NEW YORK UNIVERSITY

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

Provably and Practically Learning Topic Models

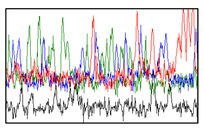
Yoni Halpern[1], Sanjeev Arora[2], Rong Ge[2], Ankur Moitra[2], David Mimno[2], Yichen Wu[2], Michael Zhu[2] and David Sontag[1]

1. Computer Science Dept. New York University, 2. Princeton University

Abstract

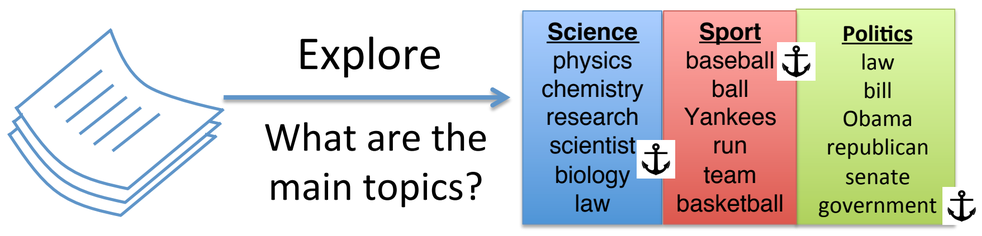
We improve on the provable method of learning topic models with anchor words described in Arora et al. (2012).

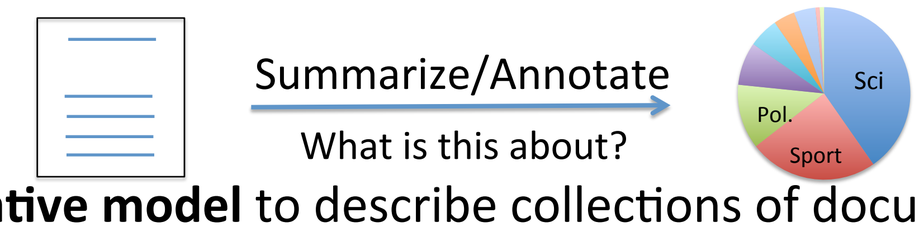
The new algorithm combines strong theoretical results and practicality:

- Provable ✓ - Robust to noise 

-Fast  -High quality topics 

Topic Models





- **Generative model** to describe collections of documents.
- Each "topic" describes a distribution of words.
- Bag of words for each document from a distribution of topics.
- **Additional anchor assumption:** For every topic there is at least one word that can only come from that topic.

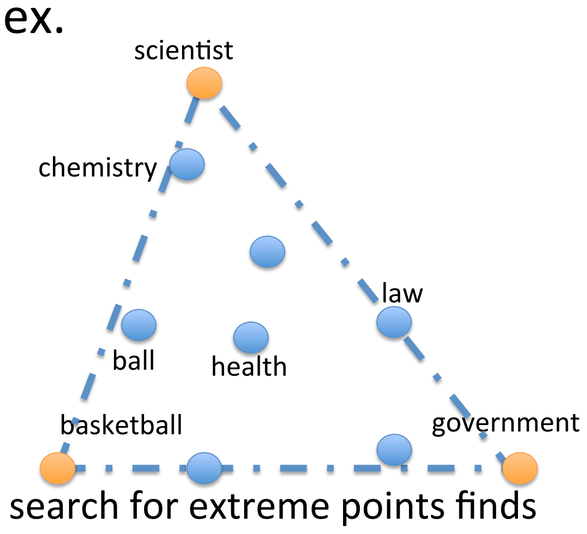
Representation

-Describe each word by a distrib. over co-occurring words.
-Simply computed from counts.

	lab	Yankees	physics	race	senate	
research	10	5	1	2	3	
law	4	3	0	1	8	
basketball	1	3	1	2	4	...
baseball	1	2	5	0	0	
		...				⋱

Finding Anchor Words

In this representation, anchors lie at the corners of a low dim. simplex.



A search for extreme points finds anchors in O(KVlogV).

Recovering Topics

Probability of topic conditioned on observing word *i* is related to the *position* of a word in the simplex.

$$P(z|w=i) = \underset{\alpha}{\operatorname{argmin}} ||Q_i - \sum_k \alpha_k Q_k||$$

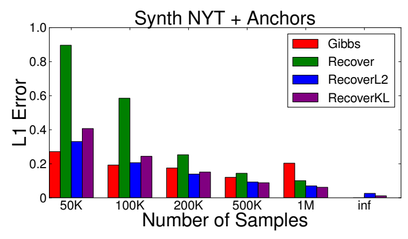
$$\text{s.t. } \alpha_k \geq 0, \sum_k \alpha_k = 1$$

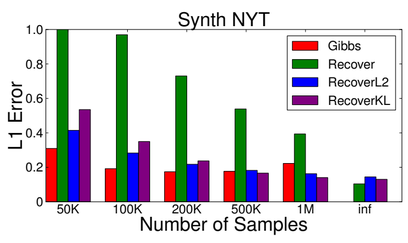
- Convex optimization problem solved using the *exponentiated gradient* algorithm.
- Inversion with Bayes Rule.

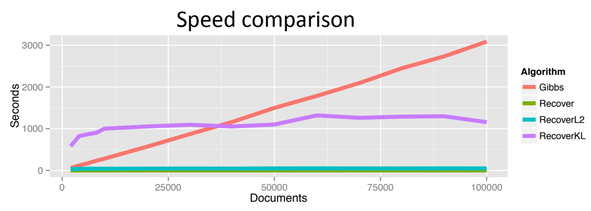
Guarantees

- Method-of-moments algorithm to recover topics from a corpus of documents.
- If data is drawn from topic model distribution: the original distribution can be recovered to *arbitrary accuracy* with a *polynomial* number of samples.

Simulations

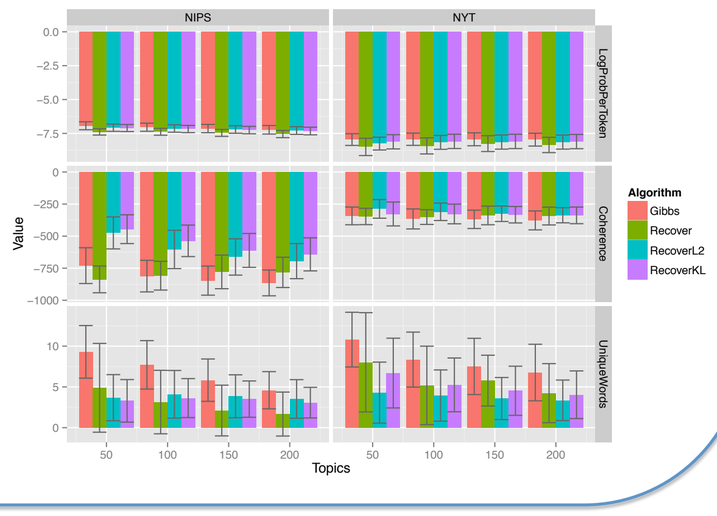






Real Data Evaluation

Example topics in New York Times
- cell, stem, research, Bush, patient
- cup, minutes, add water tablespoon oil
- oil, prices, percent, million, United States
- anthrax, official, mail, letter, worker, attack



References

- Blei, D., Ng, A., and Jordan, M. Latent Dirichlet Allocation. Journal of Machine Learning Research, pp. 993–1022, 2003.
- Arora, S., Ge, R., and Moitra, A. Learning Topic Models – Going Beyond SVD. In FOCS, 2012