Latent Semantic Topics Distribution Over WebContent Corpus

Zhengyuan Ding, Rui Jiang, Ruijie Chen, Danfeng Li,
Advisor: Simon, Bombora

NYU Center for Data Science

bombora

Problem Statement

With the explosion of electronic document archives, there is an increasing demand of automated techniques for document analy -sis. Topic modeling provides an convenient way to learn the und -erlying structure of a given corpus.

Our objective is to discover the topic distributions of Bombora's web content corpus with various generative models.

Dataset and Preprocessing

Dataset

1) 10,000 unlabeled web content documents

2) 20,000 human-labeled documents (20NewsGroups)

Data preprocessing

tokenization, stopwords removal, lemmatization, stemming, bigram collocation, non-english words removal

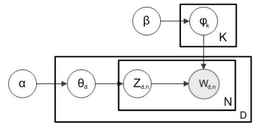
Environment

GCP virtual machine instances with 2 x NVIDIA Tesla P4 GPU.

Statistical Models

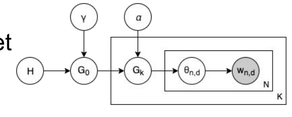
Latent Dirichlet Allocation

Treats each topic as a distribution of words. For each document, draw a mixture of topics from a Dirichlet distribution, each word in the document is an independent draw from that mixture.



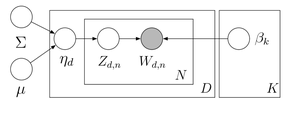
Hierarchical Dirichlet Process

An extension of LDA: also uses a Dirichlet process to capture the uncertainty in the number of topics, but doesn't require specifying number of topics in advance.



Correlated Topic Modeling

CTM break the near independent assumption between topics in LDA and models topic correlations using the covariance matrix from logistic normal distribution.

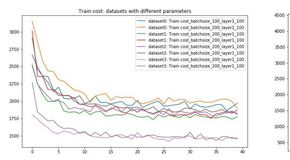


Variational Autoencoder (VAE)

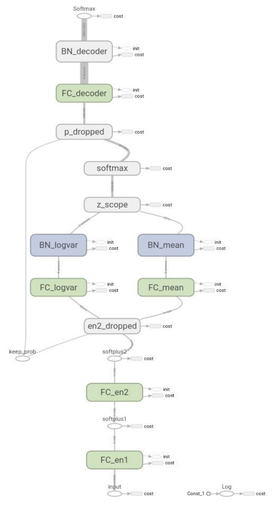
Main idea: we constructed a neural variational framework for gener -ative topic model, inspired by the variational autoencoder. The key is to build inference neural network to approximate the intractable distri -butions over the latent variables(representing topic distribution).
Object Function: composed of KL divergence and reconstruction error:

$$L(\gamma, \phi|\alpha, \beta) = -D_{KL}[q(\theta, z|\gamma, \phi||p(\theta, z|\alpha] + E_{q(\theta,z|\gamma,\phi)}[logp(w|z, \theta, \alpha, \beta]$$

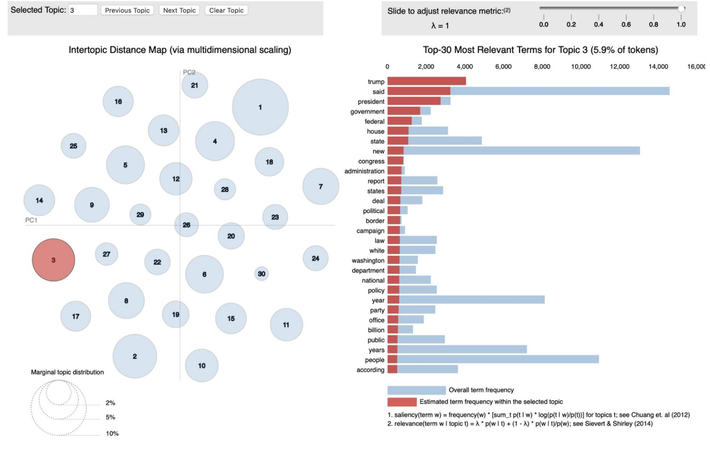
Training challenge: prone to be "trapped" to a local optimum closer to prior belief in early training epochs; standard data preprocessing methods even hurt model performance.
Model structure and performance:



Larger batch size and higher learning rate help model converge, and with tuning batch size, hidden size and vocabulary size, etc. we got much better perplexity and more interpretable topic-word distribution.



Results



	LDA	CTM	HDP	VAE
Perplexity(*PPL*)	2927.87	2317.41	1136.20	433.31
Coherence Value	0.36	0.35	0.38	0.36
Computational Efficiency(s)	157	246	178	193

Table 2: Performance of Models

Based on the experiment, we have drawn the following conclusions:

- We generated a topic graph with multiple lasso regressions on the latent variable values trained by VAE.
- Among the models we experimented, coherence scores are similar but VAE has the lowest perplexity.
- Data preprocessing methods like bigram and lemmatization work well for simpler models, but may hurt deep learning model.
- Variational inference improves computational efficiency of computing posterior distribution to some extent.

Evaluation

Human-labeled Dataset: 20NewsGroups

Topic words:

	Religion	Science	Health	Computer
LDA	god jesus church believe word bible point love sin mean	space system nasa mission orbit earth also data moon book	patient study cause food doctor pain since disease day case	card system do window run disk driver mac pc drive
VAE	jesus matthew prophecy god holy spirit isaiah db hanging israel messiah	satellite space nasa satellites telescope launch data infrared observatory spacecraft	tobacco public health aerospace commercial space illness venture space technology mariner health russia	chip serial number fpu motherboard screen card phones session key monitor clipper

Evaluation metrics:

	LDA	HDP	VAE
Coherence Score	0.488	0.506	0.518
Time Cost	68s	100s	390s