PSLDA: A novel supervised pseudo document-based topic model for short texts

Mingtao SUN, Xiaowei ZHAO, Jingjing LIN, Jian JING, Deqing WANG, Guozhu JIA

Frontiers of Computer Science, DOI: 10.1007/s11704-021-0606-3

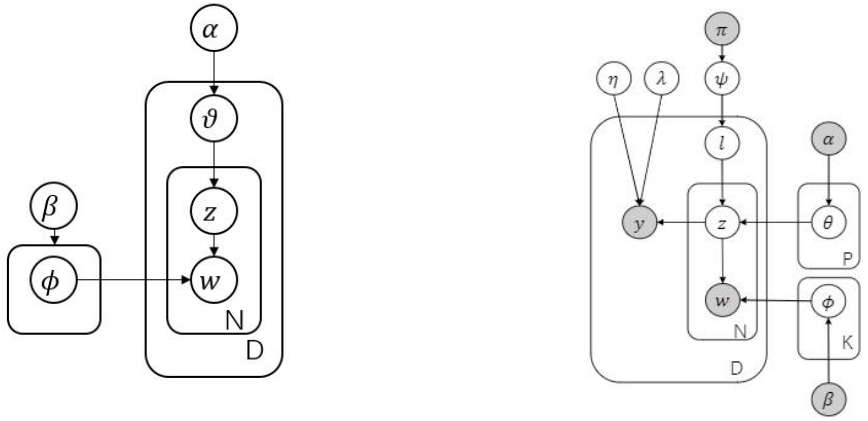
Problems & Ideas

- Problems of conventional short text topic model:

  - Sparseness of word-occurrence and the diversity of topics.

  - Previous methods suffer from high time complexity because of the pre-trained processing on the external large-scale dataset.

- Ideas: Assume that short texts are generated from the normal size latent pseudo documents, and the topic distributions are sampled from the pseudo documents.



The graphical representation of traditional topic model and PSLDA.

Main Contributions

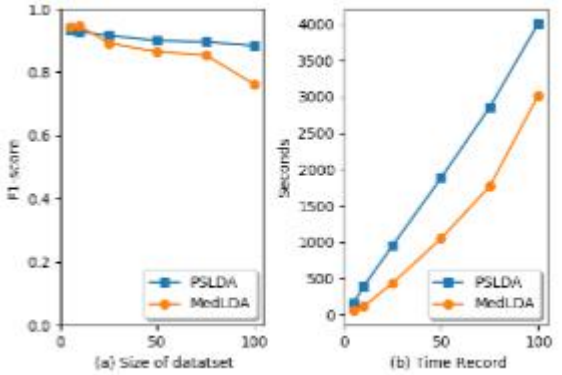
- Contributions:

  -A Pseudo-document-based Topic Model (PTM) for observed short texts

  -An expected classifier to the model, which is an alternative formulation of max-margin supervised topic model

Method	DBLP			NEWS			TWEET		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
PSLDA	0.6356$_{+2.2e-6}$	0.6376$_{+2.2e-6}$	0.6344$_{+6.4e-7}$	**0.8884**$^{**}_{+2.1e-5}$	**0.8824**$^{**}_{+1.8e-5}$	**0.8836**$^{**}_{+2.3e-5}$	**0.7252**$^{**}_{\pm 7.6e-7}$	**0.7092**$^{**}_{\pm 7.4e-6}$	**0.7136**$^{**}_{\pm 2.2e-6}$
MedLDA	0.6664$_{+1.1e-2}$	0.658$_{+9.9e-3}$	0.6572$_{+1.1e-2}$	0.7344$_{+5.3e-4}$	0.8008$_{+1.6e-4}$	0.7616$_{+3.0e-4}$	0.544$_{+6.9e-3}$	0.5548$_{+5.7e-3}$	0.5456$_{+6.4e-3}$
SLDA	0.652$_{+6.9e-5}$	0.6567$_{+4.9e-5}$	0.6567$_{+7.3e-5}$	0.853$_{+8.3e-5}$	0.854$_{+8.6e-5}$	0.854$_{+8.6e-5}$	0.63$_{+1.2e-4}$	0.643$_{+8.9e-5}$	0.6313$_{+1.3e-4}$
LDA+SVM	0.604$_{+1.2e-4}$	0.59$_{+2.2e-4}$	0.594$_{+3.5e-4}$	0.7384$_{+7.9e-5}$	0.7348$_{+3.9e-4}$	0.7356$_{+5.7e-5}$	0.5536$_{+6.8e-3}$	0.5168$_{+7.5e-4}$	0.5292$_{+7.3e-4}$
SPTM	0.661$_{+2.4e-4}$	0.667$_{+7.6e-4}$	0.663$_{+7.2e-3}$	0.760$_{+1.8e-3}$	0.761$_{+7.9e-5}$	0.759$_{+3.1e-3}$	0.551$_{+6.2e-4}$	0.558$_{+4.1e-3}$	0.550$_{+1.1e-2}$
SATM	0.657$_{+6.3e-5}$	0.662$_{+5.1e-6}$	0.654$_{+4.3e-3}$	0.697$_{+7.6e-4}$	0.702$_{+8.1e-3}$	0.686$_{+8.4e-3}$	0.599$_{+6.0e-4}$	0.605$_{+7.1e-5}$	0.594$_{+2.5e-4}$
PTM	**0.667**$^{**}_{+4.1e-3}$	**0.672**$^{**}_{+2.9e-2}$	**0.668**$^{**}_{+6.1e-4}$	0.755$_{+5.7e-3}$	0.757$_{+5.4e-4}$	0.754$_{+1.5e-2}$	0.561$_{+8.1e-4}$	0.568$_{+7.6e-3}$	0.559$_{+4.7e-3}$

The Precision, Recall and F-measure on three datasets.



Effect of dataset size on (a) F1-score and (b) training time