Applying Link Block LDA to Political Blog Posts

Derek Owens-Oas

March 31, 2018

Link block LDA:

I extend the LDA topic model of Blei, et al. by modeling each document's node and links and clustering documents with similar topics.

Parameter estimation:

Model parameters are estimated using Bayesian inference. I used Gibbs sampling to approximate the posterior distribution of each model parameter, and then took posterior means and medians.

Political blog posts:

I apply the method to text from 9430 political blog posts from January, 2012. Each document's node is the blog website and the links are hyperlinks to other blogs.

Learned Topics in the corpus:

In this section I present a selection of the topics learned. Following parameter estimation, which is unsupervised, I specify words of interest and choose the topic which assigns those words the highest probability. Then the highest scoring words of each topic are shown. Scores are from the lda R package.

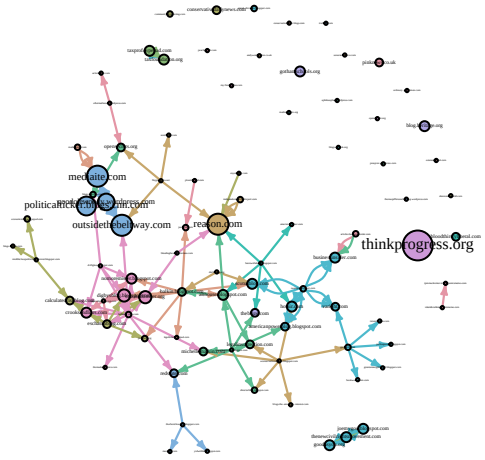
election	climate	economy	energy	health	legal	jobs	education
political	climate	market	oil	health	government	jobs	school
vote	global	fed	energy	care	rights	money	education
election	change	economy	gas	insurance	law	job	students
democratic	warming	gold	pipeline	people	people	workers	schools
democrats	energy	year	keystone	immigration	freedom	government	public
run	solar	price	environmental	abortion	laws	work	college
year	science	money	natural	reform	does	private	children
campaign	scientists	global	project	country	free	americans	student
voting	power	china	xl	applause	constitution	labor	community
day	new	prices	department	mr	legal	employees	poor

The learned topics can be identified as relating to the economy, health, and education. These are recognizable political topics. Also discovered is a topic emphasizing the election and campaign, which is underway in January 2012.

Network Visualization:

Below I present a visualization of the network of blogs relating to the "election". I consider posts assigned to any block which assigns highest probability to the election topic. Edge width represents number of links

between blogs in these posts, and node sizes represent number of posts on that blog. Colors of each node represent the block to which most of their posts at that time are assigned.



Here I find that `crooksandliars` posts and links the most about the election. `tpmmuckraker` is linked a lot about the election. Other blogs which post, link, or are linked about the election are identifiable from the plot.

Computational Cost:

I consider an analysis with K = 50 topics, V = 5000 unique words, S = 100 kept samples and 1000 total samples, B = 100 number of communities, A = 500 number of unique senders and receivers, W = 24363882 total words, N = 110000 total documents.

The space required is approximately 500.5655032 Mb.

The time required is approximately 4.0324074 days.