Extended Topic Models with Numerical Features

Gökhan Çapan, Ali Caner Türkmen

March 23, 2016

Introduction

- Unsupervised learning, recover *latent* topics in documents
- Can be thought of as clustering. Loosely equivalent to link prediction.

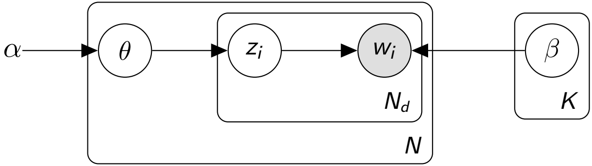
Latent Dirichlet Allocation

- Mixed Membership Model
- Assumes a fixed number of topics in a corpus
- A document includes words from multiple topics (in contrast with clustering)

LDA - Generative Model

- For each document;
  - Topic proportions vector is drawn ($\theta \sim Dirichlet(\alpha)$)
  - For each word in the document
    - A topic is drawn from topic proportions ($z_i \sim Multinomial(\theta)$)
    - The word is drawn from topic ($w_i \sim Multinomial(\beta_{z_i})$)

LDA - Bayesian Network Representation

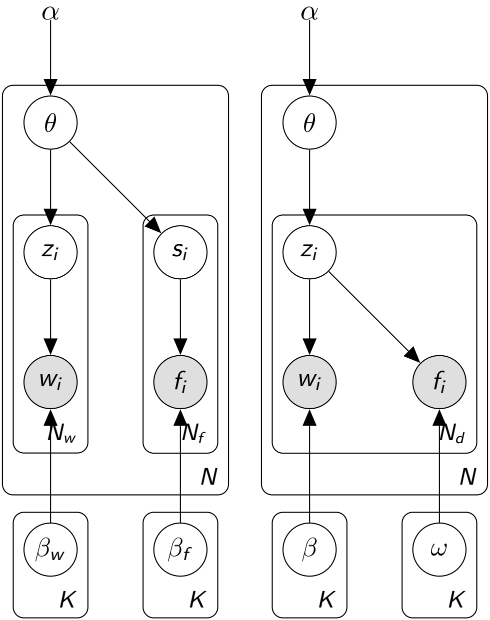


- Note the difference with document clustering (z is outside of the plate in that case, each word of a document comes from a single cluster), which is referred to as *mixture of unigrams*

LDA - Multi Modal (Aspect) Variants

- A topic generates not only words, but also other modalities
- The features can be paired with words themselves (for word sense, for example)
- The topics can generate other aspects

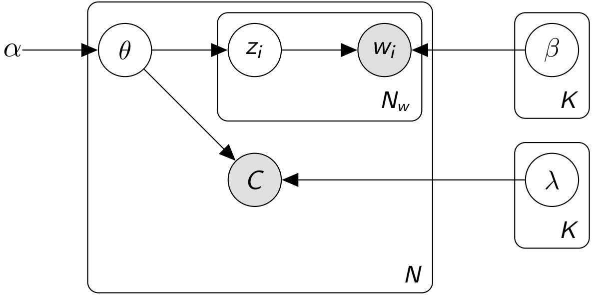
Multi-Modal LDA Variants



Proposed Model

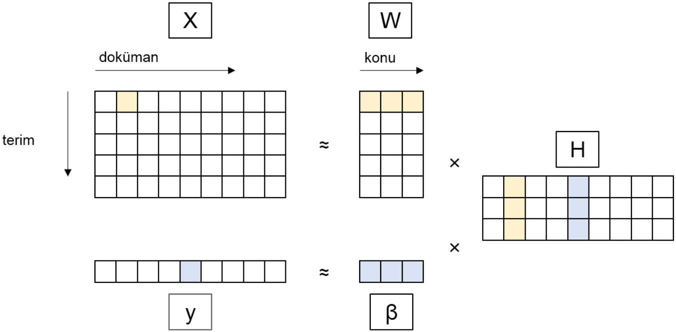
- We hypothesize that topics and etymological origins of words are related
- Using a multi-modal LDA, we can understand the relationship

A Tentative Approach



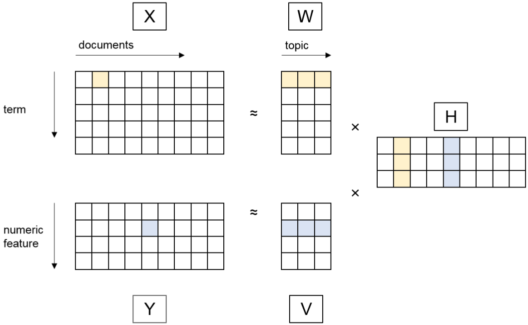
- Here, C is an $|L|$ dimensional vector of Poisson random variables, where L is the set of etymological origins

Coupled Matrix Factorization for Recovering Topics



- Represent data with well-known algebraic structures
- Jointly guide topic assignments from multimodal datasets, in a probabilistically sound framework
- Easily extensible to semi-supervised learning, kernel methods
- (T. and Cemgil, 2016)

Extended Coupled NMF for Topic Learning with Count Features



- $y_{ij}|W, H \sim GPO(\sum_t w_{it} h_{tj}, \phi)$
- $x_{kj}|V, H \sim GPO(\sum_t v_{kt} h_{tj}, \gamma)$
- Guide topic modeling with numeric count data
- Can assume priors $p(W), p(H), p(V)$ for Bayesian learning

Data Set and Features

- **Data Set:** News articles sampled from Anadolu Agency website. 1337 documents (can be expanded), 3000 tokens after adjusting for document frequency.
- **Features:** Complexity features such as word count, sentence count, average sentence length, comma count. (TBD)
- **Novel Features:** Etymological counts. Count the number of words from their etymological origins. Number of Arabic, Farsi, French words, etc. Source: TR Wiktionary Database Dump.

Learning

- EM-like updates with multiplicative NMF update rules
- Gibbs sampling assuming appropriate priors (tentative, out of scope for this project)

Conclusion

We propose two key contributions

- Put the topic modeling problem in a coupled NMF framework, extending with numerical features
- Use etymological counts for the Turkish language