Efficient Topic Modeling on Phrases via Sparsity

Weijing Huang, Wei Chen, Tengjiao Wang, Shibo Tao

School of Electronics Engineering and Computer Science, Peking University, China

@ICTAI 2017, Boston, US

Topic Modeling: An exploring tool to overview a big text corpus.

Before

Figure: This picture is from http://www.anex-uk.co.uk/, to illustrate the information overload.

Topic Modeling: An exploring tool to overview a big text corpus.

After (Ideally)

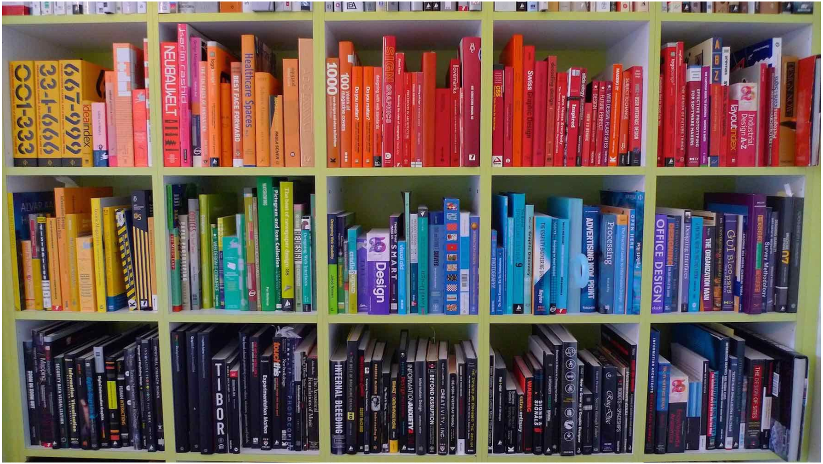


Figure: This picture is from `https://themorningnews.org`, to illustrate the organized information.

sci.electronics/52434: ... He details methods of underground and underwater wireless communications

...

sci.electronics/54237: ... Being in the "visualization" stage of a circuit design right now, I was wondering what the best way to "scale" an AC voltage ...

...

Raw Text Corpus

sci.electronics/52434: ... He details methods of underground and underwater wireless communications

...

sci.electronics/54237: ... Being in the "visualization" stage of a circuit design right now, I was wondering what the best way to "scale" an AC voltage ...

...

Raw Text Corpus

$\Rightarrow$

Topical Words: power, ground, wire, current, circuit, use, heat, used, radio, signal, cable, supply, water, voltage, high, hot, light, wiring, input, noise

Topic Modeling on Words

sci.electronics/52434: ... He details methods of underground and underwater wireless communications

...

sci.electronics/54237: ... Being in the "visualization" stage of a circuit design right now, I was wondering what the best way to "scale" an AC voltage ...

...

Raw Text Corpus

⇒

Topical Words: power, ground, wire, current, circuit, use, heat, used, radio, signal, cable, supply, water, voltage, high, hot, light, wiring, input, noise

Topic Modeling on Words

⇒

Topical Phrases: grounding conductor, ground wire, aluminum wiring, work well, neutral ground, main panel, house wiring, make sure, check local, neutral wire, electrical code

Topic Modeling on Phrases

Topic Modeling on Words

- LDA [Blei, JMLR 2003]
- very useful for understanding text corpus by providing interpretable topics

Topic Modeling on Phrases

- PhraseLDA [El-Kishky, VLDB 2014]
- provides more interpretable topics

Topic Modeling on Words

- LDA [Blei, JMLR 2003]
- very useful for understanding text corpus by providing interpretable topics

Topic Modeling on Phrases

- PhraseLDA [El-Kishky, VLDB 2014]
- provides more interpretable topics

Efficient Implementation

- SparseLDA [Yao, SIGKDD 2009], AliasLDA [Li, SIGKDD 2014], LightLDA [Yuan, WWW 2015]

Topic Modeling on Words

- LDA [Blei, JMLR 2003]
- very useful for understanding text corpus by providing interpretable topics

Topic Modeling on Phrases

- PhraseLDA [El-Kishky, VLDB 2014]
- provides more interpretable topics

Efficient Implementation

- SparseLDA [Yao, SIGKDD 2009], AliasLDA [Li, SIGKDD 2014], LightLDA [Yuan, WWW 2015]

Efficient Implementation?

- Our target

What about directly run LDA on phrases?

What about directly run LDA on phrases? OK, but not the best solution.

What about directly run LDA on phrases? OK, but not the best solution.

1. phrases are much less than words. [Tang, ICML 2014] points out *"pool performance of the LDA is expected when documents are too short"*.
2. works not well on short documents, *e.g.* , tweets, paper titles.

What about directly run LDA on phrases? OK, but not the best solution.

1. phrases are much less than words. [Tang, ICML 2014] points out *"pool performance of the LDA is expected when documents are too short"*.

2. works not well on short documents, *e.g.* , tweets, paper titles.

We solve this problem based on words and phrases together.

Our solution

SparseTP: using the sparsity of words and phrases to speed up the topic modeling on phrases.

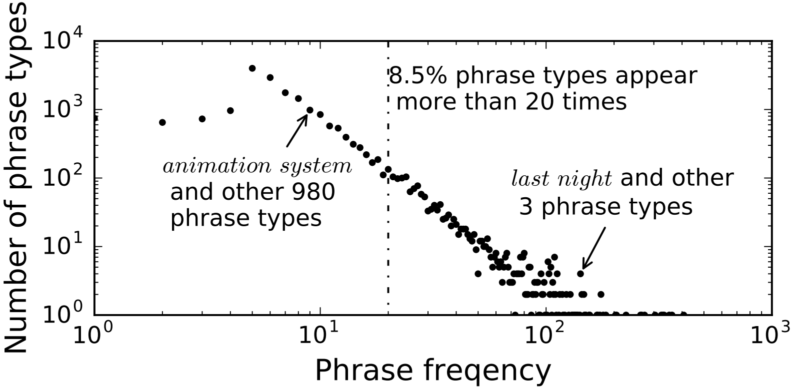


Figure: Distribution of phrase frequency (on *20 Newsgroups*).

Our solution

SparseTP: using the sparsity of words and phrases to speed up the topic modeling on phrases.

Sparsity: the ratio of zeros in the model (proposed by SparseLDA[Yao, KDD 2009]).

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
	…	…	…	…	…	…	…	…	…	…	…
animation system	1	1	1	1	1	1	1	1	1	0	…
	…	..	…	…	..	…	…	…	…	…	…

Figure: Sparsity of "*animation system*" in the topic-phrase matrix

Our solution

`SparseTP`: using the sparsity of words and phrases to speed up the topic modeling on phrases.

Sparsity: the ratio of zeros in the model (proposed by SparseLDA[Yao, KDD 2009]). We extend this conception to phrases.

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
	...	...	...	...	...	...	...	...	...	...	...
animation system	1	1	1	1	1	1	1	1	1	0	...
	...	..	...	...	..	...	...	...	...	...	...

Figure: Sparsity of "*animation system*" in the topic-phrase matrix

Our solution: workflow

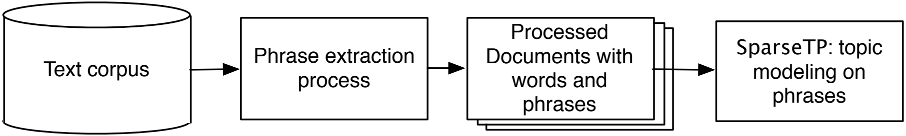
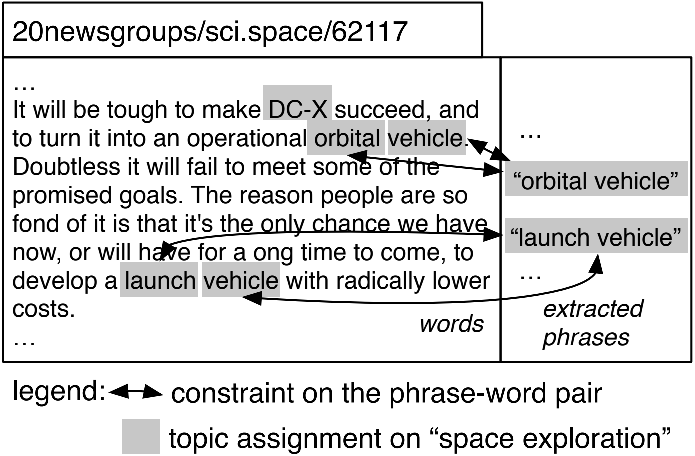
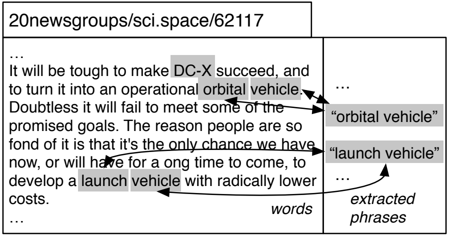


Figure: workflow

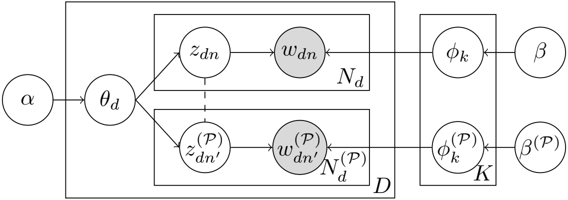
Double counting phrases: words in phrases, and phrases themselves.

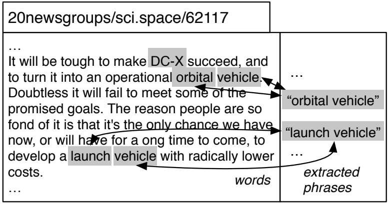


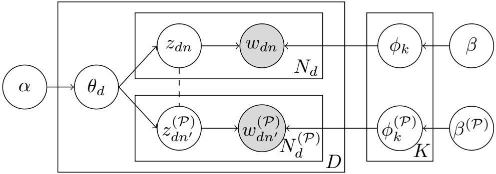
Double counting phrases: words in phrases, and phrases themselves.



$\Downarrow$







In Markov Random Field, the phrases and their component words tend to have similar topic assignemnts.

1. Edge potential function: $\exp\{\mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j})/|I(d,i)|\}$
2. Unary potential function: $p(z_{d,j} = k|\theta_d) = \theta_{d,k}$
$$p(z_{d,i}^{(\mathcal{P})} = k|\theta_d) = \theta_{d,k}$$

In Markov Random Field,

1. Edge potential function: $\exp\{\mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j})/|l(d,i)|\}$
2. Unary potential function: $p(z_{d,j} = k|\theta_d) = \theta_{d,k}$
$$p(z_{d,i}^{(\mathcal{P})} = k|\theta_d) = \theta_{d,k}$$

The joint distribution:

$$\begin{aligned} p(z_d, z_d^{(\mathcal{P})}|\theta_d) &= \frac{1}{A_d(\theta_d)} \prod_{m=1}^{N_d} p(z_{d,m}|\theta_d) \\ &\cdot \prod_{i=1}^{N_d^{(\mathcal{P})}} p(z_{d,i}^{(\mathcal{P})}|\theta_d) \cdot \exp\{\sum_{i=1}^{N_d}(\frac{1}{l(d,i)} \sum_{j \in l(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\} \end{aligned} \tag{1}$$

In Markov Random Field,

1. Edge potential function: $\exp\{\mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j})/|l(d,i)|\}$
2. Unary potential function: $p(z_{d,j} = k|\theta_d) = \theta_{d,k}$
   $p(z_{d,i}^{(\mathcal{P})} = k|\theta_d) = \theta_{d,k}$

The joint distribution:

$$
\begin{aligned}
p(z_d, z_d^{(\mathcal{P})}|\theta_d) &= \frac{1}{A_d(\theta_d)} \prod_{m=1}^{N_d} p(z_{d,m}|\theta_d) \\
&\cdot \prod_{i=1}^{N_d^{(\mathcal{P})}} p(z_{d,i}^{(\mathcal{P})}|\theta_d) \cdot \exp\{\sum_{i=1}^{N_d}(\frac{1}{l(d,i)} \sum_{j \in l(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\}
\end{aligned}
\tag{1}
$$

In (1), $A_d(\theta_d)$ is the partition function, causing the inference intractable.

$$
\begin{aligned}
A_d(\theta_d) = &\sum_{z_d} \sum_{z_d^{(\mathcal{P})}} (\prod_{m=1}^{N_d} p(z_{d,m}|\theta_d) \prod_{i=1}^{N_d^{(\mathcal{P})}} p(z_{d,i}^{(\mathcal{P})}|\theta_d) \\
&\cdot \exp\{\sum_{i=1}^{N_d^{(\mathcal{P})}}(\frac{1}{|l(d,i)|} \sum_{j \in l(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\})
\end{aligned}
\tag{2}
$$

Looking for the well-formed lower bound of the model.

1. The upper bound of $A_d(\theta_d)$

$\because 0 \leq p(z_{d,i}^{(\mathcal{P})}|\theta_d), p(z_{d,m}|\theta_d) \leq 1$

$\therefore A_d(\theta_d) \leq \sum_{z_d} \sum_{z_d^{(\mathcal{P})}} \exp\{\sum_{i=1}^{N_d^{(\mathcal{P})}} (\frac{1}{|l(d,i)|} \sum_{j \in l(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\}$

$\quad \stackrel{\text{def}}{=} \tilde{A}_d$

Looking for the well-formed lower bound of the model.

1. The upper bound of $A_d(\theta_d)$

   $\because 0 \leq p(z_{d,i}^{(\mathcal{P})}|\theta_d), p(z_{d,m}|\theta_d) \leq 1$

   $\therefore A_d(\theta_d) \leq \sum_{z_d} \sum_{z_d^{(\mathcal{P})}} \exp\{\sum_{i=1}^{N_d^{(\mathcal{P})}} (\frac{1}{|l(d,i)|} \sum_{j \in l(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\} \stackrel{\text{def}}{=} \tilde{A}_d$

2. The lower bound of the model (complete likelihood).

$$P(Z, Z^{(\mathcal{P})}, W, W^{(\mathcal{P})}|.) \geq \prod_{d=1}^{D} \frac{\exp\{\sum_{i=1}^{N_d^{(\mathcal{P})}} (\frac{1}{l(d,i)} \sum_{j \in l(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\}}{\tilde{A}_d}$$

$$\cdot \prod_{d=1}^{D} \frac{B(\vec{n}_{d,.} + \vec{n}_{d,.}^{(\mathcal{P})} + \vec{\alpha})}{B(\vec{\alpha})} \prod_{k=1}^{K} \frac{B(\vec{n}_{k,.} + \beta)}{B(\vec{\beta})} \prod_{k=1}^{K} \frac{B(\vec{n}_{k,.}^{(\mathcal{P})} + \beta)}{B(\vec{\beta})}$$

Looking for the well-formed lower bound of the model.

1. The upper bound of $A_d(\theta_d)$
   $\because 0 \leq p(z_{d,i}^{(\mathcal{P})}|\theta_d), p(z_{d,m}|\theta_d) \leq 1$
   $\therefore A_d(\theta_d) \leq \sum_{z_d} \sum_{z_d^{(\mathcal{P})}} \exp\{\sum_{i=1}^{N_d^{(\mathcal{P})}} (\frac{1}{|I(d,i)|} \sum_{j \in I(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\}$
   $\stackrel{\text{def}}{=} \tilde{A}_d$
2. The lower bound of the model (complete likelihood).

$$P(Z, Z^{(\mathcal{P})}, W, W^{(\mathcal{P})}|.) \geq \prod_{d=1}^{D} \frac{\exp\{\sum_{i=1}^{N_d^{(\mathcal{P})}} (\frac{1}{I(d,i)} \sum_{j \in I(d,i)} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\}}{\tilde{A}_d}$$

$$\cdot \prod_{d=1}^{D} \frac{B(\vec{n}_{d,.} + \vec{n}_{d,.}^{(\mathcal{P})} + \vec{\alpha})}{B(\vec{\alpha})} \prod_{k=1}^{K} \frac{B(\vec{n}_{k,.} + \beta)}{B(\vec{\beta})} \prod_{k=1}^{K} \frac{B(\vec{n}_{k,.}^{(\mathcal{P})} + \beta)}{B(\vec{\beta})}$$

It's easy to do the Gibbs Sampling on the lower bound.

Inference the model by Gibbs Sampling,

1. For phrase,

$$p(z^{(\mathcal{P})}_{d,i} = k | Z^{(\mathcal{P})}_{\neg(d,i)}, Z, W^{(\mathcal{P})}, W, \alpha, \beta, \beta^{(\mathcal{P})}) \propto (n_{d,k} + n^{(\mathcal{P})}_{d,k} + \alpha_k) \cdot \frac{n^{(\mathcal{P})}_{k,w^{(\mathcal{P})}_{d,i}} + \beta^{(\mathcal{P})}_k}{n^{(\mathcal{P})}_{k,\cdot} + \sum_{k=1}^{K} \beta^{(\mathcal{P})}_k} \boxed{\exp\{\frac{1}{|I(d,i)|} \sum_{j \in I(d,i)} \mathbb{I}(z_{d,j} = k)\}}$$

Introduced by Markov Random Field

2. For component words within phrases,

$$p(z_{d,j} = k | Z_{\neg(d,j)}, Z, W^{(\mathcal{P})}, W, \alpha, \beta, \beta^{(\mathcal{P})}) \propto (n_{d,k} + n^{(\mathcal{P})}_{d,k} + \alpha_k) \cdot \frac{n_{k,w_{d,j}} + \beta_k}{n^{(\mathcal{P})}_{k,\cdot} + \sum_{k=1}^{K} \beta_k} \boxed{\exp\{\frac{1}{|I(d,i)|} \mathbb{I}(z^{(\mathcal{P})}_{d,i} = k)\}}$$

Introduced by Markov Random Field

3. For words not in phrases, same as the Gibbs Sampling in original LDA.

Use sparsity to decouple the Gibbs Sampling equation

Take component words as an example,

$$p(z_{d,j} = k | Z_{\neg(d,j)}, Z, W^{(\mathcal{P})}, W, \alpha, \beta, \beta^{(\mathcal{P})}) \propto (n_{d,k} + n_{d,k}^{(\mathcal{P})} + \alpha_k)$$
$$\cdot \frac{n_{k,w_{d,j}} + \beta_k}{n_{k,\cdot}^{(\mathcal{P})} + \sum_{k=1}^{K} \beta_k} \exp\{\frac{1}{|l(d,i)|} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = k)\}$$

The ranking of terms according to the sparsity: (very sparse), $n_{k,v}$, $n_{d,k} + n_{d,k}^{(\mathcal{P})}$, α_k, (very dense).

Use sparsity to decouple the Gibbs Sampling equation

Take component words as an example,

$$p(z_{d,j} = k | Z_{\neg(d,j)}, Z, W^{(\mathcal{P})}, W, \alpha, \beta, \beta^{(\mathcal{P})}) \propto (n_{d,k} + n_{d,k}^{(\mathcal{P})} + \alpha_k) \cdot \frac{n_{k,w_{d,j}} + \beta_k}{n_{k,\cdot}^{(\mathcal{P})} + \sum_{k=1}^{K} \beta_k} \exp\{\frac{1}{|I(d,i)|} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = k)\}$$

The ranking of terms according to the sparsity: (very sparse), $n_{k,v}$, $n_{d,k} + n_{d,k}^{(\mathcal{P})}$, α_k, (very dense).

⇓ Decouple (Inspired by SparseLDA [Yao, KDD 2009]):

$$p(z_{d,j} = k | Z_{\neg(d,j)}, Z, W^{(\mathcal{P})}, W, \alpha, \beta, \beta^{(\mathcal{P})}) \propto s_k + r_k + q_k + t_k \tag{3}$$

Use sparsity to decouple the Gibbs Sampling equation

$\Downarrow$ Decouple (Inspired by SparseLDA [Yao, KDD 2009]):

$$p(z_{d,j} = k | Z_{\neg(d,j)}, Z, W^{(\mathcal{P})}, W, \alpha, \beta, \beta^{(\mathcal{P})})$$
$$\propto s_k + r_k + q_k + t_k$$

Table: The meaning of four parts (s_k,r_k can be maintained efficiently, q_k is determined by $n_{k,v}$, and t_k is very sparse)

Smoothing-only	$s_k = \frac{\alpha_k \beta_k}{n_{k,.} + \sum_{k=1}^{K} \beta_k}$
Document-topic	$r_k = \frac{(n_{d,k} + n_{d,k}^{(\mathcal{P})}) \beta_k}{n_{k,.} + \sum_{k=1}^{K} \beta_k}$
Topic-word	$q_k = \frac{n_{d,k} + n_{d,k}^{(\mathcal{P})} + \alpha_k}{n_{k,.} + \sum_{k=1}^{K} \beta_k} n_{k,v}$
MRF-guided	$t_k = (s_k + r_k + q_k)$ $(\exp\{\frac{1}{\lvert l(d,i)\rvert} \mathbb{I}(z_{d,i}^{(\mathcal{P})} = k)\} - 1)$

Use sparsity to decouple the Gibbs Sampling equation

Toy example:

1. The maintenance of s, r, q, and t are sublinear to K.
2. The sampling of $z_{d,j}$ is sublinear to K (because t and q occupy the major part).

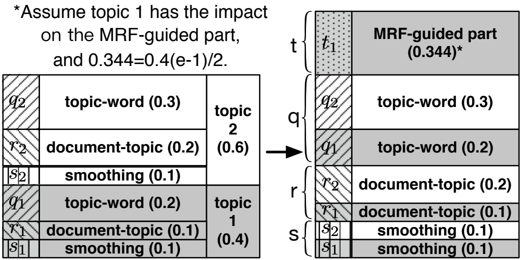


Figure: (1) s, r, q: reordering for using sparsity in LDA's sampling, introduced by SparseLDA, and also available for `SparseTP`; (2) t: additional part for using sparsity in `SparseTP`'s sampling.

Experiments

Data sets

Table: The statistics of the datasets. In average, each phrase type appears more sparse than each word type.

	$\|V\|$	$\|V^{(\mathcal{P})}\|$	$\|W\|$	$\|W^{(\mathcal{P})}\|$	$\|D\|$	$\|W\|/\|V\|$	$\|W^{(\mathcal{P})}\|/\|V^{(\mathcal{P})}\|$	$\|W\|/\|D\|$	$\|W^{(\mathcal{P})}\|/\|D\|$
20 Newsgroup	90,478	19,471	1,893,750	192,614	18,828	20.9	9.9	100.6	10.2
Argentina@Wiki	75,182	21,182	1,562,089	231,380	8,617	20.8	10.9	181.3	26.9
Mathematics@Wiki	174,211	116,012	8,310,353	1,471,464	27,947	47.7	12.7	297.4	52.6
Chemistry@Wiki	338,328	248,768	16,656,777	3,322,555	60,375	49.2	13.4	275.9	55.0
PubMed Abstracts	154,559	239,501	13,960,094	3,479,392	99,214	90.3	14.5	140.7	35.1
Twitter	3,045,146	901,719	204,867,603	40,130,035	27,576,077	67.3	44.5	7.4	1.5

Experiments

Data sets

Table: The statistics of the datasets. In average, each phrase type appears more sparse than each word type.

	$\|V\|$	$\|V^{(\mathcal{P})}\|$	$\|W\|$	$\|W^{(\mathcal{P})}\|$	$\|D\|$	$\|W\|/\|V\|$	$\|W^{(\mathcal{P})}\|/\|V^{(\mathcal{P})}\|$	$\|W\|/\|D\|$	$\|W^{(\mathcal{P})}\|/\|D\|$
20 Newsgroup	90,478	19,471	1,893,750	192,614	18,828	20.9	9.9	100.6	10.2
Argentina@Wiki	75,182	21,182	1,562,089	231,380	8,617	20.8	10.9	181.3	26.9
Mathematics@Wiki	174,211	116,012	8,310,353	1,471,464	27,947	47.7	12.7	297.4	52.6
Chemistry@Wiki	338,328	248,768	16,656,777	3,322,555	60,375	49.2	13.4	275.9	55.0
PubMed Abstracts	154,559	239,501	13,960,094	3,479,392	99,214	90.3	14.5	140.7	35.1
Twitter	3,045,146	901,719	204,867,603	40,130,035	27,576,077	67.3	44.5	7.4	1.5

Baselines:

1. Plain LDA: directly runs LDA on extracted phrases of documents.
2. TNG [Wang, ICDM 2007]: combines the phrase extraction and the topic modeling in a unified model.
3. PhraseLDA [El-Kishky, VLDB 2014]: models words and extracted phrases together.

Efficiency

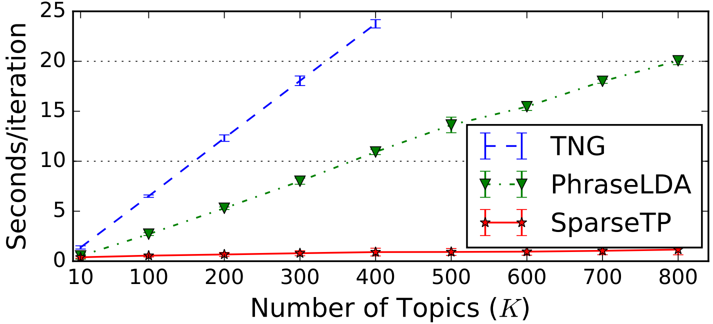


Figure: The comparison of the time efficiency of each methods' Gibbs sampling on 20 Newsgroups.

Effectiveness, evaluated by PMI.

Table: Four example phrase-topics on 20 Newsgroups, and evaluated by the PMI metric [Church, Computational Linguistics 1990].

Aligned Topic 1	Plain LDA	gun control, second amendment, self defense, keep bear arms, right people, united states, anti gun, well regulated militia, semi auto, gun owners	PMI=4.02
	TNG	in seattle, the gun, the nra, the media, to amend, in vancouver, hr house bill by mr, the weapon, the brady, gun control	PMI=3.87
	PhraseLDA	gun control, anti gun, death penalty, gun owners, violent crime, bill mr, gun control laws, crime rate, capital punishment, gun ownership	PMI=5.12
	SparseTP	self defense, gun control, semi auto, gun owners, death penalty, military weapon, dangerous ordnance, anti gun, self defense, semi autos	PMI=4.84
Aligned Topic 2	Plain LDA	— (*not detected*)	—
	TNG	— (*not detected*)	—
	PhraseLDA	first paragraph, world war ii, russian army, nazi armenians, second paragraph, world war, invading russian armies, terrorism revisionism triangle, nazi germany, took place	PMI=4.61
	SparseTP	nazi germany, nazi party, political power, commit genocide, many years, civil rights, european parliament, know nothing, right wing, nazi law	PMI=4.83
Aligned Topic 3	Plain LDA	kidney stones, per day, people think, side effects, grounding conductor, every time, much higher, best way, ground wire, news group	PMI=2.67
	TNG	the circuit, the voltage, the playback, the wire, the amp, the temperature, the pipes, the cec, the interference, the motor	PMI=5.93
	PhraseLDA	power supply, grounding conductor, ground wire, aluminum wiring, solder mask, neutral ground, house wiring, main panel, power cord, neutral wire, electrical code	PMI=6.55
	SparseTP	grounding conductor, ground wire, aluminum wiring, work well, neutral ground, main panel, house wiring, make sure, check local, neutral wire, electrical code	PMI=6.15
Aligned Topic 4	Plain LDA	hicnet medical newsletter, page volume, united states, hiv infection, lyme disease, public health, water dept, new york, health care, health insurance	PMI=5.11
	TNG	the patient, the researchers, the doctor, a patient, in spanish, a physician, the study, the medical, to percent, the health	PMI=5.39
	PhraseLDA	health care, private sector, health insurance, income tax, russian government, billion dollars, private insurance, onur yalcin, middle class, oil gas	PMI=5.34
	SparseTP	health care, health insurance, income tax, private insurance, spend money, make money, middle class, third party, insurance companies, state farm	PMI=5.49

Effectiveness, evaluated by PMI.

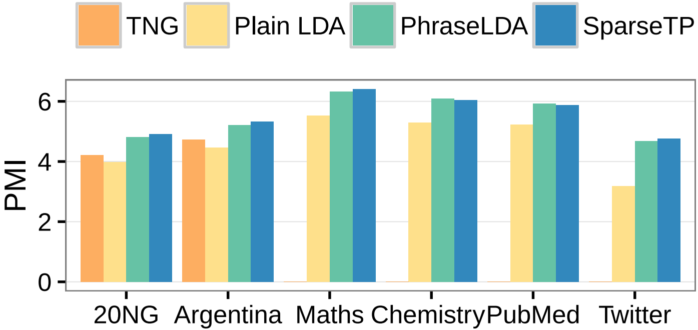


Figure: The quality of the learned topics on different datasets, evaluated by the PMI metric.

Effectiveness, evaluated manually.

Table: Overall Quality of Learned Topics on 20 Newsgroups, evaluated manually

	Plain LDA	TNG	PhraseLDA	SparseTP
#Interpretable Topics	29	34	58	**63**
Precision@30	0.20	0.22	0.48	**0.52**
sci.crypt	0.73	0.77	0.87	1.00
sci.med	0.60	0.70	0.87	0.87
sci.space	0.67	0.73	0.80	0.93
talk.politics.guns	0.83	0.50	0.90	0.90
talk.politics.mideast	0.00	0.00	0.77	1.00

SparseTP improves the efficiency without sacrificing the effectiveness.

Conclusions

We propose a novel topic model on phrases `SparseTP`, which

1. models the words and phrases by linking them in Markov Random Field when the word is a component of the phrase;
2. provides a well-formed lower bound of the model for Gibbs sampling;
3. utilizes the sparse distribution of words and phrases on topics to speed up the inference;
4. verifies the effectiveness and the efficiency by experiments.

Thanks!

Q&A