yang_2017_adapting_topic_models_using_lexical_associations_with_tree_priors

Year

2017

Author(s)

Yang, Weiwei and Boyd-Graber, Jordan and Resnik, Philip

Title

Adapting Topic Models using Lexical Associations with Tree Priors

Venue

EMNLP

Topic labeling

Partially automated

Focus

Secondary

Type of contribution

Established approach

Underlying technique

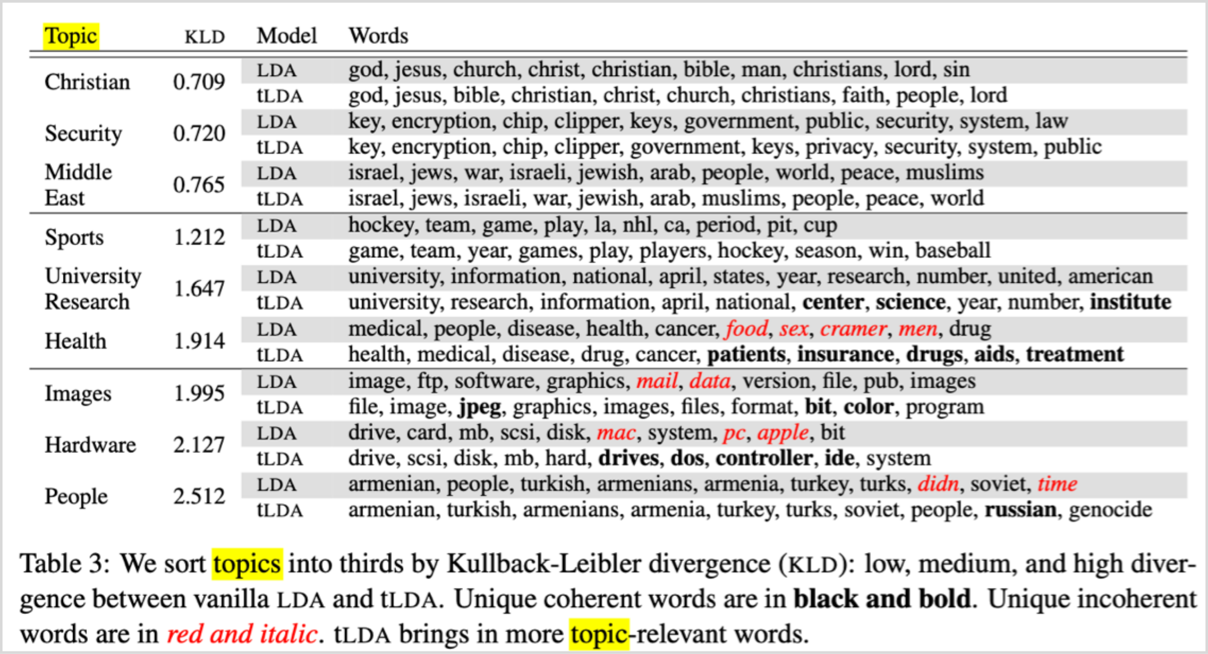
Manual labeling assisted by gold standard labels from dataset

Topic labeling parameters

\

Label generation

Manual assignment go gold standard label from dataset to generated topics.



Topic	KLD	Model	Words
Christian	0.709	LDA	god, jesus, church, christ, christian, bible, man, christians, lord, sin
		tLDA	god, jesus, bible, christian, christ, church, christians, faith, people, lord
Security	0.720	LDA	key, encryption, chip, clipper, keys, government, public, security, system, law
		tLDA	key, encryption, chip, clipper, government, keys, privacy, security, system, public
Middle East	0.765	LDA	israel, jews, war, israeli, jewish, arab, people, world, peace, muslims
		tLDA	israel, jews, israeli, war, jewish, arab, muslims, people, peace, world
Sports	1.212	LDA	hockey, team, game, play, la, nhl, ca, period, pit, cup
		tLDA	game, team, year, games, play, players, hockey, season, win, baseball
University Research	1.647	LDA	university, information, national, april, states, year, research, number, united, american
		tLDA	university, research, information, april, national, **center**, **science**, year, number, **institute**
Health	1.914	LDA	medical, people, disease, health, cancer, *food*, *sex*, *cramer*, *men*, drug
		tLDA	health, medical, disease, drug, cancer, **patients**, **insurance**, **drugs**, **aids**, **treatment**
Images	1.995	LDA	image, ftp, software, graphics, *mail*, *data*, version, file, pub, images
		tLDA	file, image, **jpeg**, graphics, images, files, format, **bit**, **color**, program
Hardware	2.127	LDA	drive, card, mb, scsi, disk, *mac*, system, *pc*, *apple*, bit
		tLDA	drive, scsi, disk, mb, hard, **drives**, **dos**, **controller**, **ide**, system
People	2.512	LDA	armenian, people, turkish, armenians, armenia, turkey, turks, *didn*, soviet, *time*
		tLDA	armenian, turkish, armenians, armenia, turkey, turks, soviet, people, **russian**, genocide

Table 3: We sort topics into thirds by Kullback-Leibler divergence (KLD): low, medium, and high divergence between vanilla LDA and tLDA. Unique coherent words are in **black and bold**. Unique incoherent words are in *red and italic*. tLDA brings in more topic-relevant words.

Motivation

\

Topic modeling

tLDA (Boyd-Graber et al., 2007) with tree priors

Baseline: LDA

We construct tree priors […] Then tLDA identifies topics with these tree priors in Amazon reviews and the 20NewsGroups datasets.

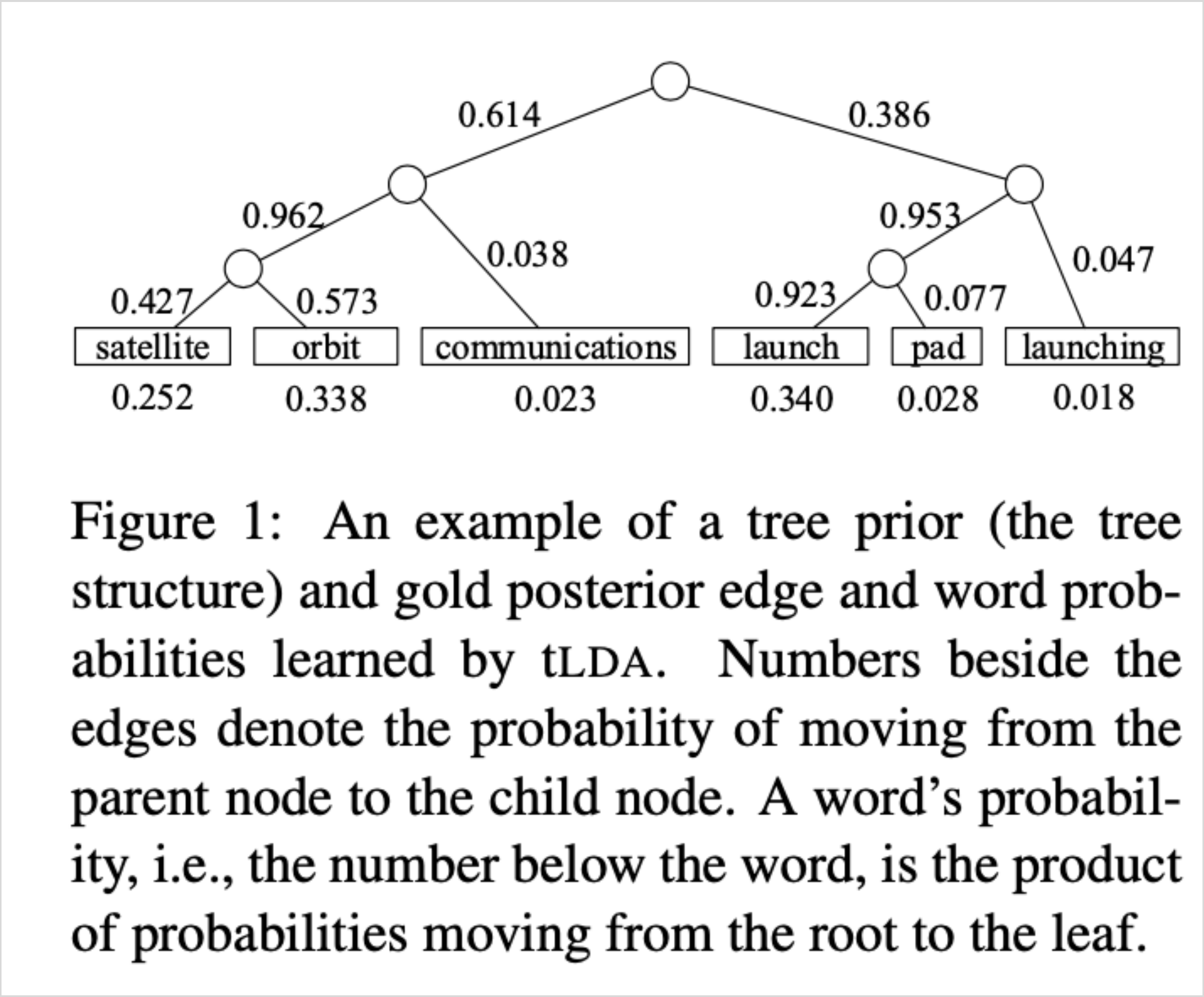


Figure 1: An example of a tree prior (the tree structure) and gold posterior edge and word probabilities learned by tLDA. Numbers beside the edges denote the probability of moving from the parent node to the child node. A word's probability, i.e., the number below the word, is the product of probabilities moving from the root to the leaf.

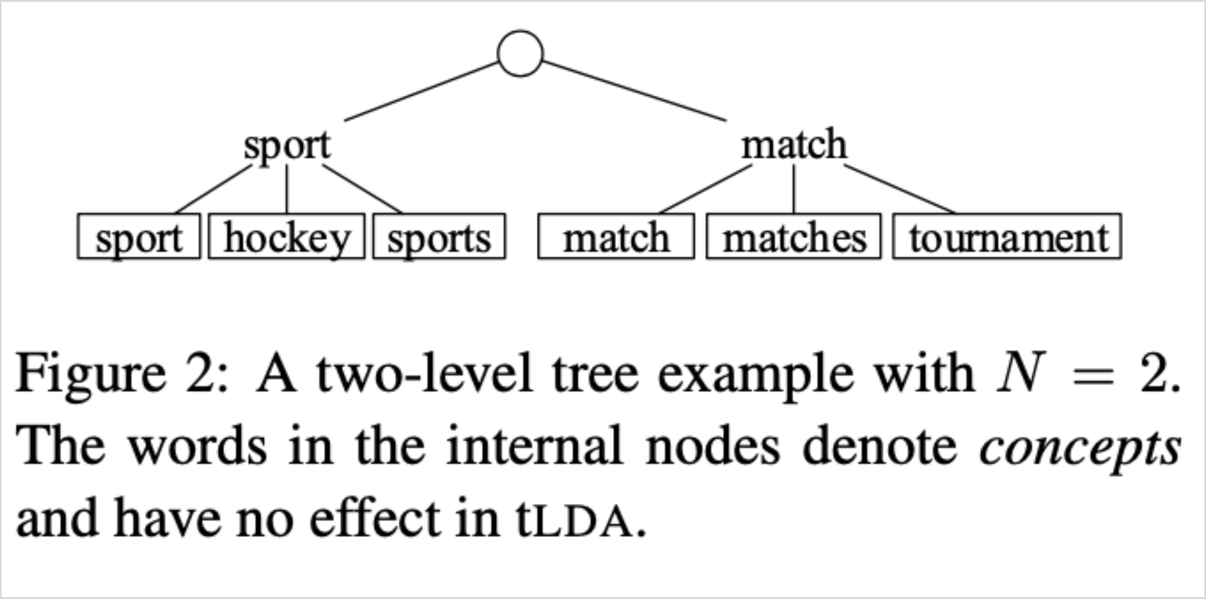


Figure 2: A two-level tree example with $N = 2$. The words in the internal nodes denote *concepts* and have no effect in tLDA.

A topic in tLDA is a multinomial distribution over the paths from the root to leaves.
An internal node, i.e., the circles in Figure 1, is a multinomial distribution over its child nodes.
The probability of a path is the product of probabilities of picking the nodes in the path

Topic modeling parameters

Nr. of topics (K): 20
α: 0.01
β: 0.01

Nr. of topics

20 (one per newsgroup class)

Label

One of 20 labels extracted from the posts newsgroups.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Domain (paper): Topic modeling (with tree priors)
Domain (dataset): News

Problem statement

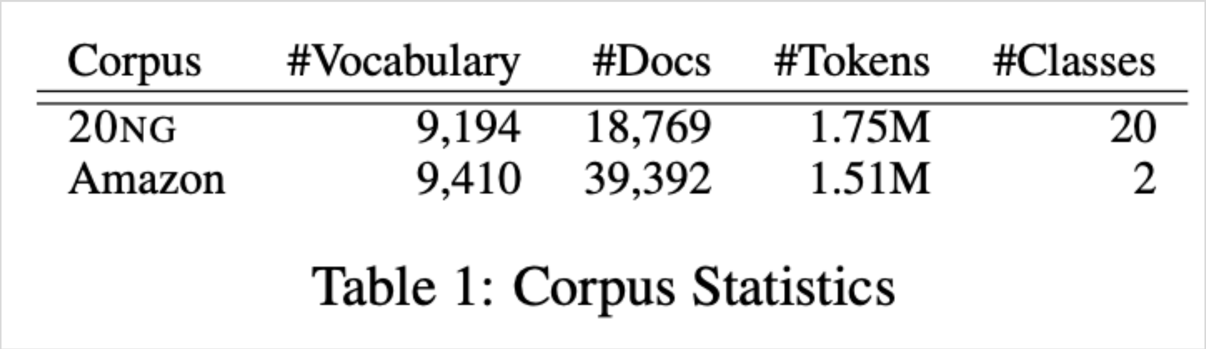
Incorporating an approximation of human topic interpretability into the topic model optimisation process.
Achieving this by integrating lexical association into topic optimisation using tree priors in order to improve topic interpretability.
Taking advantage of both first order word associations and the higher-order associations

captured by word embeddings.

Constructing tree priors with combinations of two types of word association scores (skip-gram probability and G2 likelihood ratio) and three construction algorithms (two-level, hierarchical clustering with and without leaf duplication).
Using treeLDA (Boyd-Graber et al., 2007) to derive topics using a given tree prior.

Corpus

Dataset 2
Origin: Various news outlets
Nr. of documents: 18.769
Details: 20NewsGroup dataset (Lang, 1995)



Corpus	#Vocabulary	#Docs	#Tokens	#Classes
20NG	9,194	18,769	1.75M	20
Amazon	9,410	39,392	1.51M	2

Table 1: Corpus Statistics

Document

Pre-processing

- Tokenization
- Stopword removal
- Words sorted by their document frequencies and the top words are returned
- Words that appear in more than 30% of the documents are removed
- Reviews with 3 stars are discarded from Dataset 2.

```
@inproceedings{yang_2017_adapting_topic_models_using_lexical_associations_with_tree_priors,
    title = "Adapting Topic Models using Lexical Associations with Tree
Priors",
    author = "Yang, Weiwei  and
      Boyd-Graber, Jordan  and
      Resnik, Philip",
    booktitle = "Proceedings of the 2017 Conference on Empirical Methods in
Natural Language Processing",
    month = sep,
    year = "2017",
    address = "Copenhagen, Denmark",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/D17-1203",
    doi = "10.18653/v1/D17-1203",
    pages = "1901--1906",
    abstract = "Models work best when they are optimized taking into account
the evaluation criteria that people care about. For topic models, people often
care about interpretability, which can be approximated using measures of
lexical association. We integrate lexical association into topic optimization
using tree priors, which provide a flexible framework that can take advantage
of both first order word associations and the higher-order associations
captured by word embeddings. Tree priors improve topic interpretability without
hurting extrinsic performance.",
}
```

#Thesis/Papers/Initial