Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling

Angela Fan

Finale Doshi-Velez

Luke Miratrix

(Harvard University)

Text as Data Conference, 2017

Princeton

Disclaimer:
This is an engineering talk about improving LDA

One observation and two simple tools

- ★ **Assessing topic quality is tricky, potentially misleading**, when vocabulary is shifting or when stop words are present.
- ★ **A simple manipulation of some priors** can cheaply nudge things in good directions.
- ★ **A simple measure of topic quality** that correlates well with information-content of topics.

We focus on the simplest model, but believe our general approach is applicable to all the flavors of this model.

I think we have a problem

Canonical LDA Model

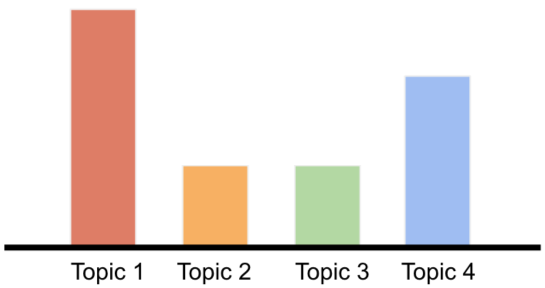
Blei et al. (2003)
+ millions more

DOCUMENT

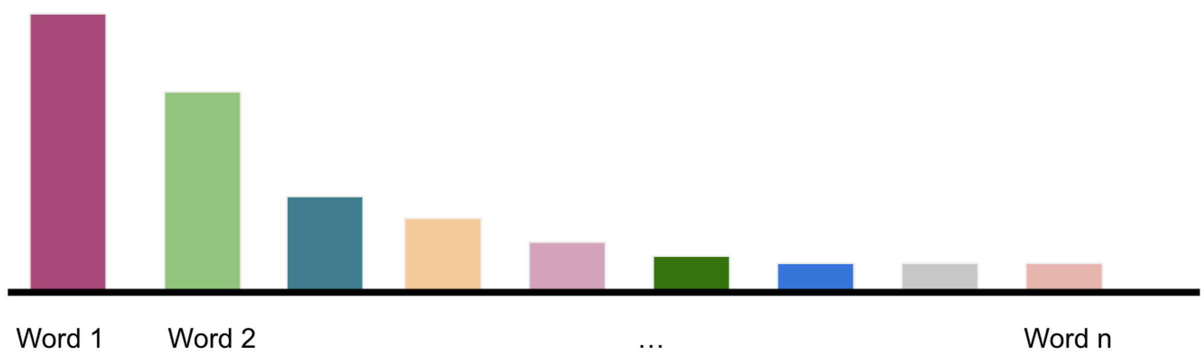
The author of a conventional novel constructs a world through language, creating a reality for the reader to immerse himself in. In such a text, the author is in control, and his writing is expected to guide the reader's thoughts and interpretations. However, Paul Auster, in the *New York Trilogy*, tests the boundaries of what constitutes a traditional detective novel. He pushes the restrictions of a genre whose every word is defined, with "no word that is not significant" and specific, predefined roles for the writer and reader to play. The "writer and detective are interchangeable, [and] the reader sees the world [only] through the detective's eyes" (9). Auster's characters know that they are characters and seek to escape the confines of the text to write the stories of their own lives, to assume authorship rather than simply remain characters. This action of escape suggests that readers, too, can take control of the text so that we are not reading Auster's novel but stories in which our own emotions and imaginations are inscribed.

But what are all these white words? What is that about?

DOCUMENTS ARE MULTINOMIAL DISTRIBUTIONS OF TOPICS



TOPICS ARE MULTINOMIAL DISTRIBUTIONS OVER THE VOCABULARY



What is a topic?
vs. How do we view topics?

A **topic** is a vector of probabilities over a set vocabulary.

We **view** a topic by taking the top K words

- either by total probability mass
- or by other measures such as “lift scores”

The quality of a topic is really the quality of this view

Stop words are a hassle and are not just canonical

Canonical: *and, the, but*…

Context-specific: *child, son, autism*, …

(in a corpus about children and autism)

Stop words are prevalent and can contaminate topics unless they are appropriately handled:

- Do we like this?

  from, approximately, fell, his, hospitalized, is
  him, falling, injured, in

  (in a corpus about workplace accidents)

Correlation with content words (e.g., “the” before nouns) increases their prevalence in co-occurance based approaches.

Conventional approaches to stop word removal are inadequate

Deletion methods:

- ★ Easy to use.
- ★ Miss contextual stop words.
- ★ Generally produces noisy lists.

Modeling methods:

- ★ These more complex methods can be difficult to integrate into complicated LDA models attending other things.
- ★ High barrier to entry.
- ★ Require tuning, not a slam-dunk.

Canonical stop word removal:
A convenient rug for hiding method failure.

A simple test:

When using some method, try *not* removing canonical stop words.

If you don't like what you see, why do you think that your method running *with* stop word removal is doing the right thing?

My propaganda:

“Regularize, don’t remove.”

Regularizing word appearance based on word frequency can make a real difference.

Not a new idea:

- TF-IDF scores
- Rescaling document-term matrices

See Miratrix and Ackerman (2016) for further discussion.

More problems: What does doing a “good job” mean?

There are a million ways to fit an LDA model.

How do you decide if you did well?

A Possible Gold Standard:

- Force humans to tell you.

Classic machine-based measures:

- Perplexity
    - Found to not well correlate with human judgement. Chang et al. (2009)
- Coherence
- Pointwise Mutual Information
    - Well, we will show a similar story…

Stop words can break common measures of topic quality

★ Common metrics like PMI and Coherence score topics with many stop words more highly than informative topics

★ They do not work when comparing models with differing vocabularies

★ These measures of quality do not correlate with human assessment of stop word contamination

(See our paper for more about why.)

PMI: Newman, Bonilla and Buntine, 2011, Coherence: Mimno et al., 2011

A quick example of this failure

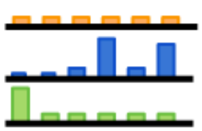
Common Topic Quality Metrics

counterintuitive performance when topics are stopword-heavy

(closer to 0 is better)

		Coherence	PMI
	**STANDARD LDA**	-554.2	-1.56

topic: social diagnosis as an or only autism child

		Coherence	PMI
	**INFORMATIVE PRIOR**	-1119.6	-2.42

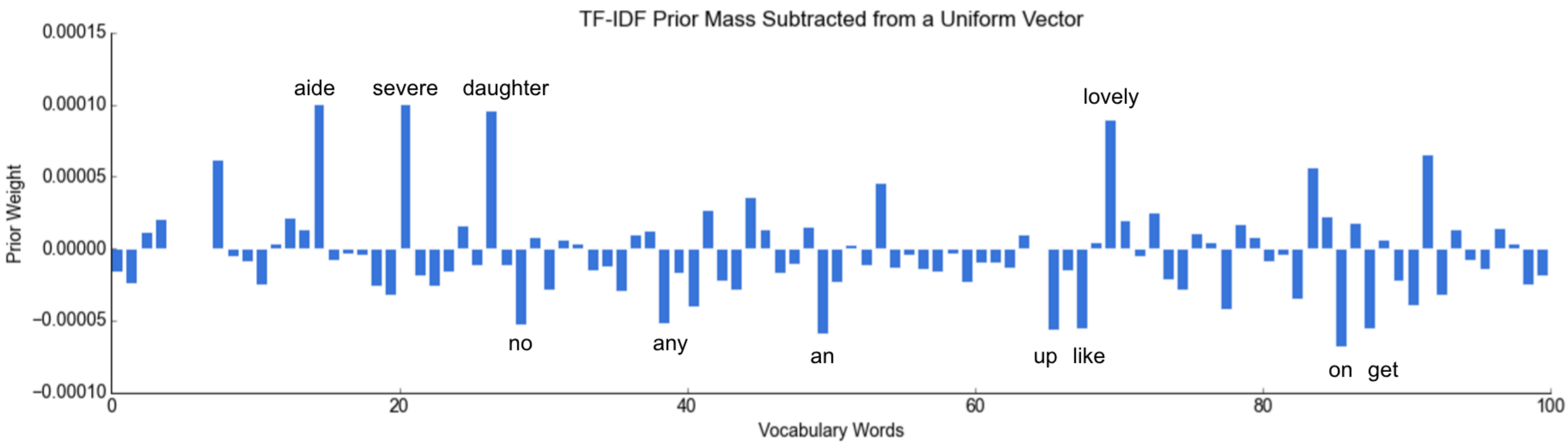
domain topic: learning attention symptoms similar problem development negative disorder positive school

stopword topic: child autism or on you it parent as son have

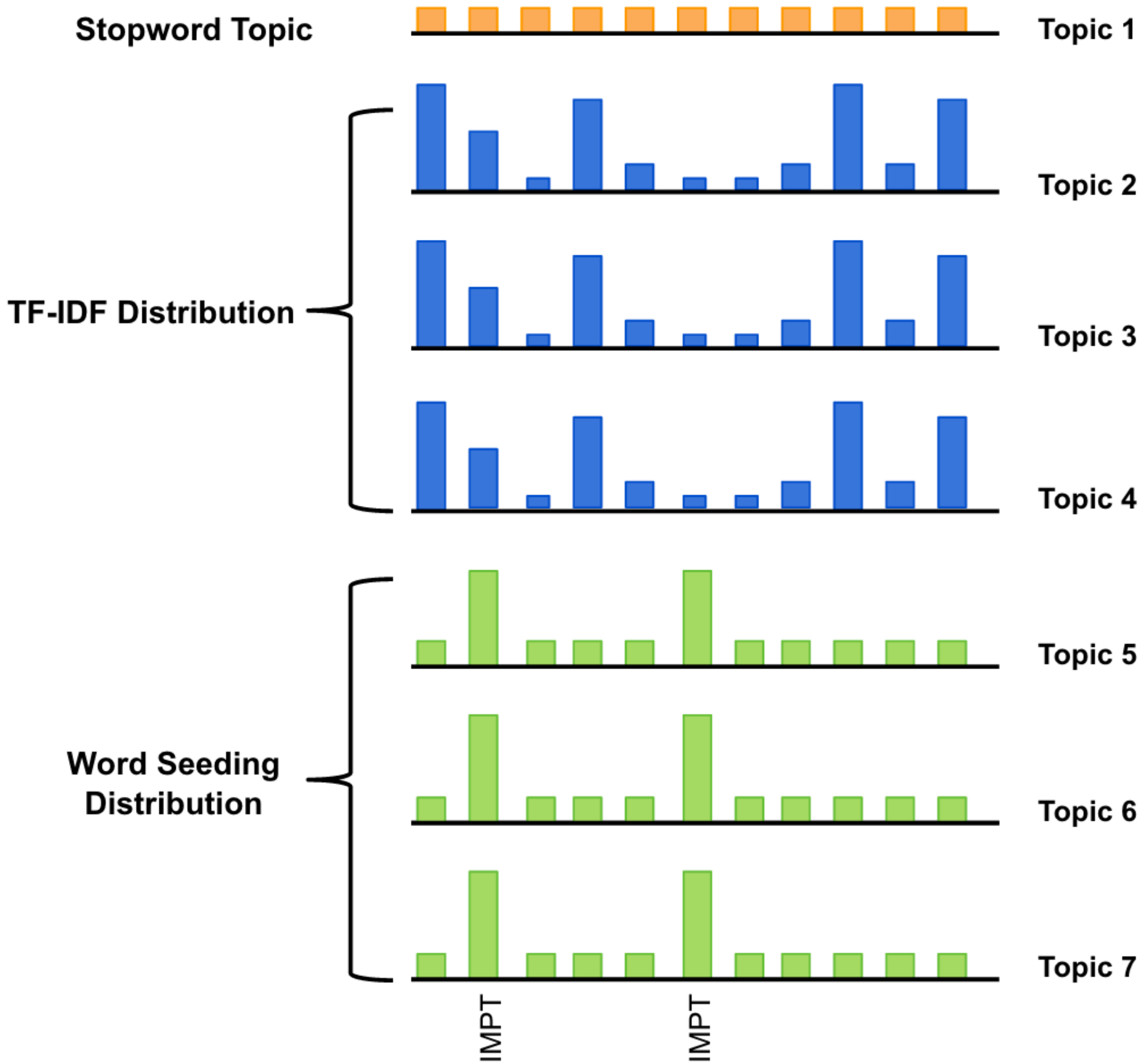
Playing with Priors

The priors of LDA

Main idea: use priors to promote words that are likely informative



Differing and Informative priors



Word Frequency or TF-IDF-weighted Topics: Discouraging high-frequency words

Idea: Put individual weights on words proportional to

Inverse word frequency

TF-IDF scores

to shrink rates of overall high-frequency words in "content topics" towards zero.

Stop word topics: A release valve

Idea: Make alternate topics without this shrinkage, giving stop words a place to go.

(These priors tend to be the canonical ones found in standard LDA.)

We will see that this plus the prior strategy successfully sequesters stop words to their topics.

Keyword seeding topics: pushing topics towards relevance

Idea: Tweak topics to prefer those words, and words that co-occur with them

While curating stop words is a nuisance, often generic whitelists of “good” words relevant to a corpus are easy to find.

We argue this is different from stop-word *removal:* don’t need to be comprehensive, for example.

Priors better than canonical deletion

Human Evaluation Study

evaluators were asked to circle low-information words

		% Human Stopwords
	**STANDARD LDA**	70.7
	**CANONICAL DELETION**	25.9
	**INFORMATIVE PRIOR**	

Only 12.7% in non-stopword topics!

So does it work?

Three Sample Datasets

Autism forums: 656,972 posts from three online support communities for autism patients and their caretakers.

OSHA Accidents: 49,558 entries from the Department of Labor Occupational Safety and Health database of casualties. Each entry describes a workplace accident.

NIPS abstracts: 403 abstracts from the Neural Information Processing Systems Conference 2015 accepted papers

Comparison to Baselines

We highlight three common baselines:

- No Deletion
- Stopword Deletion with a canonical stop word list (NLTK)
- Hyperparameter Optimization of the LDA priors

A qualitative peek at the output (Autism)

No Deletion Baseline:

social diagnosis as an or
only are autism that child

Stopword Deletion Baseline:

schools lea information need special
son statement parents support class

Hyperparameter Opt Baseline:

the to school needs support
statement we permit chairman he

TF-IDF & Keyword Seeding Prior:

learning attention symptoms similar problem
development negative disorder positive school

Example Stopword Topic:

child autism or on you
it parent as son have

Another peek (Accidents)

No Deletion Baseline:

from approximately fell his hospitalized is him falling injured in

Stopword Deletion Baseline:

report trees surface backing inc degree determined forks fork board

Hyperparameter Opt Baseline:

the employee lift number operator operating approximately jack to by

TF-IDF & Keyword Seeding Prior:

work rope tree landing protection caught lift edge open story

Example Stopword Topic:

hospitalized employee by for at when ft fall his fell

Evaluating performance

We measure

1) Percent of canonical stop words in word lists

2) Percent of words marked by domain experts as important

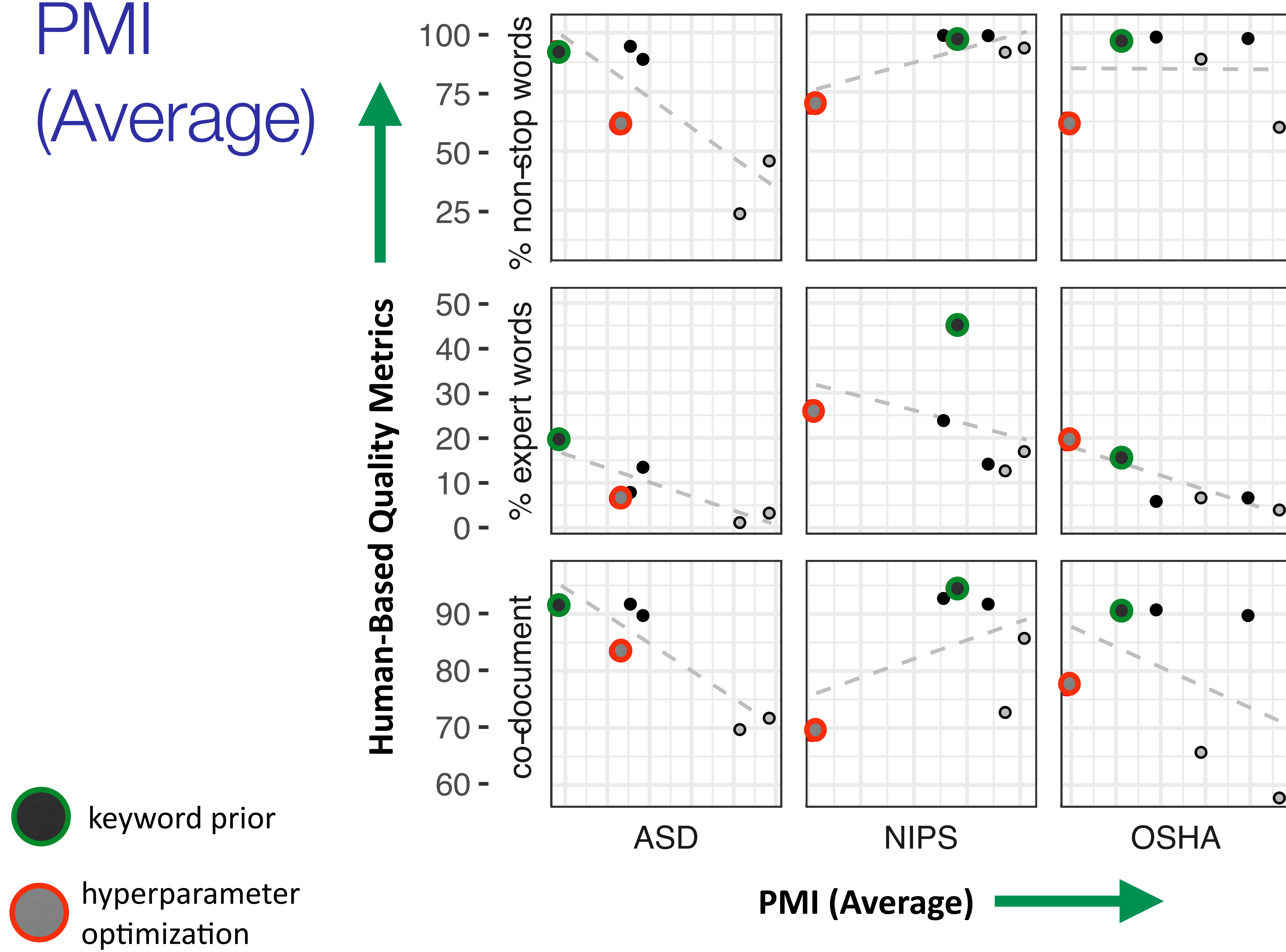
3) Co-occurence of domain expert words with topic words

and compare these scores to
PMI and Coherence

Two questions:

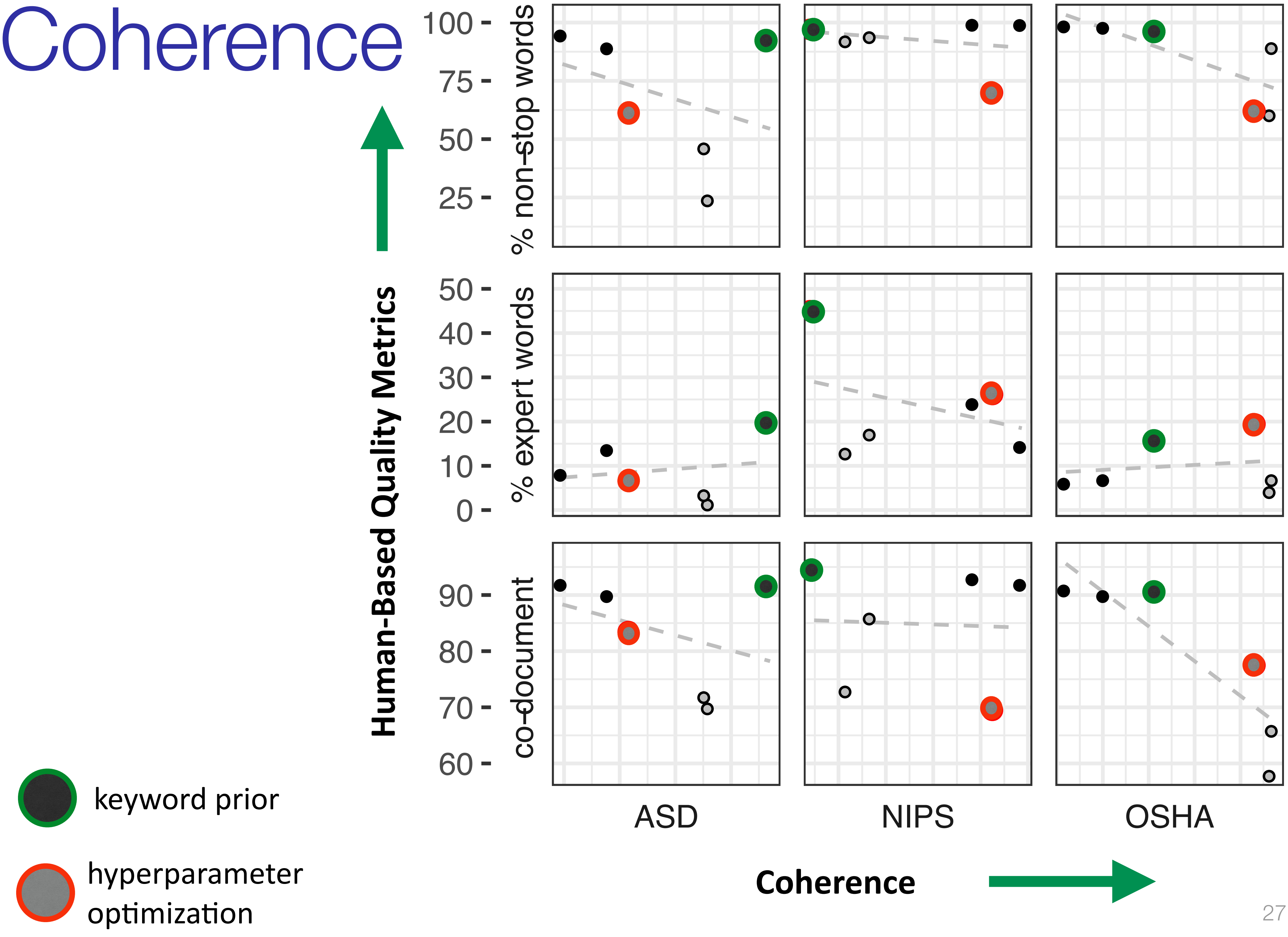
- How do these scores compare to machine assessment?
- What methods work best?

PMI (Average)



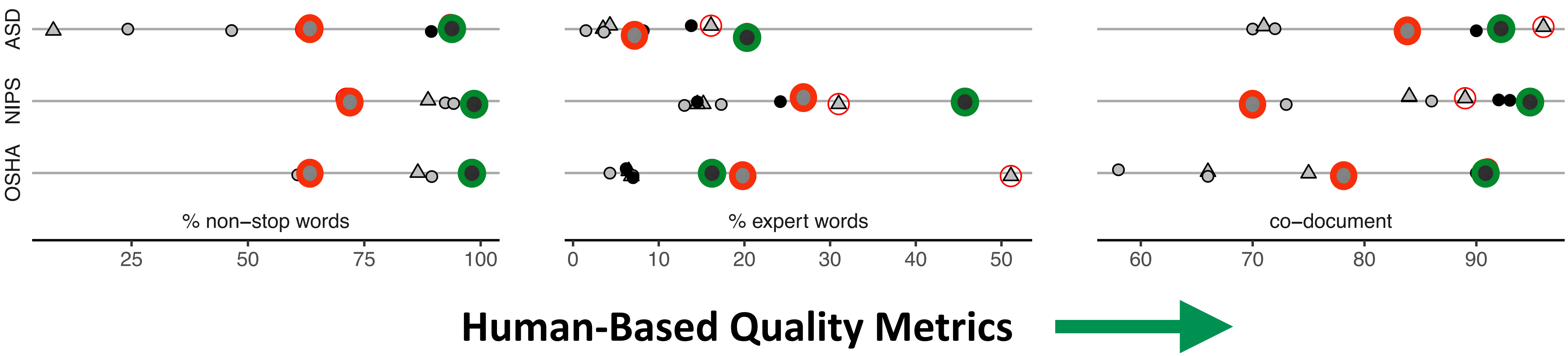
Human-Based Quality Metrics
% non-stop words
% expert words
co-document
100
75
50
25
50
40
30
20
10
0
90
80
70
60
ASD
NIPS
OSHA
PMI (Average)
keyword prior
hyperparameter optimization

Coherence



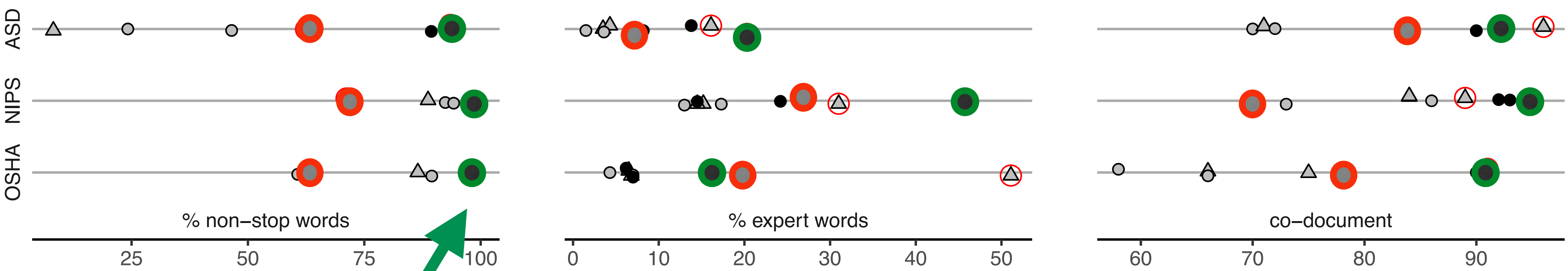
100
75
50
25
% non-stop words
50
40
30
20
10
0
% expert words
90
80
70
60
co-document
Human-Based Quality Metrics
ASD
NIPS
OSHA
Coherence
keyword prior
hyperparameter optimization

The ranking of the modeling methods



Informative priors with keyword augmentation
tend to outperform hyperparameter optimization

Ranking of Prior Methods



Human-Based Quality Metrics

As expected, informative priors are particularly good at isolating stop words, and almost all stop words appear in the designated stop word topic.

Another machine metric
to predict this type
of topic quality

Evaluating topic quality automatically is difficult

... and requires multiple metrics

We propose looking at **lift** as well, which we find correlates well with human-based metrics for the stop word problem

Average lift - a new score for topic quality

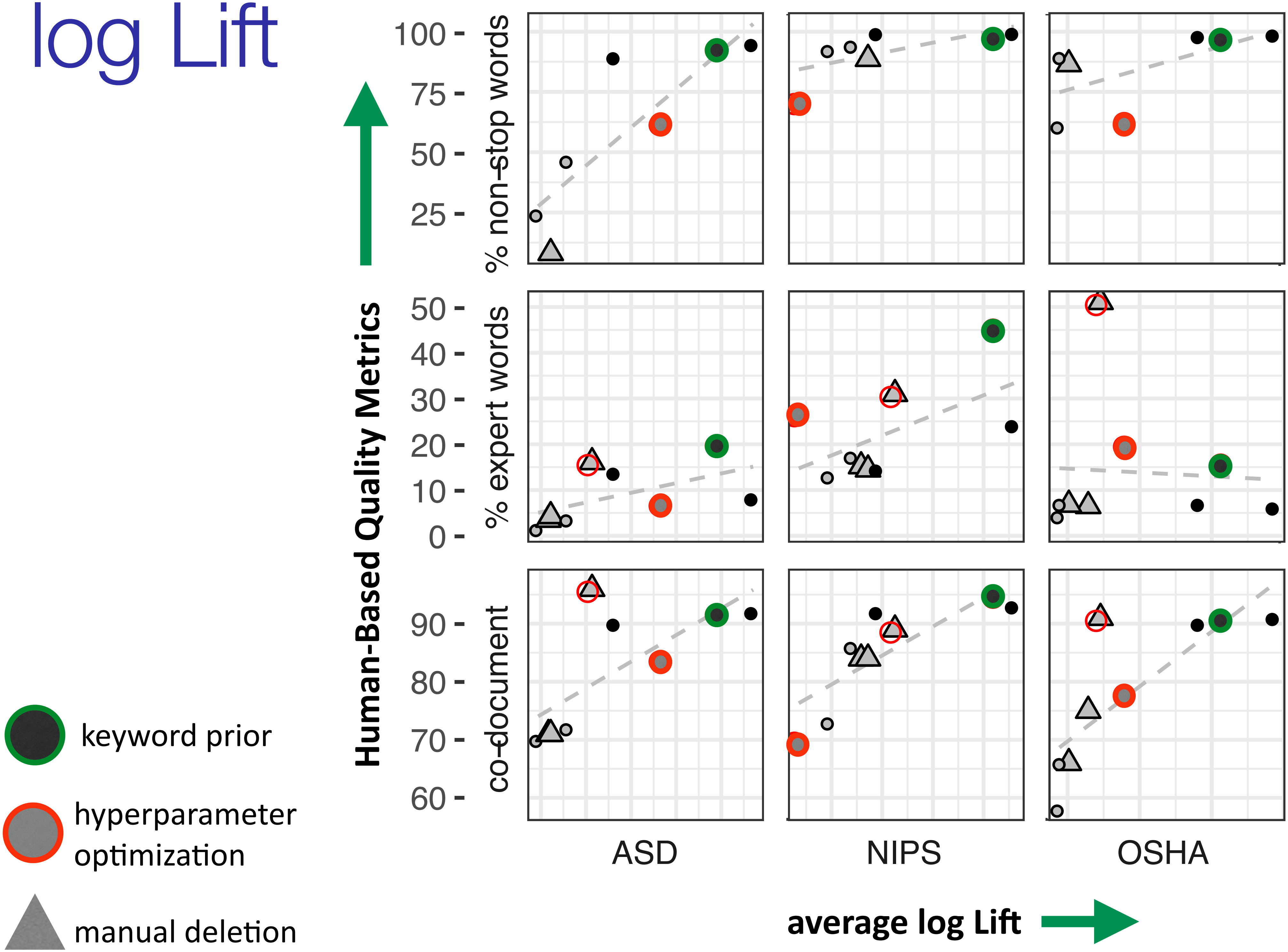
For topic t we average the top J lift scores:

$$\mathit{lift}_t = \frac{1}{J}\sum_{j=1}^{J} \log \mathit{lift}_{(j)t} \text{ with } \mathit{lift}_{jt} = \frac{\beta_{tj}}{b_j}$$

Motivation:

- We want topics to be well separated.
- We want words that have a high differential rates of appearance in their primary topics

log Lift



Human-Based Quality Metrics
% non-stop words
100
75
50
25
% expert words
50
40
30
20
10
0
co-document
90
80
70
60
ASD
NIPS
OSHA
average log Lift
keyword prior
hyperparameter optimization
manual deletion

Conclusions

- Standard ways of scoring topics are not necessarily awesome (no surprise there)
- Stop words should be handled in an integrated fashion.
- One way is by manipulating priors
  - Priors are easily added with a couple lines of code.
  - No modifications needed to inference procedures.
  - It is computationally cheap, which is important for complex models.

Next steps / current uses

Extending to STM

- Reagan Rose (in audience) working to extend these principles to STM adapting to its different latent structures.

Matching on topic proportions can be improved

- Matching on *content* topic proportions appears to give improved match quality.

Thank you

This work is built on many things. For complete references please see our working paper on arXiv:

https://arxiv.org/abs/1701.03227

Acknowledgements:

My collaborators, Angela Fan (the primary driver of this work) and Finale Doshi-Velez.

Hanna Wallach, for thoughts on the project and pushing us to consider hyperparameter optimization. Chirag Lakhani and Tim Miller (and others) for scoring topics and topic

Full & illegible table of results

Corpus	Model	Coherence 10 wds	30 wds	Avg. PMI	log Lift	% Stopword	% expert	Co-Doc Appear.
ASD	No Deletion Baseline	-45.5	-554.2	-1.56	1.94	76%	2%	70%
	Stopword Deletion Baseline				2.17	0%	4%	71%
	TF-IDF Deletion Baseline				2.22	92%	4%	71%
	Keyword Topics Baseline	-48.2	-580.1	-1.42	2.61	54%	4%	72%
	Deletion + Hyp. Opt.				3.13	0%	16%	96%
	Hyperparameter Opt.	-105.8	-1107.9	-2.12	4.73	38%	7%	84%
	Word Frequency Prior	-115.2	-1278.3	-2.02	3.65	15% (11%)	14% (14%)	90%
	TF-IDF Prior	-143.3	-1611.8	-2.08	6.71	10% (5%)	9% (8%)	92%
	Keyword Seeding Prior	-102.8	-119.6	-2.42	5.98	9% (6%)	20% (20%)	92%
NIPS	No Deletion Baseline	-71.2	-790.7	-2.06	2.96	8%	13%	73%
	Stopword Deletion Baseline				3.58	0%	15%	84%
	TF-IDF Deletion Baseline				3.72	11%	14%	84%
	Keyword Topics Baseline	-71.0	-765.2	-1.97	3.42	6%	17%	86%
	Deletion + Hyp. Opt.				4.25	0%	31%	89%
	Hyperparameter Opt.	-72.7	-633.2	-2.96	2.35	29%	27%	70%
	Word Frequency Prior	-76.5	-606.5	-2.14	3.91	3% (1%)	16% (14%)	92%
	TF-IDF Prior	-86.7	-656.8	-2.35	6.60	4% (0%)	24% (24%)	93%
	Keyword Seeding Prior	-87.1	-825.7	-2.28	6.27	3% (2%)	48% (46%)	95%
OSHA	No Deletion Baseline	-68.2	-831.9	-2.66	2.89	39%	4%	58%
	Stopword Deletion Baseline				3.29	0%	6%	75%
	TF-IDF Deletion Baseline				3.02	14%	7%	66%
	Keyword Topics Baseline	-68.5	-819.9	-3.01	2.91	10%	7%	66%
	Deletion + Hyp. Opt.				3.46	0%	51%	91%
	Hyperparameter Opt.	-74.8	-899.1	-3.60	3.85	37%	20%	78%
	Word Frequency Prior	-154.4	-1738.6	-2.80	4.83	6% (2%)	8% (7%)	90%
	TF-IDF Prior	-171.9	-1951.2	-3.21	5.87	5% (1%)	7% (6%)	91%
	Keyword Seeding Prior	-129.8	-1447.4	-3.36	5.18	5% (2%)	17% (16%)	91%