A Parallel LDA-Based Approach to Identifying Trends of Cultural Products in the Music Industry

[1]Richard de Groof, [2]Haiping Xu, [3]Jurui Zhang and [4]Raymond Liu
[1,2]*Computer and Information Science Department, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA*
Email: {rdegroof, hxu}@umassd.edu
[3,4]*Department of Marketing, University of Massachusetts Boston, Boston, MA 02125, USA*
Email: {jurui.zhang, raymond.liu}@umb.edu

Abstract

In the music industry, feedback provided through online social media by listening audiences could be incredibly valuable for local artists. With tremendous amount of textual data available online, automatic tools are required for efficient data processing. Latent Dirichlet Allocation (LDA), as a useful text mining mechanism, supports identification of latent topics in massive textual data. However, traditional usages of this methodology require manual interpretation of the posterior distributions of topics across documents and words. In this sponsored project funded by the UMass President's Office, we introduce a novel post-processing step to the LDA process, which allows us to derive significant keywords that are highly reflective of the underlying topics. As a result, our process is more effective in identifying trends in the music industry by deriving meaningful keywords in unorganized collections of textual data. To make our approach more efficient, we designed a parallelized LDA process, called pLDA. Our approach not only speeds up the inference procedure used in LDA, but also better approximates the original sequential process for text mining.

Objectives

1. To provide recommendations to the WUMB music station through an analysis of popular musicians featured at the station, and also lesser-known local Boston artists found at thedelimagazine.com.
2. To provide an efficient methodology for organizing a large amount of textual data with little manual labeling.

Preliminaries

Fig. 1 shows the LDA framework, which is a three-level hierarchical Bayesian model that defines distributions of documents and words in terms of mixtures of latent topics. LDA presents Dirichlet and multinomial prior distributions for the assignment of words (w) to topics (z) and topics to documents, φ and θ, respectively, with Dirichlet parameters β and α. The complete joint probability of a corpus and the topic assignments can be calculated as in Eq. (1).

$$P(W,Z,\theta,\phi;\alpha,\beta)= \prod_{i=1}^{K} P(\phi_i,\beta)\prod_{j=1}^{M}[P(\theta_j;\alpha)\prod_{t=1}^{N}[P(z_{j,t}\,|\,\theta_j)P(w_{j,t}\,|\,\phi_{z_{j,t}})]] \quad (1)$$

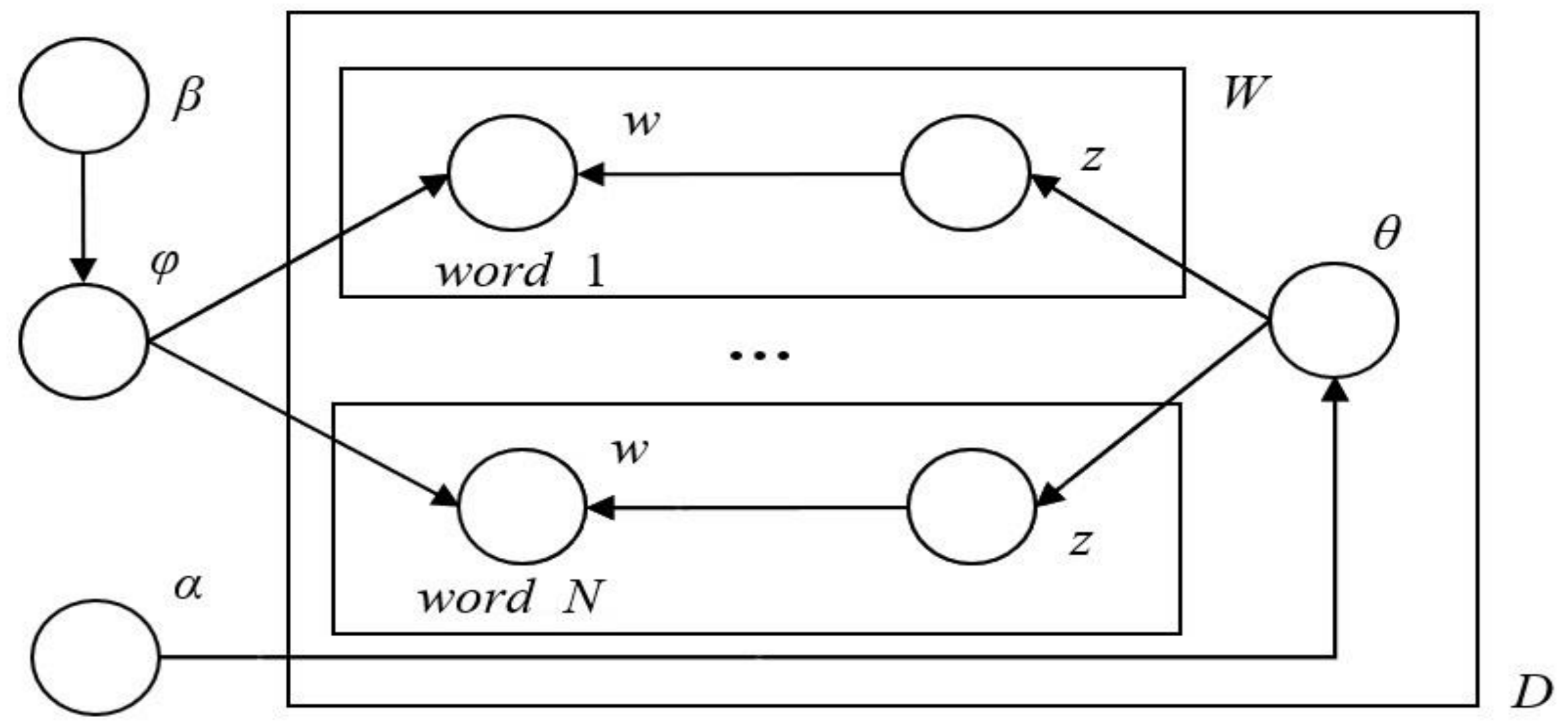

Fig 1. LDA probabilistic model

By integrating out θ and φ and using properties of the Dirichlet and multinomial distributions, Eq. (1) can be reduced to Eq. (2).

$$P(Z_{(m,n)}=k,Z_{-(m,n)},W;\alpha,\beta)\propto (n_{m,(.)}^{k,-(m,n)}+\alpha_k)\frac{(n_{(.),n}^{k,-(m,n)}+\beta_n)}{\sum_{r=1}^{N}(n_{(.),r}^{k,-(m,n)}+\beta_r)} \quad (2)$$

We adopt Collapsed Gibbs Sampling (CGS) to calculate the probability of topics. In Eq. (2), *m* and *n* represent document and word, respectively. The superscript *k* represents the assignment of the *k*-th topic; while the subscripts represent this assignment to document *m* and word *n*. The topic assignment at the current measurement is excluded from the computation as indicated by –(*m, n*). Like a random walk, CGS proceeds from the current position considering only all other stationary assignments, and makes a random draw from the probability distribution over topics generated by Eq. (2). The next assignment is a likely choice considering this arrangement.

Significant Keywords and Parallelized LDA

$$\theta_{m,k}=(n_{m,(.)}^{k,-(m,n)}+\alpha_k) \quad (3)$$

$$\phi_{k,n}=\frac{(n_{(.),n}^{k,-(m,n)}+\beta_n)}{\sum_{r=1}^{N}(n_{(.),r}^{k,-(m,n)}+\beta_r)} \quad (4)$$

The purpose of CGS is to derive posterior distributions θ and φ. Once CGS has reached a stationary point where the topic assignments are stable, we calculate the distributions as in Eq. (3) and (4), where θ is normalized by all topics occurring for that document. Applying LDA in practice, one or the other matrices has been the basis for text classification or the identification of relevant words. We derive *Significant Keywords* using matrix multiplication of θ with φ as in Eq. (5) and (6) .

$$\sum_{k=1}^{K}\frac{P(d\,|\,z_k)P(z_k)}{P(d)}P(w\,|\,z_k)=\sum_{k=1}^{K}\frac{P(d,w\,|\,z_k)P(z_k)}{P(d)} \quad (5)$$

$$\sum_{k=1}^{K}\frac{P(d,w,z_k)P(z_k)}{P(z_k)P(d)}=\sum_{k=1}^{K}\frac{P(d,w,z_k)}{P(d)}=\frac{P(d,w)}{P(d)}=P(w\,|\,d) \quad (6)$$

To make the computation more efficient, we introduce a parallelized LDA, called pLDA, where processing occurs across partitions of the documents. In addition, we synchronize the counts in real-time, and the results better approximate the original sequential process.

Once we have extracted the significant keywords, we identify clusters of documents according to common occurrences or non-occurrences of words with a high *P(w|d)* value, which is reflective of the underlying topic. After the topics have been identified, we associate each document with a sentiment score. In our case study, we use our approach to classify many thousands of comments from the Facebook pages of well-known and local artists according to hot topics in the music industry. Assuming that artist popularity is related to a combination of the public's perception of these different factors, we can find the primary topics.

Case Study

We analyze music review comments posted on social media (e.g., Facebook pages) using pLDA, and calculate the sentiment scores derived using the Stanford Core NLP. The primary topics we found in our analysis were *streaming*, describing live streaming services such as Spotify; *album*, where artists described the release of a new album; and *shows*, where artists talked about upcoming live performances. The sentiment scores are combined with the artist popularity as measured by the number of Facebook followers into a linear system of equations as in Eq. (8).

$$\begin{gathered} ArtistPopularity_i=\alpha_k streaming_i+\alpha_l album_i+\alpha_m shows_i \\ \cdots \\ ArtistPopularity_j=\alpha_k streaming_j+\alpha_l album_j+\alpha_m shows_j \end{gathered} \quad (8)$$

Solving the equations using weighted least squares regression with the weights being the number of comments, we can derive the topic significance value for each coefficient, representing the importance to artist popularity. Fig. 2 and Fig. 3 show the corresponding results.

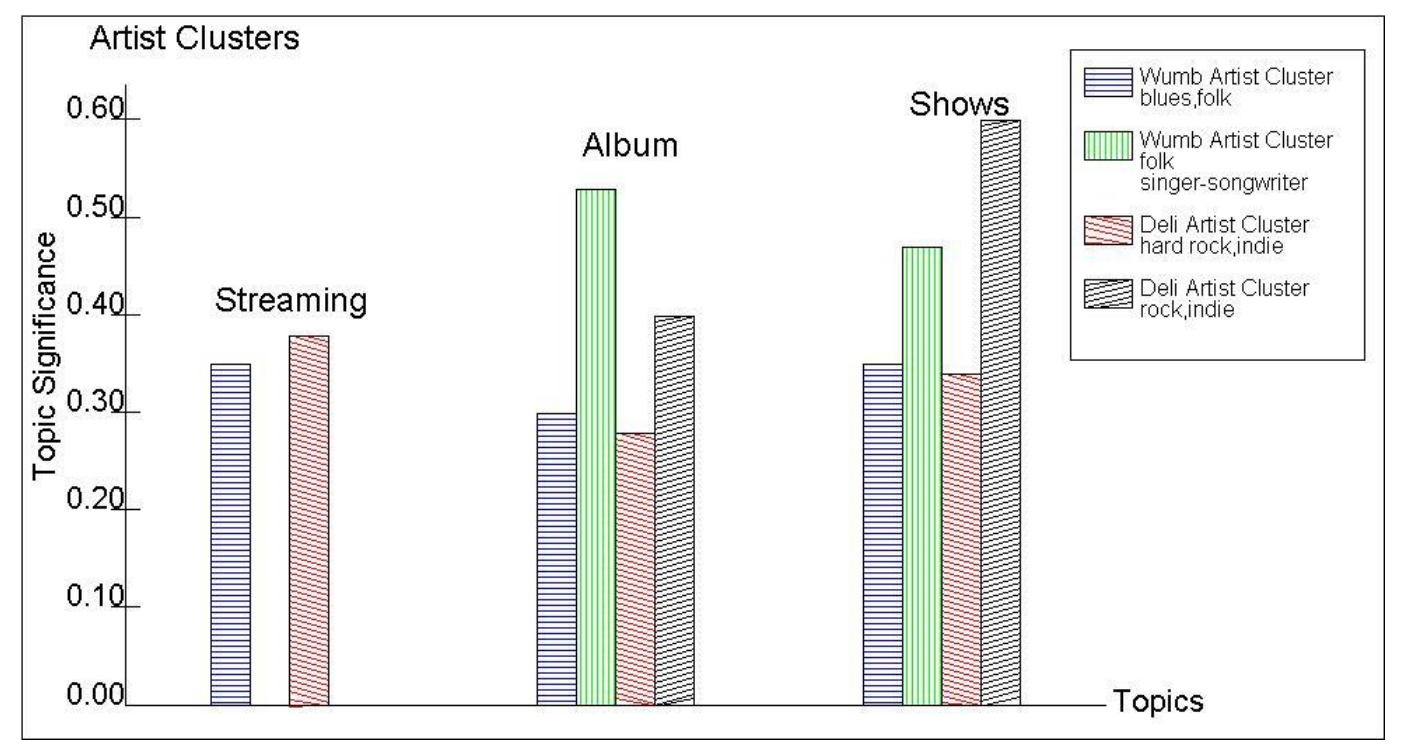

Fig 2. Deli and WUMB Clusters

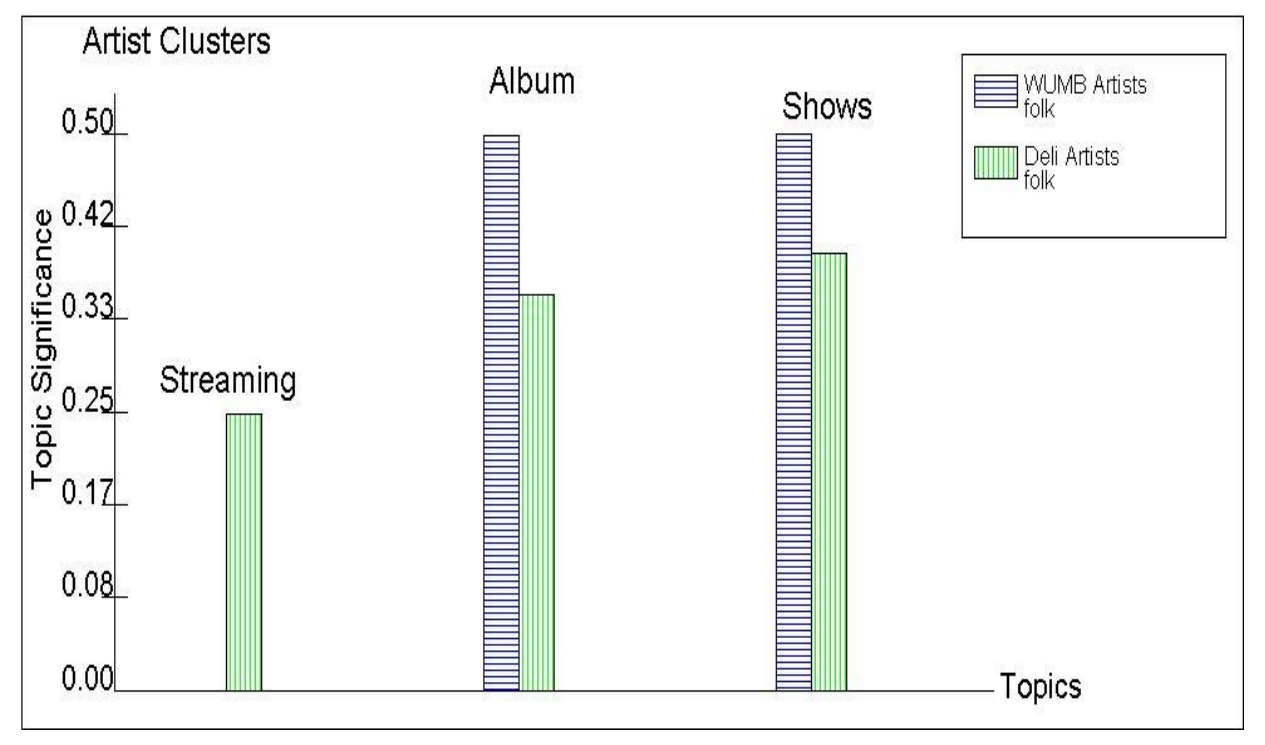

Fig 3. Deli and WUMB Folk Clusters

Conclusions

Our methodology of deriving significant keywords allows pLDA to be applied to the task of topic categorization quickly and accurately. Using our approach, we found that, in the folk music scene and elsewhere, online streaming services are relevant to artist popularity. The artists are turning to such new venues instead of the conventional approach of promoting albums through record labels. The results are particularly relevant to the WUMB radio station, as their artists are mostly categorized as folk music.

Acknowledgement

This material is based upon work supported by the President's 2017 Creative Economy (CE) Fund, University of Massachusetts.