Neural Sparse Topic Model

xieq@whu.edu.cn

Outline

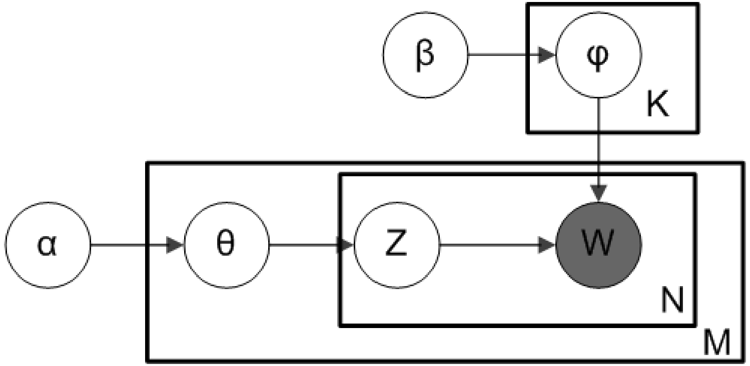
- Background
- Related work
- Our method

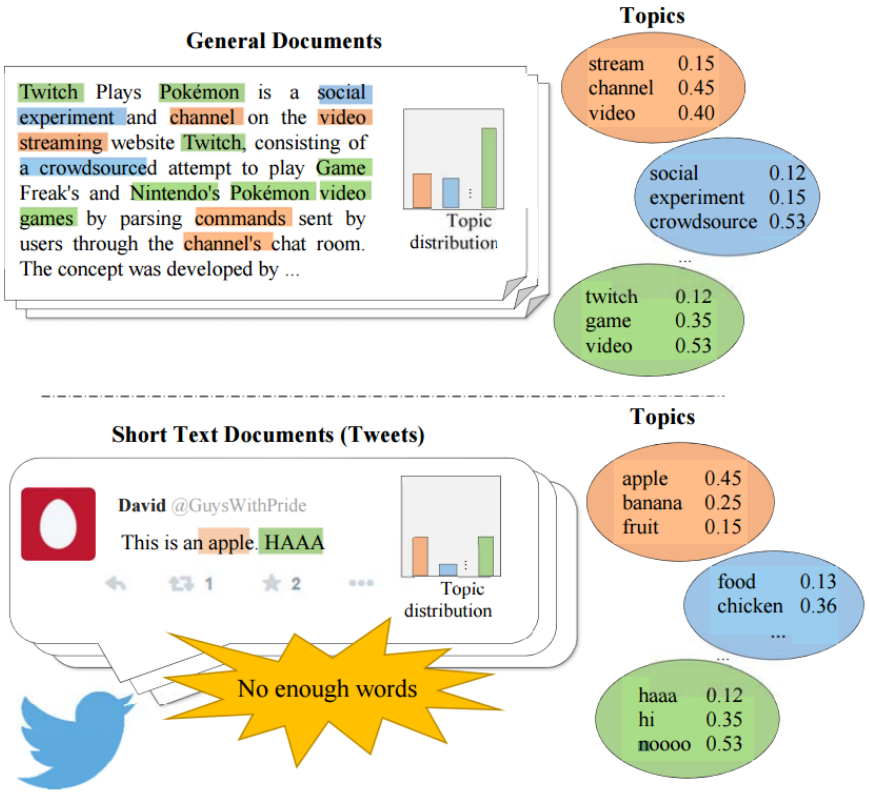
BACKGROUND

Traditional Topic Models

Unsupervised representation method

Over-complex inference procedure



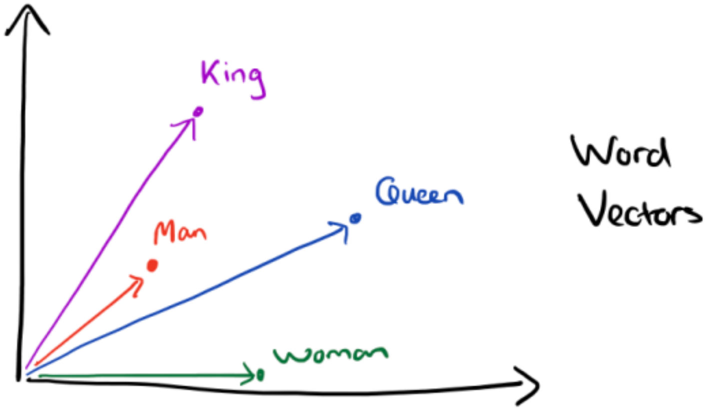


Relying on word occurrence information

RELATED WORK

Topic Models with Word Embeddings

Word semantic relations



Gaussian LDA for Topic Models with Word Embeddings (ACL 2015), Carnegie Mellon University

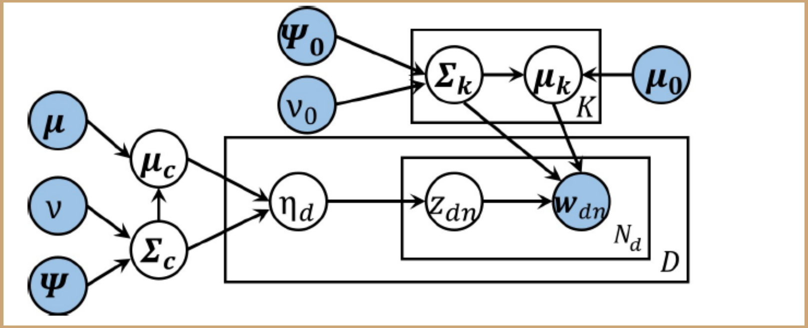
- Generating continues word vectors collapsed
- Gibbs sampling algorithm
- Group semantically related words into topics.

1. 对于每个主题 $k = 1, ..., K$:
   (a) 生成主题的协方差矩阵 $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\psi}, \nu)$;
   (b) 生成主题的均值 $\boldsymbol{\mu}_k \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\kappa}\boldsymbol{\Sigma}_k\right)$
2. 对每篇文档 $d = 1, ..., M$:
   (a) 生成文档主题分布 $\theta_d \sim \mathrm{Dir}(\alpha)$;
   (b) 对文档中的每个词 $i = 1, ..., N_d$:
      i. 生成词的主题 $z_{di} \sim \mathrm{Mult}(\theta_d)$;
      ii. 生成词向量 $\boldsymbol{v}_{di} \sim \mathcal{N}\left(\boldsymbol{\mu}_{z_{di}}, \boldsymbol{\Sigma}_{z_{di}}\right)$.

A Correlated Topic Model Using Word Embeddings (IJCAI 2017), Renmin University of China

- Exploit the additional word-level correlation information
- Directly model topic correlation in the continuous word embedding space

1. Draw $\boldsymbol{\Sigma}_c \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$.
2. Draw $\boldsymbol{\mu}_c \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\tau_c}\boldsymbol{\Sigma}_c)$.
3. For each Gaussian topic $k = 1, 2, \cdots, K$:
   (a) Draw topic covariance $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_0, \nu_0)$.
   (b) Draw topic mean $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{1}{\tau}\boldsymbol{\Sigma}_k)$.
4. For each document $d = 1, 2, \cdots, D$:
   (a) Draw $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.
   (b) For each word index $n = 1, 2, \cdots, N_d$:
      i. Draw a topic $z_{dn} \sim Multinomial(f(\boldsymbol{\eta}_d))$.
      ii. Draw a word $\boldsymbol{w}_{dn} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{dn}}, \boldsymbol{\Sigma}_{z_{dn}})$.

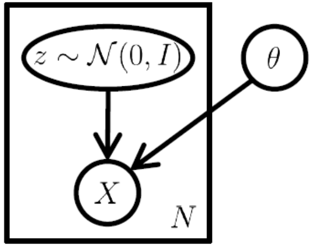


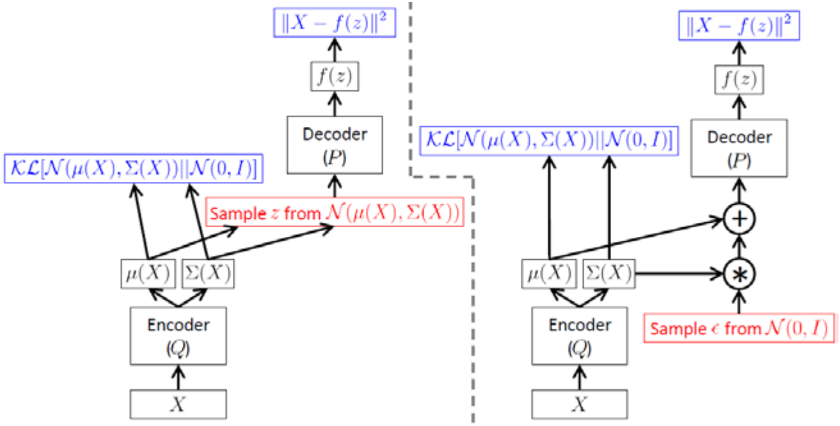
Neural Variational Inference for topic models

Deep generative models: VAE, GAN

Neural network + generative models

Training by BP





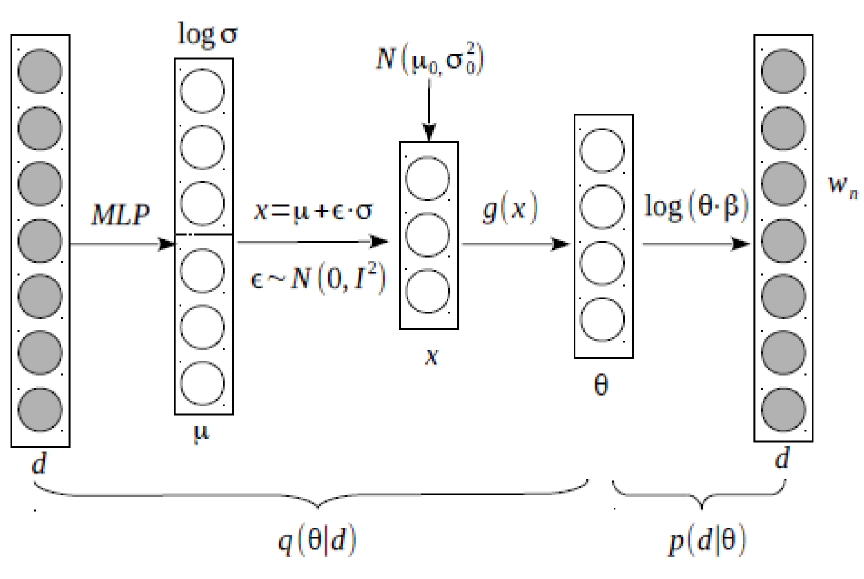
Autoencoding Variational Inference for Topic Models (ICLR 2017)

- AEVB based inference
- Deal with the component collapsing problem
- ProdLDA: a change of only one line of code

Model	Topics
ProdLDA	motherboard meg printer quadra hd windows processor vga mhz connector armenian genocide turks turkish muslim massacre turkey armenians armenia greek voltage nec outlet circuit cable wiring wire panel motor install season nhl team hockey playoff puck league flyers defensive player israel israeli lebanese arab lebanon arabs civilian territory palestinian militia
LDA NVLDA	db file output program line entry write bit int return drive disk get card scsi use hard ide controller one game team play win year player get think good make use law state health file gun public issue control firearm people say one think life make know god man see
LDA DMFVI	write article dod ride right go get night dealer like gun law use drug crime government court criminal firearm control lunar flyers hitter spacecraft power us existence god go mean stephanopoulos encrypt spacecraft ripem rsa cipher saturn violate lunar crypto file program available server version include software entry ftp use
LDA Collapsed Gibbs	get right back light side like see take time one list mail send post anonymous internet file information user message thanks please know anyone help look appreciate get need email jesus church god law say christian one christ day come bike dod ride dog motorcycle write article bmw helmet get
NVDM	light die burn body life inside mother tear kill christian insurance drug different sport friend bank owner vancouver buy prayer input package interface output tape offer component channel level model price quadra hockey slot san playoff jose deal market dealer christian church gateway catholic christianity homosexual resurrection modem mouse sunday

Discovering Discrete Latent Topics with Neural Variational Inference (ICML 2017)

- Gaussian Softmax
- Gaussian Stick-Breaking
- Recurrent Stick-Breaking
- Neural Topic Models
- Recurrent Neural Topic Models



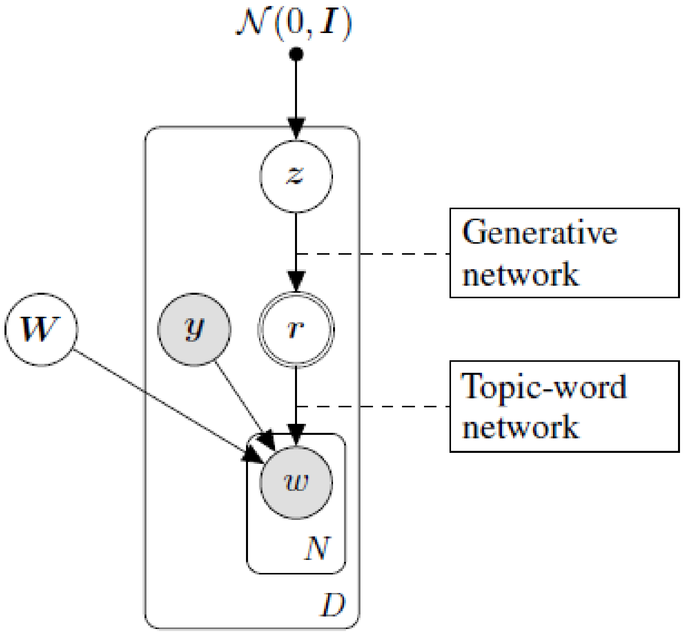
For each document i of length N_i:

(a) $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I})$

(b) $\boldsymbol{r}_i = f_g(\boldsymbol{z}_i)$

(c) For each word j in document i, $j = 1..., N_i$:

$$w_{ij} \sim p(w_{ij} \mid \boldsymbol{W}, \boldsymbol{r}_i),$$



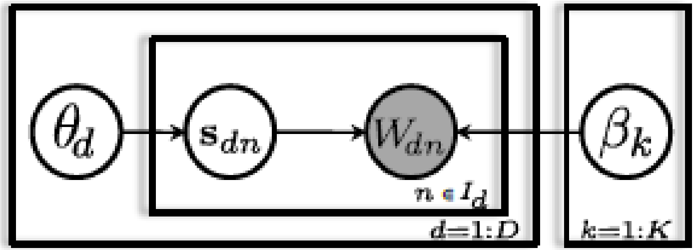
(a) Generative model

A Neural Framework for Generalized Topic Models

OUR METHOD

Sparse Topical Coding

1. Sample the document code θ from a prior $p(\theta) \sim Laplace(\lambda_1)$.
2. For each observed word n:
   (a) Sample the word code s_n from a conditional distribution $p(s_n|\theta) \sim supergaussian(\theta, \lambda_2)$.
   (b) Sample the observed word count w_n from a distribution $p(w_n|{s_n}^T \beta_n) \sim Poisson({s_n}^T \beta_n)$

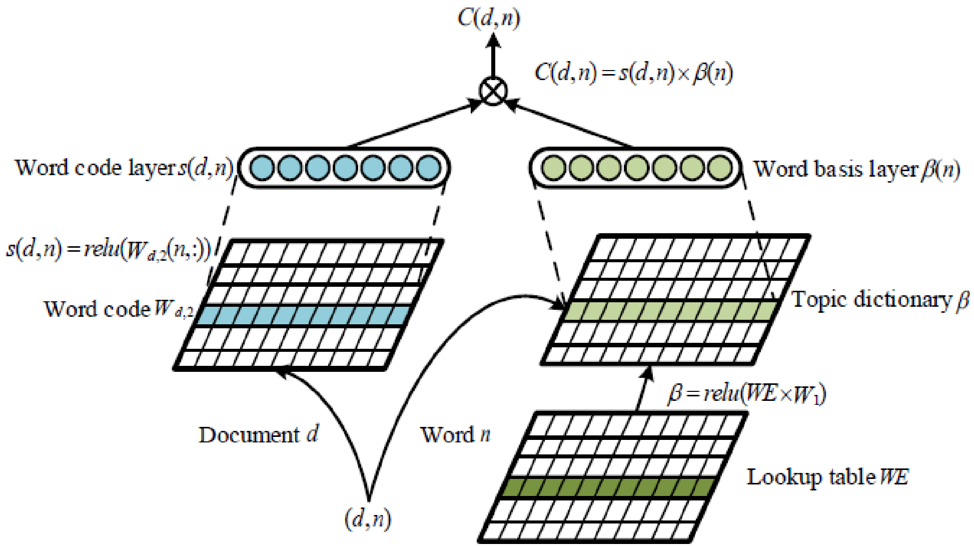


$$\min_{\Theta,\boldsymbol{\beta}} \sum_{d,n\in I_d} \ell(\mathbf{s}_{dn}, \boldsymbol{\beta}) + \lambda \sum_d \|\boldsymbol{\theta}_d\|_1 + \sum_{d,n\in I_d} (\gamma\|\mathbf{s}_{dn} - \boldsymbol{\theta}_d\|_2^2 + \rho\|\mathbf{s}_{dn}\|_1)$$

$$\text{s.t. : } \boldsymbol{\theta}_d \geq 0,\ \forall d;\ \mathbf{s}_{dn} \geq 0,\ \forall d, n \in I_d;\ \ \boldsymbol{\beta}_k \in \mathcal{P},\ \forall k, \qquad (2)$$

Neural Sparse Topical Coding (NSTC)

1. For each word n in document d:
   (a) Sample a latent variable word code $s_n \sim f_g(d, n)$.
   (b) Sample the observed word count w_n from $p(w_n|{s_n}^T\beta_n) \sim Poisson({s_n}^T\beta_n)$



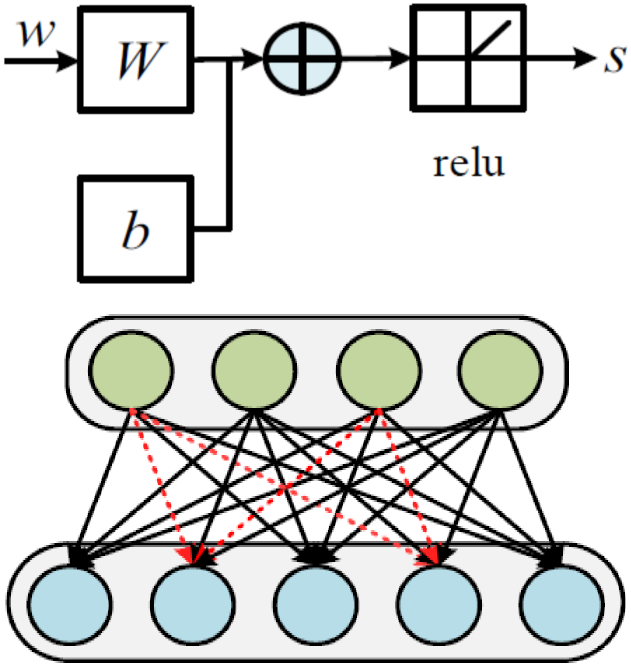
Extension: NSTCE

- Deep l1 encoder

$$F(w; W, b) = relu(W * w + b)$$

- Make the prediction of neural network predictor as close as possible to the optimal set of coefficients
- Jointly optimizing all parameters

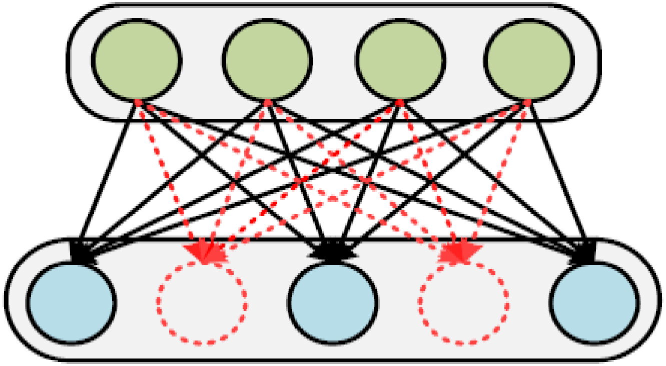
$$L = l(w_{d,n}, C(d,n)) + \lambda ||W_{d,2}||_1 + \alpha ||s(d) - F(w; W, b)||_2^2$$



Extension: NGSTC

- Group Sparse Regularization
- Make a neural network extension of GSTC

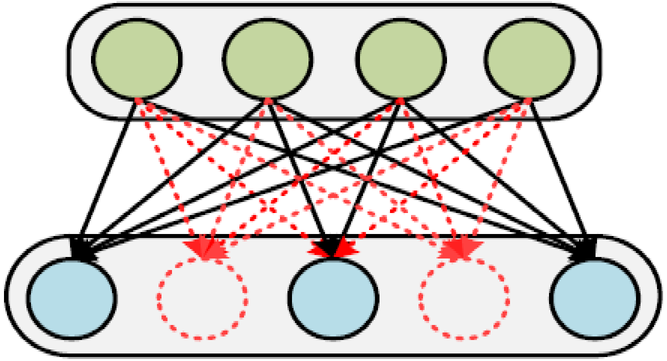
$$L = l(w_{d,n}, C(d,n)) + \lambda \sum_{k=1}^{K} ||W_{d,2}^{k}||_2$$



Extension: NSTCSG

- Sparse Group Lasso
- Make a neural network extension of STCSG

$$L = l(w_{d,n}, C(d,n)) + \lambda_1 ||W_{d,2}||_1 + \lambda_2 \sum_{k=1}^{K} ||W_{d,2}^{k}||_2$$



Training

- SGD: NSTC, NSTCE
- Proximal stochastic gradient descent (PSGD): NGSTC, NSTCSG
- Performing Euclidean projection of the intermediate solution via SGD on the loss:

$$\min_{W_{d,2}^{t+1}} R(W_{d,2}^{t+1}) + \frac{1}{2}||W_{d,2}^{t+1} - W_{d,2}^{t+\frac{1}{2}}||_2^2$$

- GL:

$$prox_{SGL}(W_{d,2}) = (1 - \frac{\lambda_2}{||sign(W_{d,2}^{k}, \lambda_1)||_2})_+ sign(W_{d,2}^{nk}, \lambda_1)$$

- SGL:

$$prox_{GL}(W_{d,2}) = (1 - \frac{\lambda}{||W_{d,2}^{k}||_2})_+ W_{d,2}^{nk}$$

EXPERIMENTS

DATA AND SETTINGS

Datasets

- 20Newsgroup: is comprised of 18775 newsgroup articles with 20 categories
- Web snippet: 12340 Web search snippets in 8 categories

Dataset	Label	Docs	Words	Vocabulary
Web Snippet	8	12265	10.72	5581
20Newsgroups	20	18775	135	60698

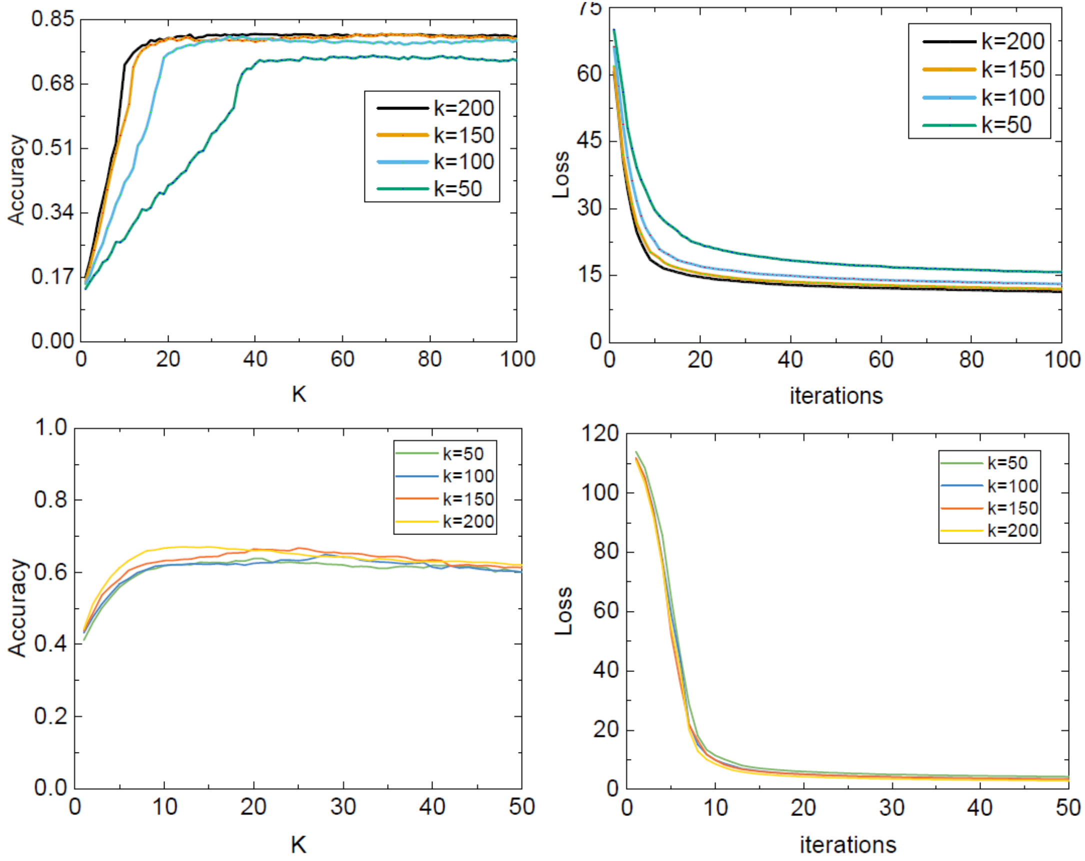
- LDA: $n = 200, \alpha = 0.1, \beta = 0.01.$
- STC: $\lambda = 0.3, \rho = 0.0001, n = 100.$
- DocNADE: the hidden size 50, the learning rate 0.0004 , the bath size 64, n=50000
- GLDA: default values for the parameters

BASELINES

SETTINGS

- ➢ 300-dimensional word embeddings by GloVe
- ➢ learning rate 0.0001
- ➢ OVW: normal distribution [0,1]

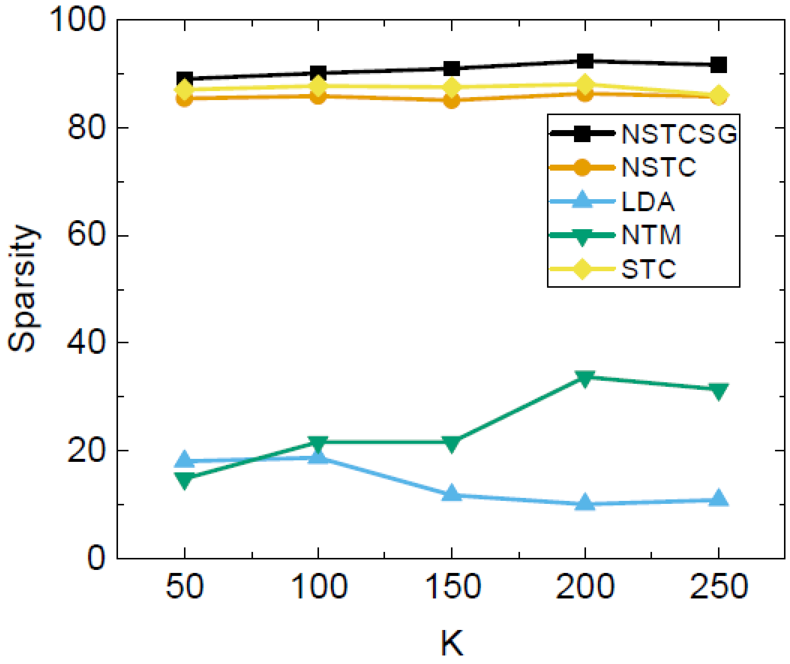
RESULTS

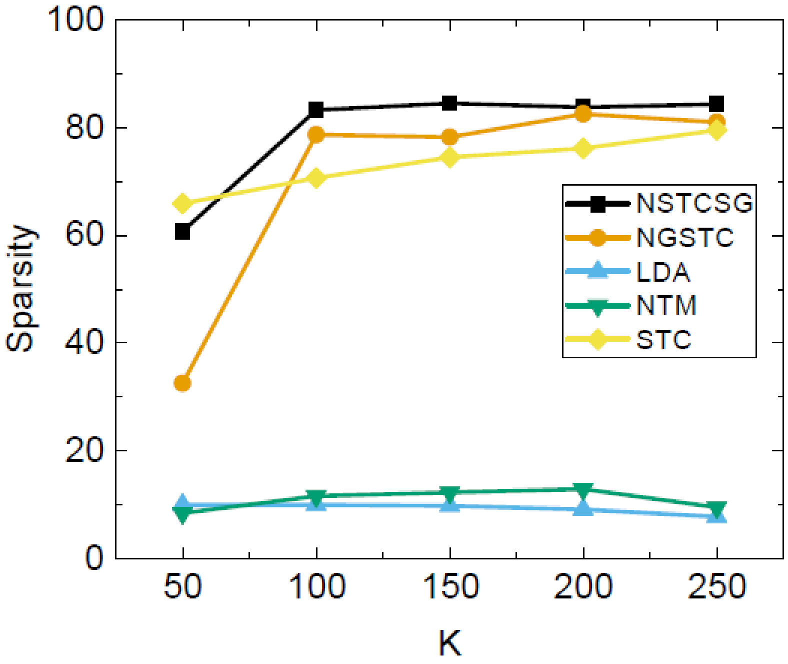


Classification accuracy

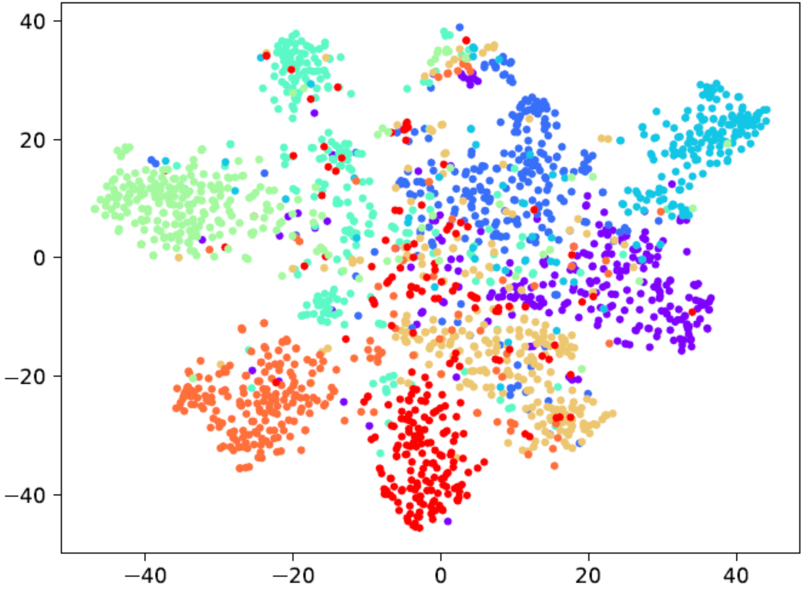
Dataset	Snippet					20NG				
k	**50**	**100**	**150**	**200**	**250**	**50**	**100**	**150**	**200**	**250**
LDA	0.682	0.592	0.573	0.615	0.583	0.545	0.615	0.607	0.613	0.623
STC	0.678	0.699	0.724	0.731	0.723	0.602	0.631	0.647	0.652	0.654
DocNADE	0.618	0.667	0.66	0.732	0.747	0.682	0.615	0.592	0.583	0.573
GLDA	0.618	0.667	0.66	0.732	0.747	[illegible]	0.615	0.592	0.583	[illegible]
NSTC	[illegible]	[illegible]	0.791	[illegible]	[illegible]	0.634	[illegible]	[illegible]	[illegible]	[illegible]
NSTCE	0.739	0.778	0.801	0.803	0.810	0.631	0.681	0.682	0.701	0.721
NGSTC	0.773	0.792	0.813	0.811	0.821	0.67	0.681	0.701	0.712	0.737
NSTCSG	0.788	0.813	0.821	0.823	0.829	0.665	0.687	0.691	0.717	0.735

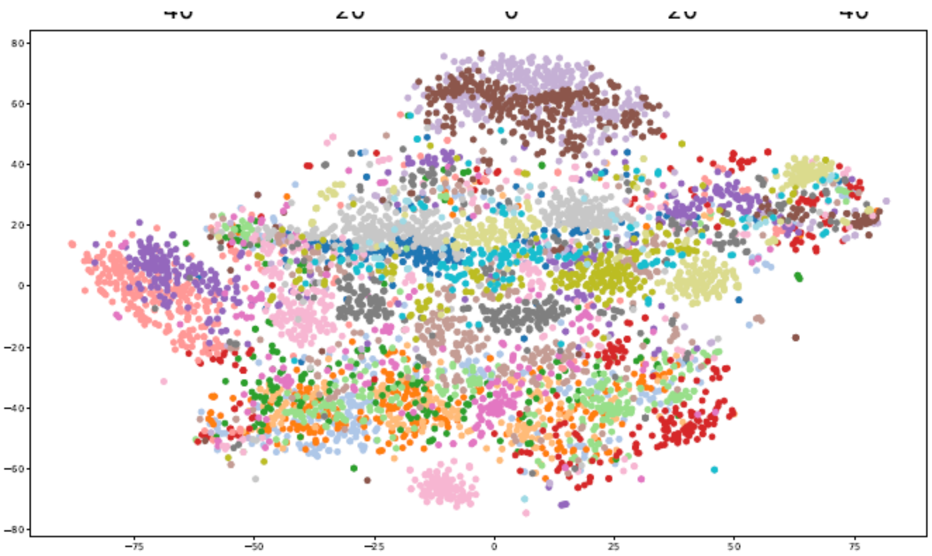
Classification accuracy





Sparse Ratio





Quality of Extracted Representations

computer	sport	drug	weapon	space-flight	atheism	medication	politics	electronics
computer	hockey	tobacco	nuclear	nasa	matthew	cancer	turkey	compass
windows	games	drug	guns	flyers	state	insurance	south	wire
ibm	motorcycl	fallacy	crime	space	atheism	technology	bill	electronic
drive	team	aids	booming	air	book	life	adress	open
disk	play	hiv	controller	statelite	god	hiv	congress	export
system	groups	dades	firearms	send	jesus	des	rockefeller	machines
dos	came	illeg	military	launch	truth	patients	cosmo	byte
key	rom	same	wiring	apartment	faq	water	american	center
hardware	ball	adict	neutral	la	church	health	slave	si

Quality of Extracted Topics

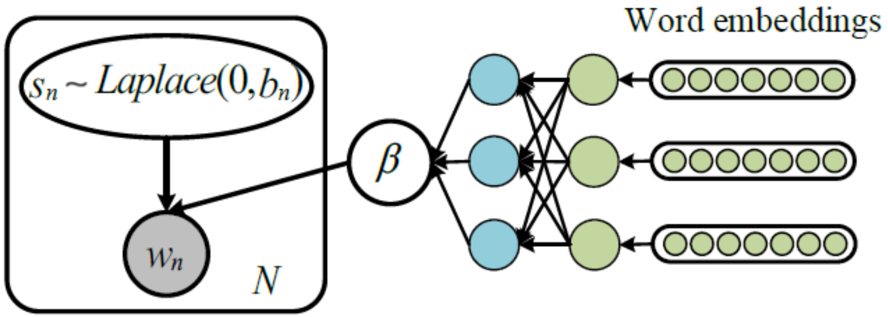
NVSTM

STRUCTURE

Neural Variational Sparse Topic Model

1. For each word n in document d:
   (a) Sample a latent variable word code $s_n \sim Laplace(0, b_n)$.
   (b) Sample the observed word count w_n from $p(w_n|{s_n}^T\beta_n) \sim Poisson({s_n}^T\beta_n)$

$$L(\gamma|\beta) = D_{KL}[q(s|\gamma)||p(s|w,\beta,b)] - logp(w|s,\beta)$$



Neural Variational Sparse Topic Model

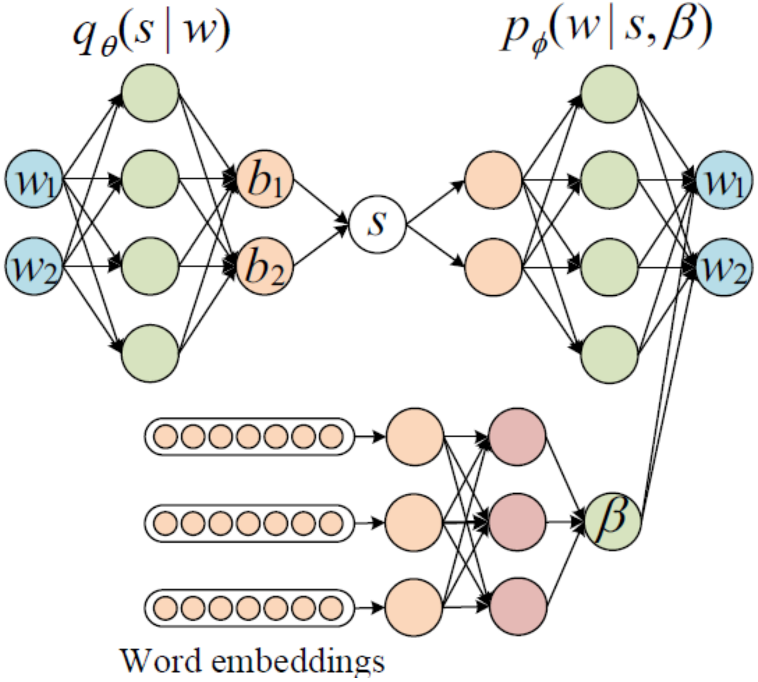
- Rewrite ELBO:

$$L(\theta, \phi|\beta) = -D_{KL}[q_\theta(s|w)||p(s)] + E_{q_\theta(s|w)}(logp_\phi(w|s, \beta))$$

- Reparameterization Trick:

$$s_n \sim Laplace(0, b_n) \to s_n = -b_n sign(\varepsilon) ln(1 - 2|\varepsilon|), \varepsilon \sim U(0, 1)$$

$$L(\Theta) = \sum_{i=1}^{d} \sum_{j=1}^{N} (1 + log2b_{ij}) + \sum_{i=1}^{d} \frac{1}{N} \sum_{j=1}^{N} logp(w_{ij}|s_{ij}, \beta_j)$$



Algorithm 1 Training Algorithm for NVSTM

Input: initialize θ, ϕ, W

repeat
    $w^M \leftarrow$ Random mini-batch of M word counts from full datasets
    $\varepsilon \leftarrow$ Random samples from noise distribution $p(\varepsilon)$
    $g \leftarrow \nabla_{\theta,\phi,W} L(\theta, \phi; w^M, \varepsilon)$
    $\theta, \phi, W \leftarrow$ Update parameters using SGD
until convergence

TRAINING

EXPERIMENTS

Datasets

- 20Newsgroup
- Web snippet
- BBC
- Biomedical

Dataset	Label	Docs	Words	Vocabulary
20Newsgroups	20	18775	135	60698
Web Snippet	8	12265	10.72	5581
BBC	5	2225	11.97	2453
Biomedical	20	19989	7.95	6887

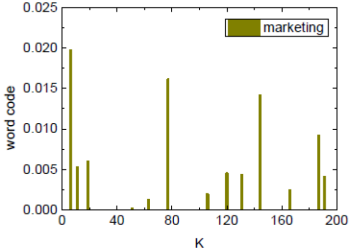
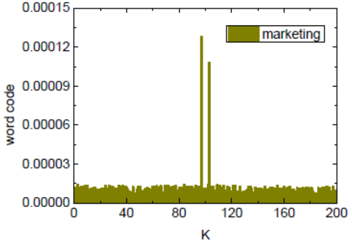
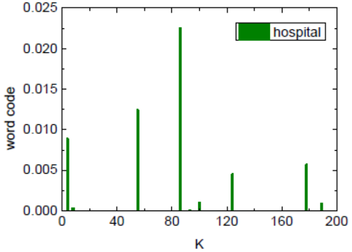
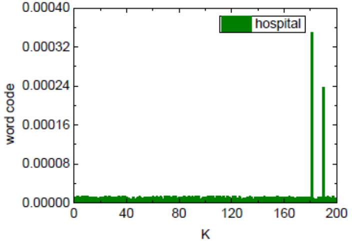
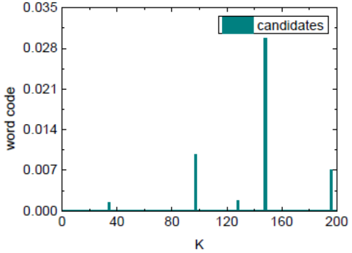
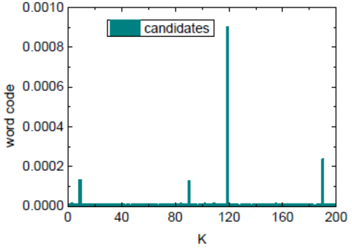
- LDA: a classical probabilistic topic model
- STC: a sparsity-enhanced topic model
- NTM: a neural network based topic model
- DocNADE: An unsupervised neural network topic model
- GLDA: LDA + word embedings

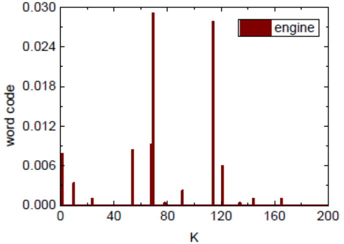
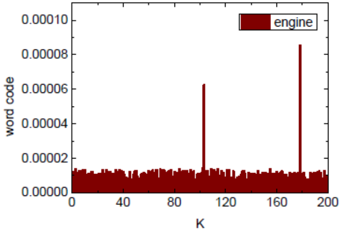
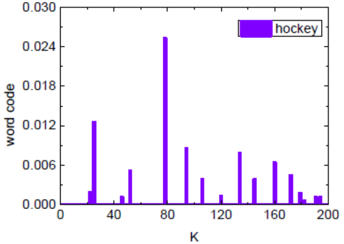
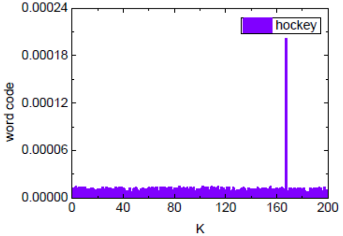
BASELINES

Dataset	Snippet					20NG				
k	**50**	**75**	**100**	**125**	**150**	**50**	**100**	**150**	**200**	**250**
LDA	0.682	0.615	0.592	0.583	0.573	0.545	0.615	0.607	0.613	0.623
STC	0.678	0.686	0.699	0.724	0.701	0.602	0.631	0.647	0.652	0.654
NTM	0.660	0.667	0.723	0.732	0.747	0.623	0.627	0.641	0.632	0.667
DocNADE	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GLDA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NVSTM	0.742	0.808	0.799	0.805	0.818	0.654	0.671	0.672	0.683	0.691

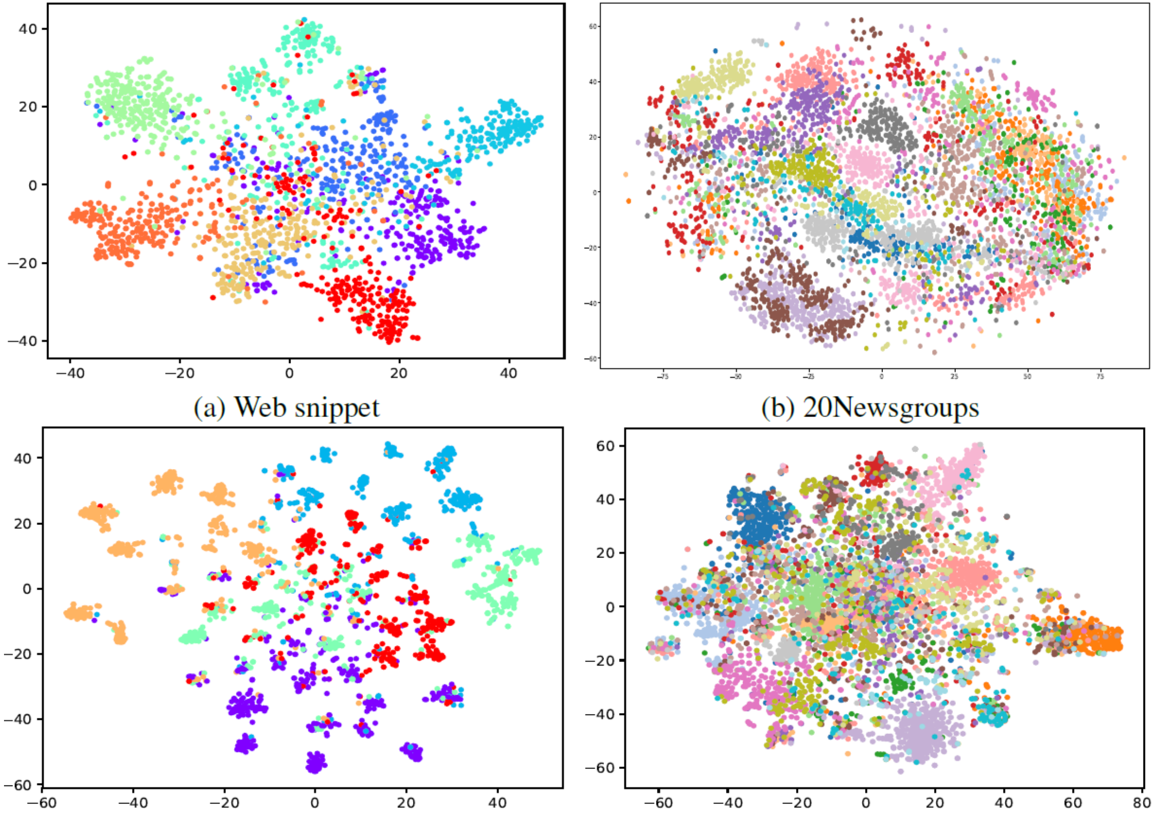
Dataset	BBC					Biomedical				
k	**20**	**30**	**40**	**50**	**60**	**50**	**100**	**150**	**200**	**250**
LDA	0.714	0.724	0.736	0.732	0.746	0.536	0.534	0.547	0.534	0.541
STC	0.552	0.593	0.583	0.634	0.604	0.351	0.405	0.439	0.464	0.494
NTM	0.660	0.667	0.723	0.732	0.747	0.623	0.627	0.641	0.632	0.667
DocNADE	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GLDA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NVSTM	0.762	0.775	0.783	0.796	0.813	0.567	0.623	0.645	0.671	0.664

Classification accuracy

Category	NVSTM	LDA
Business		
Health		
Politics -society		

Engineering		
Sports		

Sparse Ratio



(a) Web snippet

(b) 20Newsgroups

Quality of Extracted Representations

Category	Topic
Business	T6: marketing parascope development business sustainable partnerships movieactors developing partnership T63: finance loans equity loan mortgage financing banking investment mortgages T67: investing ratneshwar investments investment investors invest equity niddk income T133: products source product quality premium csail content manufacture socialsciences T144: development sciserv ecommerce developing innovation developers business marketing projects T176: trade trading markets commodities commodity stocks market parascope currencies
Computers	T38: processor microprocessor processors llnl signonsandiego cpu microprocessors intel cores T108: memory laptop computer computers processor nutritionsource laptops intel disk T112: firefox mozilla netscape macintosh linux windows adobe verizon zdnet T118: systems system control security controls remote automatic monitoring automation T121: msn yahoo firefox aol gmail java algorithm algorithms signonsandiego T159: quantum computing space nasa cpu computational computers astrophysics physics
culture -arts- entertainment	T3: ocos parascope space socialsciences living world academyawards planet intradoc T5: film films indie filmmaker filmmakers movie comedy screening filmmaking T10: sound audio voice acoustic recordings recording listening bass song T16: photography poetry poems prose poet writing getthejob poem photographer T58: sculptor painter artist sculpture sculptures paintings artists artwork surrealist T177: art sculpture socialsciences sculptures painter paintings sculptor painting pcguide
education - science	T41: mathematics physics maths professors students undergraduate science teachers ncidod T59: undergraduate degree student undergraduates faculty students acts particles mathematics T82: teaching school mathforum english teacher mathematics education college schools T102: lecture book lectures papers essay journal seminar conference books T109: topics mathforum essays lectures articles journals emedicinehealth literature syllabus T147: science scientific research journals published theories sciences publications articles

Quality of Extracted Topics

Thanks for your attention