Sara Kangaslahti, Danny Ebanks, R. Michael Alvarez
California Institute of Technology

ANALYZING SOCIAL MEDIA CONVERSATIONS AT SCALE WITH TENSOR LDA

ABSTRACT

The data exchanged daily on social media platforms such as Twitter presents an important testbed for topics in social science research including messaging coordination around social movements such as #MeToo. The growth of social media data and prohibitive cost of annotation has rendered traditional supervised machine learning methods infeasible for analysis. This cost arises from the enormity and dynamic nature of some of the most compelling textual data.

In order to analyze these datasets, unsupervised topic modeling methods such as Latent Dirichlet Allocation (LDA) have gained widespread popularity, since they can extract the important information without requiring labeling or prior knowledge. Unlike supervised methods, unsupervised learning methods like LDA do not require a pre-defined set of topics, but the existing LDA models face numerous computational limitations. To overcome these, we build on previous scalable spectral methods and propose an online GPU-based tensor Latent Dirichlet Allocation method (tLDA). We achieve optimal scalability by centering and batching the data, as well as providing an end-to-end GPU pipeline, from data pre-processing to model estimation.

In order to demonstrate the utility of the method, we show qualitative results derived from applications of this method for studying topic evolution in domains relevant to social research -- particularly a large #MeToo Twitter dataset, composed of over 8 million tweets. We find that topics related to political events were generally ephemeral, whereas topics related to supporting women in the #MeToo Movement were persistently prominent as the topics dynamically evolved.

To validate the method, we show relative gains in topic coherence of 7 to 44 percent over traditional LDA methods and 2 to 8 percent over previous Tensor LDA methods. At the same time, our method is up 7.6x faster than traditional LDA implementations and 20x faster than previous tensor LDA methods. On GPU, we show 10x improvement against the fastest parallelized CPU implementations of LDA. Finally, we demonstrate scaling in this paper up to 30 million tweets.

WEDNESDAY, APRIL 13 2022
3:30pm - 4:30pm, CHS 33-105 (External Speaker)
https://ucla.zoom.us/j/93739481772?pwd=MVZvMjNuQkUzdk1TM2hWWFZKUGZZQT09
Meeting ID: 937 3948 1772 | Passcode: 339592