Modeling Online Discourse with Coupled Distributed Topics

Nikita Srivatsan, Zachary Wojtowicz, Taylor Berg-Kirkpatrick

nsrivats@cmu.edu, zdw@andrew.cmu.edu, tberg@cs.cmu.edu

Carnegie Mellon University
School of Computer Science

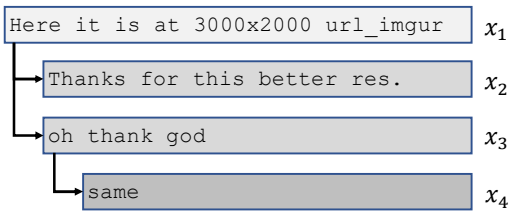
ArXiV

https://arxiv.org/abs/1809.07282

Summary

We propose a deep, globally normalized topic model that captures discursive interactions along observed reply links, typical of online data, in addition to traditional topic information. Our model incorporates latent distributed representations arranged in a deep architecture, which enables efficient GPU-based variational inference. We apply our model to a new dataset of 13M comments mined from Reddit.

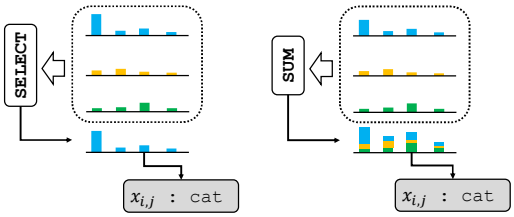
Online Discourse



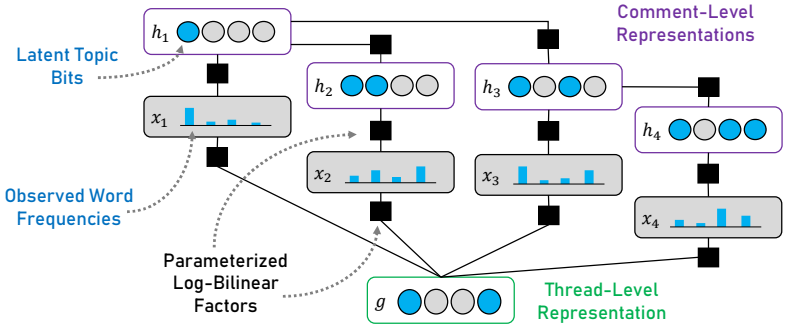
Visualization of a branching Reddit thread with observed reply links

Distributed Topics

Discrete (LDA) vs distributed (RS) topic modeling approaches for explaining some particular word



Model



Factor graph visualization of Discursive Distributed Topic Model (DDTM)

Learning

Joint distribution (intractable due to global normalization)

$$p(x, h; \theta)$$

Form of marginal likelihood and optimization

$$p(x; \theta) = \sum_h \frac{\exp(E(x,h;\theta))}{Z(\theta)}$$

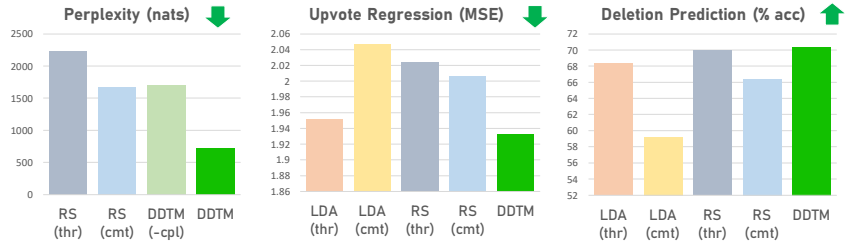
$$\max_\theta \left(\max_{q^{(E)}} \left(\mathbb{E}_{q^{(E)}}[E(x,h)] - \mathbb{E}_{q^{(E)}}[\log q^{(E)}(h)] \right) - \max_{q^{(Z)}} \left(\mathbb{E}_{q^{(Z)}}[E(x',h')] - \mathbb{E}_{q^{(Z)}}[\log q^{(Z)}(x,h)] \right) \right)$$

Factorization structures of variational approximations

$$q^{(Z)}(x, h; \phi, \gamma) \quad q^{(E)}(h; \psi)$$

Likelihood approximation breaks down into two variational lower bounds

Results

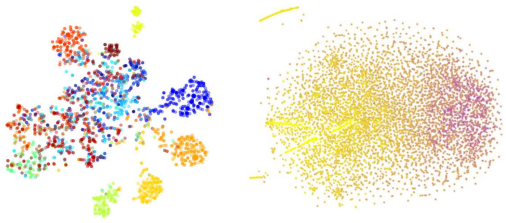


Performance of our model (DDTM), with and without (-cpl) coupling potentials, vs baselines at both thread (thr) and comment (cmt) resolutions

Analysis

t-SNE plot of thread-level representations colored by subreddit of origin

t-SNE plot of comment-level representations colored by comment length



Strongest word emissions for particular thread-level topic bits (single bit active)

Bit 1	Bit 2	Bit 3
maduro	comey	btc
venezuelan	pede	gameplay
ballot	macron	tutori
puerto	pg13	cyclist
catalonia	maga	dev
rican	globalist	currenc
quak	ucf	kitti
skateboard	committe	bitcoin
venezuela	cuck	rpg
quebec	distributor	crypto

Strongest word emissions for particular comment-level topic bits (single bit active)

Bit 1	Bit 2	Bit 3
irl	faq	funniest
riamverysmart	tldrs	mah
legend	pms	tfw
omfg	165	teleport
riski	til	fav
aboard	keyword	hoo
favr	questions	plz
madman	feedback	bah
skillset	chat	whyd
tunnel	pm	dumbest

Strongest word emissions based on inferred latent comment-level topic representations (multiple bits active)

Embedding 1	Embedding 2	Embedding 3
reev	reddit	moron
keanu	shill	douchebag
christoph	question	stupid
murphy	background	dipshit
walken	user	snitch
vincent	subreddit	jackass
chris	answer	dickhead
til	relev	idioci
wick	discord	hypocrit
roger	guild	riddanc

Qualitative analysis of strongest word emissions for selected topic bits