A Practical Algorithm for Topic Modling with Provable Guarantees

Sanjeev Arora
Rong Ge
Yoni Halpern
David Mimno
Ankur Moitra
David Sontag
Yihcen Wu
Michael Zhu

Presented by: Vanush Vaswani and Kristy Hughes

Information Overload

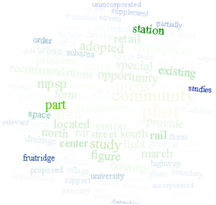
Effective Organisation

Topics

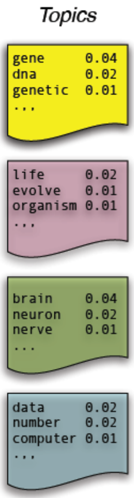

transportation
parking
system
mobility
movement
network
travel
safe
trips
plan
trucks
mode
people
priority
demand
impacts
walking
reduce
regional
transit
vehicle
goods
management
safety
improve
strategies
destinations
accessibility
programs
provide
convenient
bicycling


workforce
innovation
education
economy
research
technology
training
systems
industry
skills
metropolitan
industries
entrepreneurs
clusters
skilled
data
work
important
institutions
technologies
universities
employers
improving
logistics


commercial
industrial
areas
retail
residential
mixed
activity
district
existing
office
districts
surrounding
designated
environmental
concentration
business
structures
large
area
goods
shopping
location
oriented
services
accommodate
businesses
adjacent


plan
community
areas
station
retail
adopted
existing
special
opportunity
mpsp
part
located
study
street
south
rail
north
center
figure
march
fruitridge
studies
university

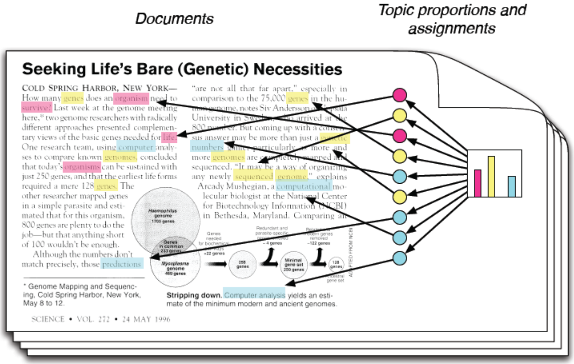
Topics



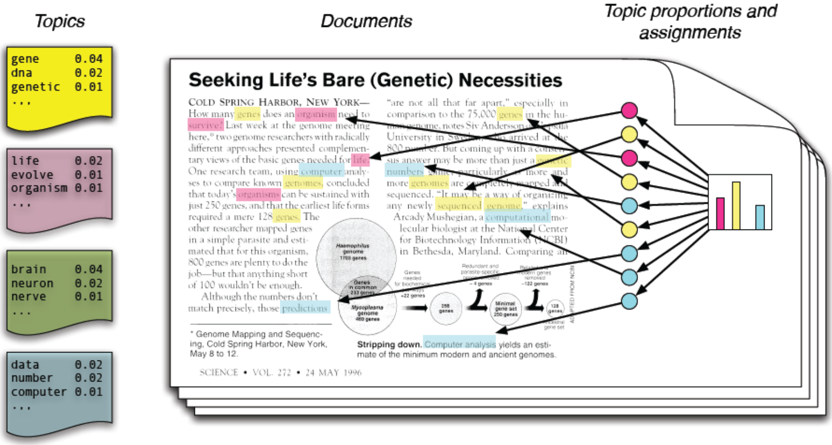
Topics are distributions over words

Documents

Documents have distribution of topics

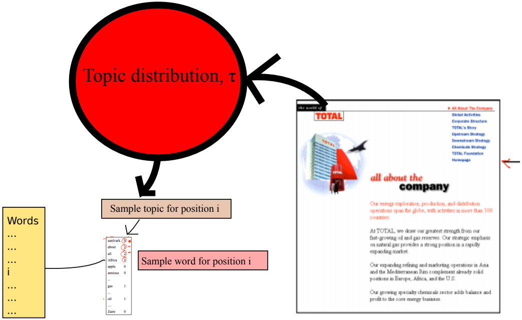


Topic Modelling



Task

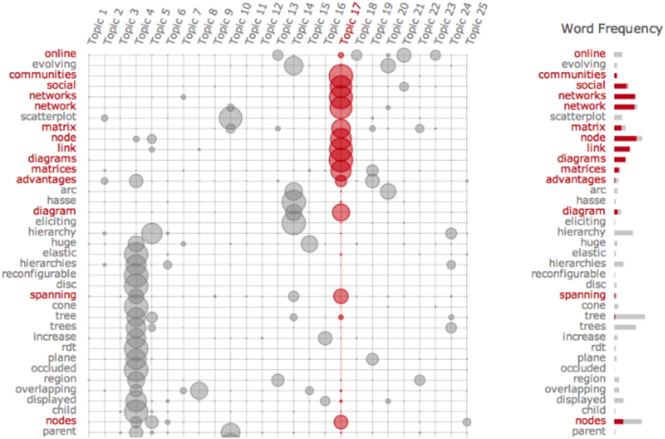
- Assume documents are generated by probabilistic model with unknown variables
- Infer hidden structure onto document
- Situate new document into model



TODO: Redo pic

Word-topic Matrix

Extracted: Word-topic matrix



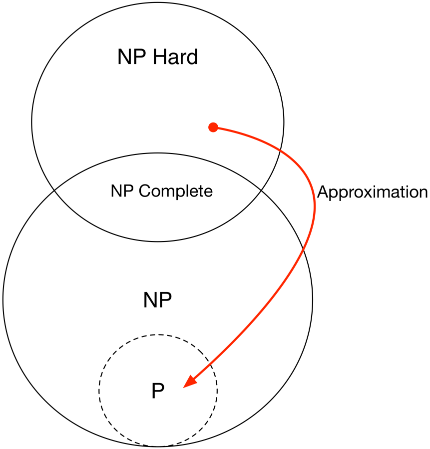
Aim: Find document-topic matrix

Anchor Words

- Word-topic distributions are separable
- There is a word unique to each topic
- Indicates document is partially about that topic
- Can learn parameters in polynomial time provided there is a large enough number of documents

Approximate Inference & Provable Guarantees

- Document-topic inference:
    - NP-hard
- Approximate techniques
- Provably polynomial-time?



Algorithm

Input: Corpus $\mathcal{D}$, Number of topics K

Output: Word-topic matrix A, topic-topic matrix R

1. Compute word-word co-occurrence matrix
2. Normalize the matrix
3. Find anchor words
4. Recover topics

Assumptions:

- Topics may be correlated
- Word-topic distributions are separable

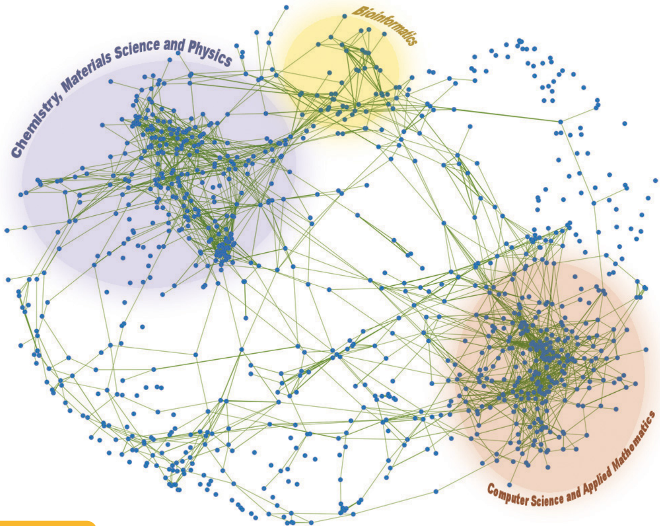
Contributions

1. Anchor Selection
   - Combinatorial rather than ILP
   - Stable in the presence of noise
   - polynomial sample complexity
2. Recovery step
   - Previous matrix-inversion approach sensitive to noise
   - Replaced with Gradient-based inference
3. Empirical comparison of algorithms

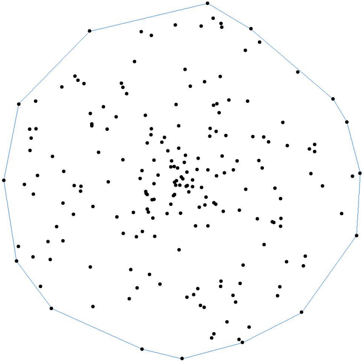
Word-word co-occurrence matrix

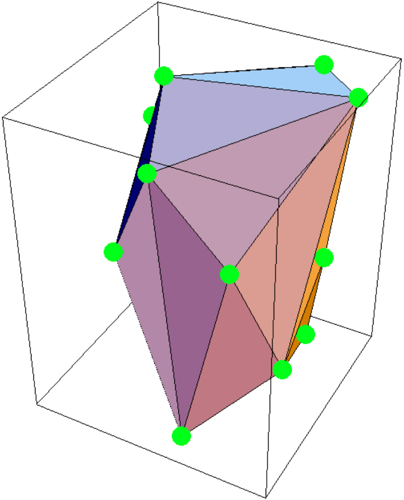
	bank	California	Canada	career	careers	employers	employment	federal	human	job	jobs	listings	openings	opportunities	positions	recruiters	resources	resume	resumes	retirement	search	state	texas	unemployment	work
bank	--																								
California		--																							
Canada	1		--																						
career	3	3		--																					
careers			2	9	--																				
employers		2		11	7	--																			
employment	3	26	22	66	10	16	--																		
federal	1	1	5		1		11	--																	
human		4	12	1	1			4	--																
job	34	14	2	49	8	13	92	13	2	--															
jobs		18	6	62	11	27	204	19	2	74	--														
listings		4	2	15	4	9	68	2	55	44		--													
openings		4		7	2	9	28			49	30		--												
opportunities	4	8	3	51	9	13	181	9		84	106	25	19	--											
positions		1		8	2	10	19			16	20	9	13	21	--										
recruiters				10	4	3	9			5	4	2	2	5	2	--									
resources		4	12		1			4	74	3	2						--								
resume		4	3	5		2	3	1	1	10	3			1	2	1		--							
resumes				8	3	3	11			5	16	1		8	5			15	--						
retirement		1	1						3			2		1						--					
search			3	4	6			10			18	6		6	2			3	1		--				
state			4	1			18		1	12	7	6		3			1			2		--			
texas	2			1			18			12	6		1	2						9			--		
unemployment																						2	2	--	
work			2	1		3	3	2		2	8	2	4	7	5				1			1	2		--

Words as vertices



Convex Hull





Computing Convex Hull

- Efficient for 2 dimensions - O($n \log n$)
- Inefficient for $n > 2$ dimensions
- Complexity depends on method and approximation used
- Previous method: ILP
- New method: Recursive greedy
  1. Compute subspace span of current convex hull
  2. Find point furthest from this sub-span
  3. Add point to convex hull
  4. Repeat until K points found

TODO: Work out how the whole convex hull - words as vertices work. I think what we have here is wrong because there is no approximation

Topic Recovery Task

- Recovers the topics
- Represented as topic-word distributions
- Topic uniquely identified by anchor word

Previous method

1. Discard rows not containing anchor words from word-word co-occurrence matrix (Q)
2. Permute matrix into a diagonal matrix
3. We know that $Q = ARA^T$ where A is the word-topic matrix and R is the topic-topic matrix
4. Solve $Q = ARA^T$ using matrix inversion

Word-word co-occurrence probability matrix

For an anchor word (row) in the co-ocurrence matrix

$$Q_{s_k,j} = \sum_{k'} p(z_1 = k'|w_1 = s_k) p(w_2 = j|z_1 = k')$$

= 1 because of the anchor word property

$$= p(w_2 = j|z_1 = k) = C_{i,k}$$

For any other row

$$\bar{Q}_{i,j} = \sum_{k} p(z_1 = k|w_1 = i) p(w_2 = j|z_1 = k)$$

But this is clearly a convex combination of anchor words

$$\bar{Q}_{i,j} = \sum_{k} C_{i,k} \bar{Q}_{s,k}$$

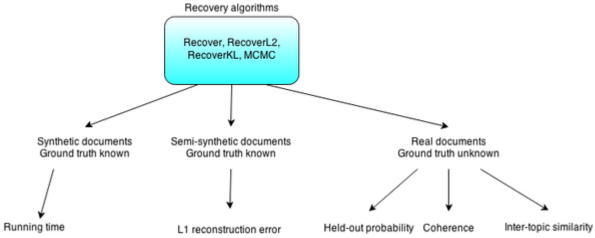
New method

- Row normalize Q into $\bar{Q}$
- Recover A and R using Bayes rule

$$p(w_1 = i | z_1 = k) = \frac{p(z_1 = k | w_1 = i) p(w_1 = i)}{\sum_i p(z_1 = k | w_1 = i') p(w_1 = i')}$$

1. Introduction
2. Topic Modelling
3. Algorithm
4. Efficiently Finding Anchor Words
5. Topic Recovery via Bayes' Rule
6. Experimental Results
7. Conclusion

Experiments



Metrics

- Ground truth known
  - Reconstruction error
- Ground truth unknown
  - Held out probability - probability of an unseen document under the learned model
  - Coherence - measure of semantic quality
  - Inter-topic similarity - measure of uniqueness of topics

Results - Efficiency

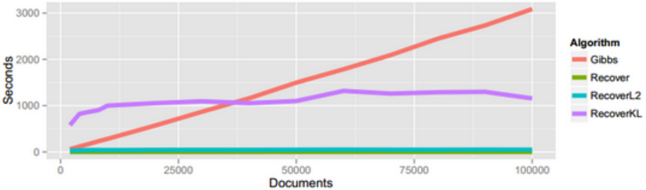


Figure 1. Training time on synthetic NIPS documents.

Results - Reconstruction error

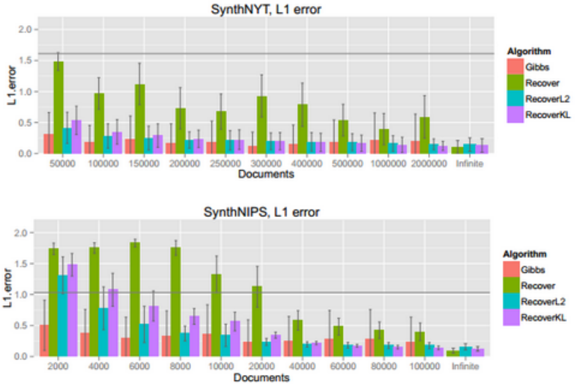


Figure 2. ℓ_1 error for learning semi-synthetic LDA models with $K = 100$ topics (**top**: based on NY Times, **bottom**: based on NIPS abstracts). The horizontal lines indicate the ℓ_1 error of K uniform distributions.

Results - Held out probability, coherence and unique words

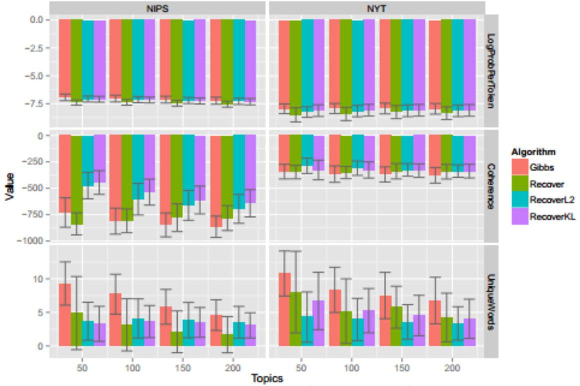


Figure 5. Held-out probability (per token) is similar for RecoverKL, RecoverL2, and Gibbs sampling. RecoverKL and RecoverL2 have better coherence, but fewer unique words than Gibbs. (Up is better for all three metrics.)

Summary

- New algorithms for topic recovery have been presented
  - Empirical results compare to MCMC
  - Uses anchor word assumption and Bayes' rule for theoretical and empirical performance improvements
- Simple to implement, maintains provable guarantees
- Attractive feature: *Running time independent of the size of the corpus*

Future Work

- Use output of algorithms for further optimization
- Parallel implementations

Comments

- Main contribution: algorithms for topic inference with provable guarantees for running time
  - Effect: running time effectively independent of corpus size

Thanks!

Any questions please email either of us:

Vanush Vaswani
vvas9619@uni.sydney.edu.au

Kristy Hughes
khug2372@uni.sydney.edu.au