Supplementary Material for "Distributed Online Learning for Latent Dirichlet Allocation"

JinYeong Bak, Dongwoo Kim and Alice Oh
Department of Computer Science
Korea Advanced Institute of Science and Technology
Daejeon, South Korea
{jy.bak, dw.kim}@kaist.ac.kr, alice.oh@kaist.edu

This supplementary material is for "Distributed Online Learning for Latent Dirichlet Allocation" which is submitted to NIPS 2012 workshop on parallel and large-scale machine learning.

1 Distributed Online Learning for LDA Algorithm

Here are the algorithms for the three parts of Distributed Online Learning for LDA.

Algorithm 1 Distributed Online Learning for LDA - Driver

repeat
  Set α, η, $\boldsymbol{\lambda}$, ρ to distributed cache
  Set mini-batch size S, number of topic *K* and number of words *W* to distributed cache
  Create new MapReduce job for new documents
  Run MapReduce job
  Get new document's γ_d and $\boldsymbol{\lambda}_{new}$ from MapReduce job output
  Update α, η and change $\boldsymbol{\lambda}_{new}$ to $\boldsymbol{\lambda}$
until New documents are not observed

Algorithm 2 Distributed Online Learning for LDA - Mapper

Input
KEY - document ID $d \in \{1, \ldots, S\}$
VALUE - document content
Configure
Load α , $\boldsymbol{\lambda}$, number of topic *K* and words *W* from distributed cache
Calculate $\forall k, w \ \exp E_q[\log \beta_{kw}]$
Map
Initialize $\gamma_{dk} = 1$
Read document *d* content
repeat
  Set $\phi_{dwk} \propto \exp\{E_q[\log \theta_{dk}] + E_q[\log \beta_{kw}]\}$
  Set $\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}$
until convergence with γ
Output
variational parameter for document *d*'s θ_d : γ_d
sufficient statistics for $\boldsymbol{\lambda}$: $n_{dw}\phi_{dwk}$

Algorithm 3 Distributed Online Learning for LDA - Reducer

Input

1: KEY - sufficient statistics
2: VALUE - $\forall d, w, k\ n_{dw}\phi_{dwk}$

Configure

1: Load η , $\boldsymbol{\lambda}$, ρ, mini-batch size *S*, number of topic *K* and document *D* from distributed cache

Reduce

1: Set $\tilde{\lambda}_{kw} = \eta + \frac{D}{S}\sum_s n_{sw}\phi_{swk}$
2: Set $\boldsymbol{\lambda}_{new} = (1-\rho)\boldsymbol{\lambda} + \rho\tilde{\boldsymbol{\lambda}}$

Output

1: variational parameter for topic : $\boldsymbol{\lambda}_{new}$

2 Twitter Conversation Topics

In this section, we show some topics from the Twitter conversation corpus discovered by DoLDA. These topics are from the settings of $k = 100$.

Table 1: Several topics from Twitter conversation corpus using distributed online LDA

Topic 15	**Topic 21**	**Topic 36**	**Topic 38**	**Topic 45**
tea	dm	breakfast	gym	black
party	text	toast	body	wear
vote	number	butter	weight	white
country	send	cereal	running	dress
tax	email	pancakes	workout	shirt
labour	sent	branch	chest	clothes
soup	phone	pjs	knee	pants
government	ok	jar	diet	light
ed	check	nutella	leg	look
political	skype	cinnamon	exercise	suit

Topic 54	**Topic 60**	**Topic 74**	**Topic 79**	**Topic 95**
game	xx	drink	movie	room
play	xxx	drinking	film	clean
team	aw	apple	singing	mood
football	babe	water	harry	dirty
mate	hun	bottle	read	shower
player	aww	orange	michael	kitchen
match	lovely	juice	book	house
think	thankyou	tall	films	wash
beat	xo	pie	potter	laundry
good	ahh	blonde	cinema	basement

References

[1] M. Hoffman, D. Blei, and F. Bach. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23:856–864, 2010.