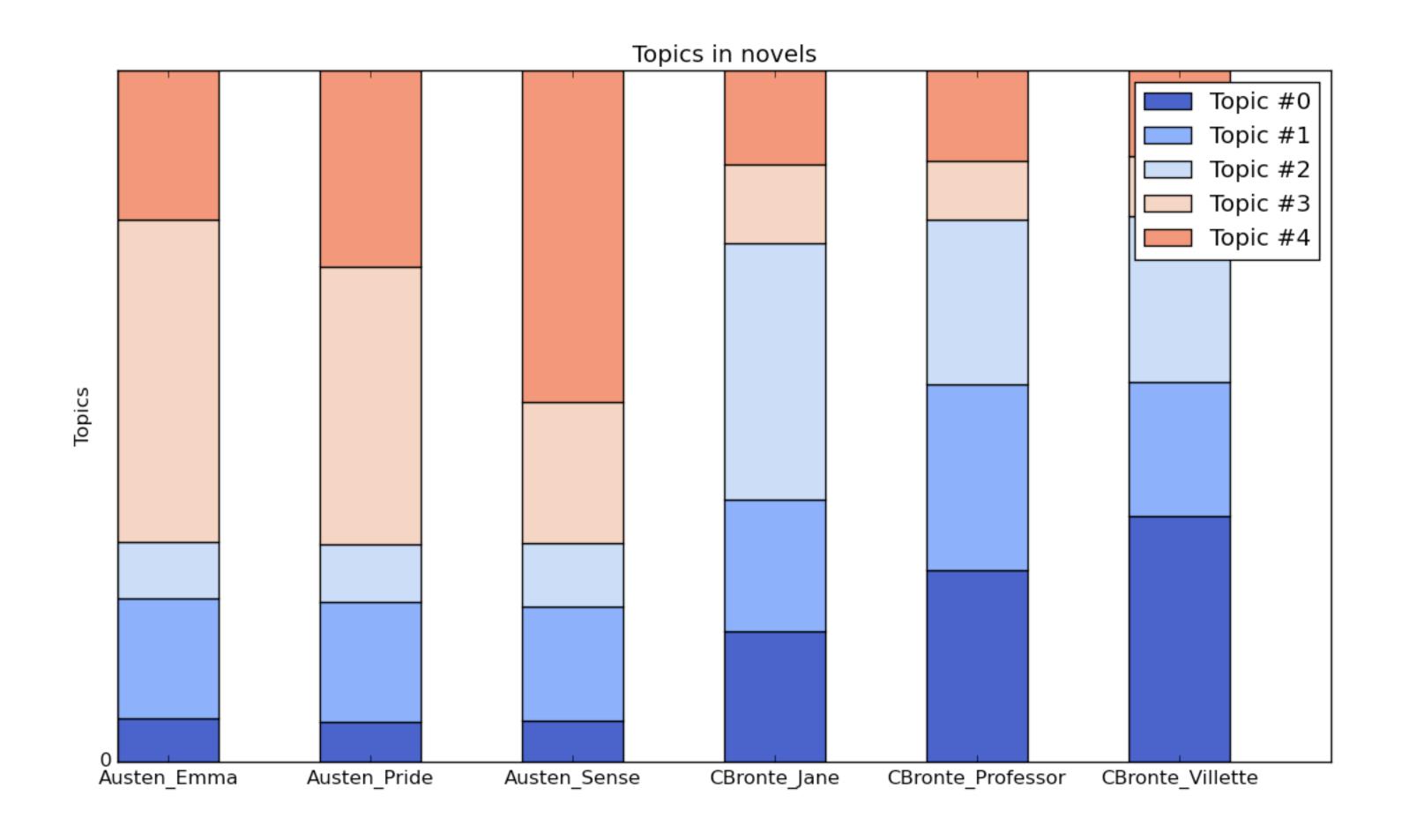
What is the meaning of each topic?

How prevalent is each topic?

How do the topics relate to each other?

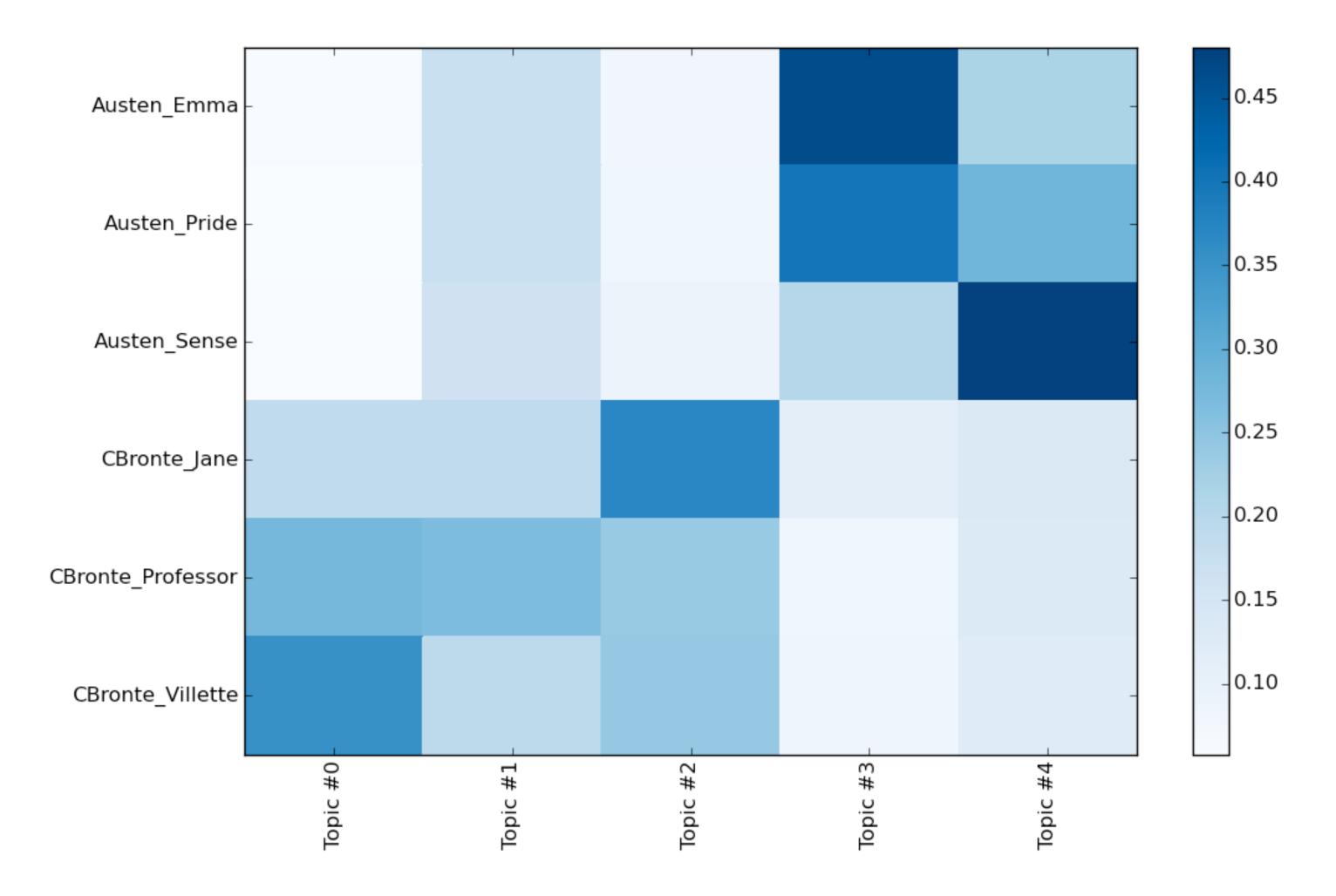
How do the documents relate to each other?





https://de.dariah.eu/tatom/topic\_model\_visualization.html





https://de.dariah.eu/tatom/topic\_model\_visualization.html



Topic #0	Topic #1	Topic #2	Topic #3	Topic #4
hand	looked	night	mr	elinor
good	found	room	miss	mother
madame	side	door	mrs	sister
life	speak	long	emma	marianne
heart	girl	house	jane	time
thought	gave	rochester	good	mrs
de	word	round	elizabeth	felt
day	made	hour	thing	letter
monsieur	sense	heard	dear	make
eye	eyes	back	great	john

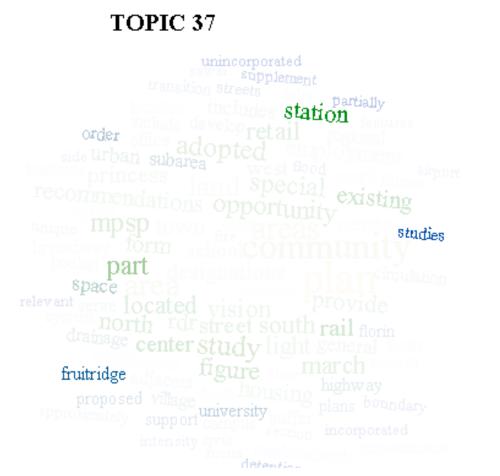
https://de.dariah.eu/tatom/topic\_model\_visualization.html



# supply administrate programs distance destinations efficient measures convenient bicycling projects maintain improve mark management choices safety information effective land goodstrave safe identify vehicle systems link freight reduce regional transit work mobility auto system direct transportation trips support transportation plan trucks agencies network modes maintain improvements plan trucks agencies network modes maintain improvements people efficiency

# successfully production trained providers logistics advanced focus tech factors advanced focus tech factors good clusters data System skill important skilled future work role important jobs Systems institutions fully success venture degree researcheconomy tracking region degrees workforce apital technologies prosperity high innovation industry successful employers education skills universities gain workers educated industries metropolitan companies educations talent monored leaders model foundations philanthropic class philanthropic class manual eaders model foundations philanthropic class successful providers logistics focus tech factors advanced focus tech factors advanced foundations systems institutions improving improving attracking region advanced foundations professionals start ingest philanthropic class focused report states focused report

# environmental concentration significant distribution expansion communities communities offer district. Cas link housing district. Cas link housing existing industrial large commercial area entire mix low shopping mixed retail goods sizes site activity sites. Surrounding cultural supper estimated due nodes land districts close location office serve light locate oriented services light mixture businesses adjacent provimity efforts.

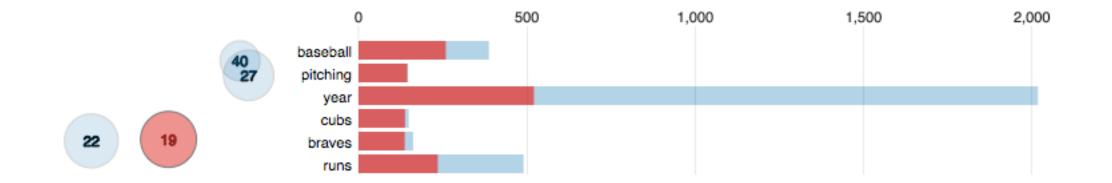


Please don't...

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



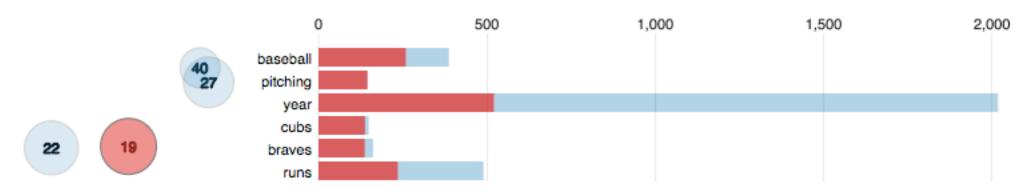
#### LDAvis



https://github.com/cpsievert/LDAvis



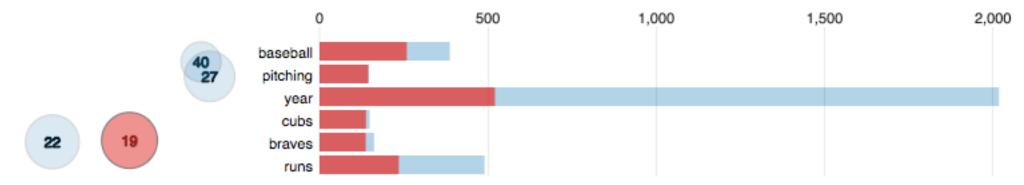
## 



https://github.com/bmabey/pyLDAvis

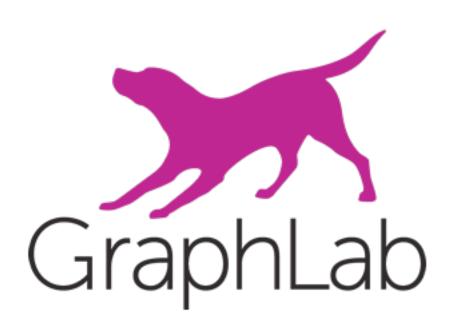


## 



https://github.com/bmabey/pyLDAvis









#### Demo Time!



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

	coding	tech news	video games
game	10	10	50
apple	20	40	20
angry birds	7	7	30
python	<b>50</b>	5	10

	coding	tech news	video games
game	10	10	<b>50</b>
apple	20	40	20
angry birds	7	7	30
python	<i>5</i> 0	5	10

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



	coding	tech news	video games
game	10	10	<b>50</b>
apple	20	40	20
angry birds	7	7	30
python	<b>50</b>	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

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



	coding	tech news	video games
game	10	10	50
apple	20	40	20
angry birds	7	7	30
python	<b>50</b>	5	10

P(T   angry birds)	0.03	0.03	0.94
P(T   pyhton)	0.77		
P(T)	0.33	0.23	0.45

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



	coding	tech news	video games	distinctiveness
game	10	10	<b>50</b>	
apple	20	40	20	
angry birds	7	7	30	0.56
python	<b>50</b>	<b>5</b>	10	

P(T   angry birds)	0.03	0.03	0.94
P(T   pyhton)	0.77		
P(T)	0.33	0.23	0.45

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



	coding	tech news	video games	distinctiveness
game	10	10	<i>5</i> 0	0.15
apple	20	40	20	0.18
angry birds	7	7	30	0.56
python	<i>5</i> 0	<b>5</b>	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

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



	coding	tech news	video games	distinctiveness
game	10	10	50	0.15
apple	20	40	20	0.18
angry birds	7	1	30	0.56
python	<i>5</i> 0	<b>5</b>	10	0.41

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

distinctiveness weighted by the term's overall frequency

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



	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	7	30	0.56	0.13
python	<b>50</b>	<b>5</b>	10	0.41	0.26

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

distinctiveness weighted by the term's overall frequency

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



	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	7	30	0.56	0.13	0.07
python	<i>5</i> 0	<b>5</b>	10	0.41	0.26	0.11

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

distinctiveness weighted by the term's overall frequency

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

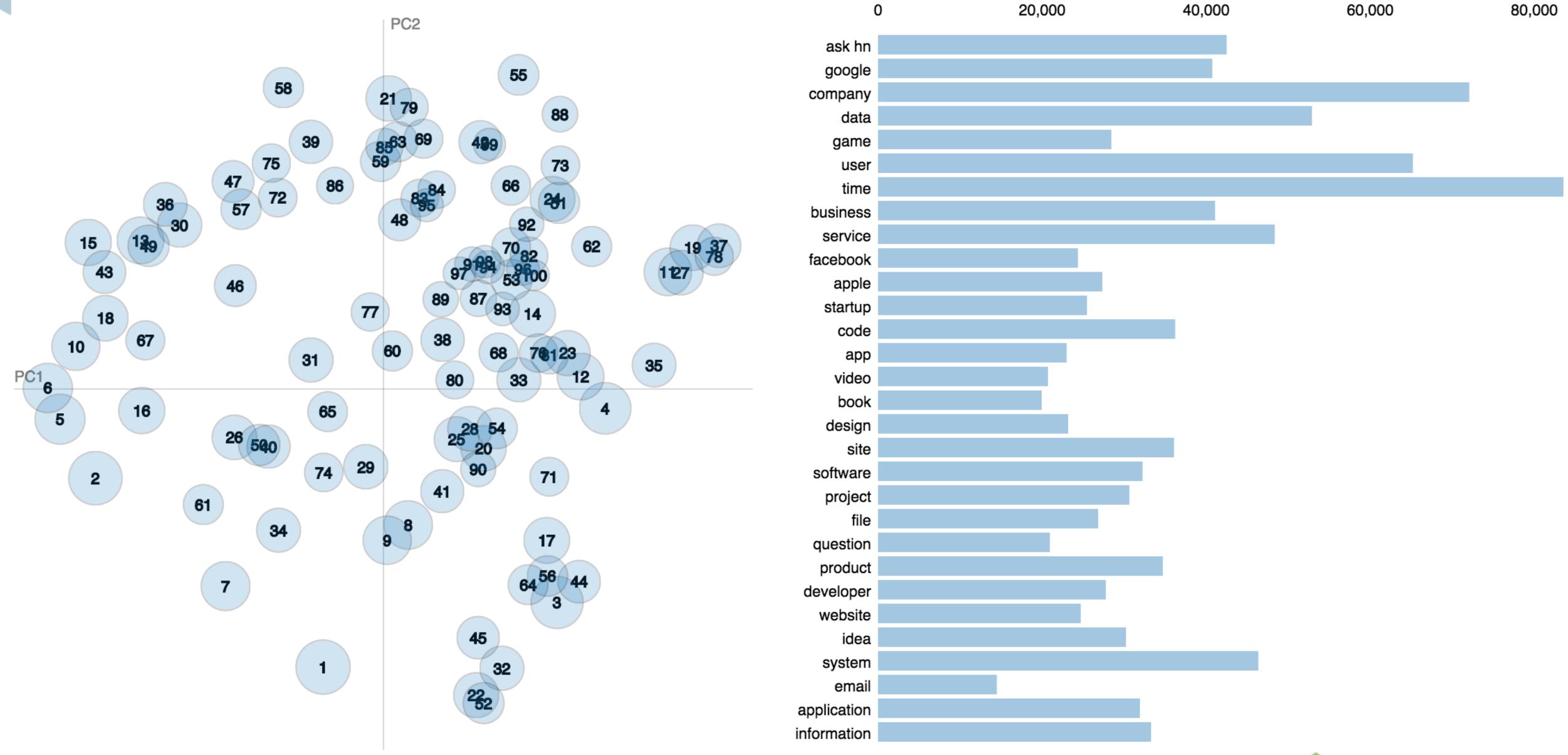


measure how much information a term conveys about topics...

measure how much information a term conveys about topics...

## globally

#### Top-30 Most Salient Terms<sup>1</sup>



## Thank you!

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

http://nbviewer.ipython.org/github/bmabey/hacker\_news\_topic\_modelling/

Ben Mabey Obmabey (S)





