Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Presented by Krithika Srinivasan

14 February 2019

Outline

- Introduction
- Methods Before LDA
- LDA
  - Explanation
  - The Maths
  - Inference
  - Parameter Estimation
  - Smoothing
- Results
- Conclusion
- Strengths of LDA and the paper
- Weaknesses of LDA and the paper
- Discussion Points

Introduction

What Is Topic Modelling?

What Is Topic Modelling?

- Short descriptions of members of a large collection of text data
- These descriptions help process large amounts of data
- Statistical relationships in the collection are still preserved

History of Topic Modelling

- Early methods in Information Retrieval represented documents in a corpus as vectors of real numbers which were counts of words
- Tf-idf built on that by creating a matrix of ratios of this same count to the respective inverse count
- Latent Semantic Indexing took the resulting matrix and performed Singular Value Decomposition on it. This revealed variations in the data respective to the tf-idf features
- From LSI, probabilistic LSI (pLSI) was derived

Probabalistic LSI

- pLSI is the precursor to LDA
- Its end result is a probability distribution on a fixed set of topics for every document
- Its main drawback is also that it can also show distributions for just one document

Terminology

- Words are the basic unit of discrete data in this paper
- A document is a sequence of words
- A corpus is a collection of documents

Latent Dirichlet Allocation

What does LDA Do?

An Example

- Here are five sentences
  - I like sunny weather
  - Yesterday was really cold
  - My dogs are fussy eaters
  - I love looking at pictures of lazy cats
  - My dog and I are terrified of thunderstorms

- Sentences 1 and 2 belong to Topic A
- Sentences 3 and 4 belong to Topic B
- Since there's only one topic, the distribution of these topics on these sentences is 100%
- Sentence 5 is a mix of both topics. Let's say sentence 5 is 50% Topic A and 50% Topic B
- Topics are distributions of words too
- So let's say Topic A is 20% sunny, 10%cold, 20% thunderstorms…i.e. to do with the weather
- And Topic B is 40% dogs, 20% cats, 10% fussy, 10% lazy etc. which pertains to animals

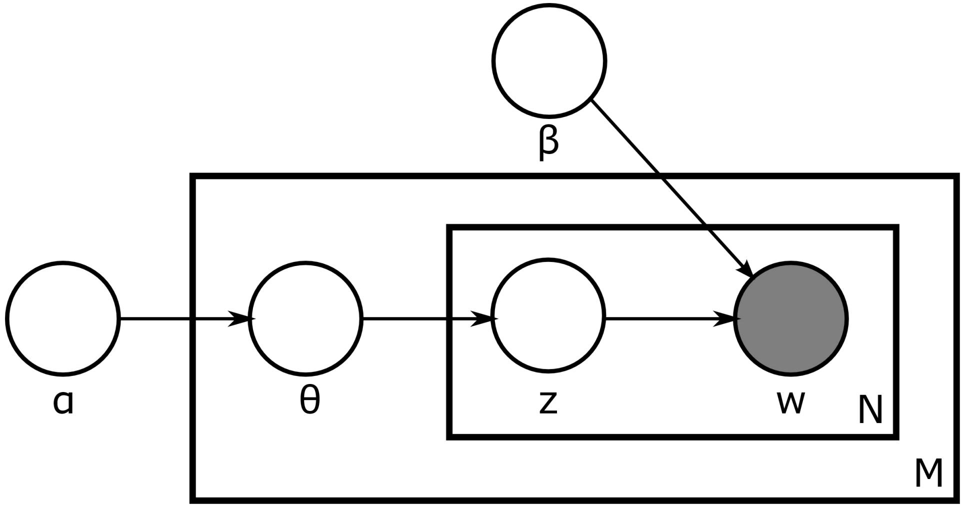
How does LDA work?

- The documents are produced with the following steps in LDA
- First, you decide on the number of words *N* the document will have. E.g. We want five words
- Then, you pick a topic mixture for the document, e.g. 70% Topic A and 30% Topic B
- Now, each word in the document from w_1 to w_N is generated by
  - First picking a topic based on the probability i.e. a 7/10 probability we pick the weather topic and a 3/10 probability that we go with animals
  - And then, we pick a word based on the probability distribution in each topic. For example, if the topic is weather, then there’s a 20% chance we generate the word ‘sunny’, 10% chance for ‘cold’

How does LDA work?

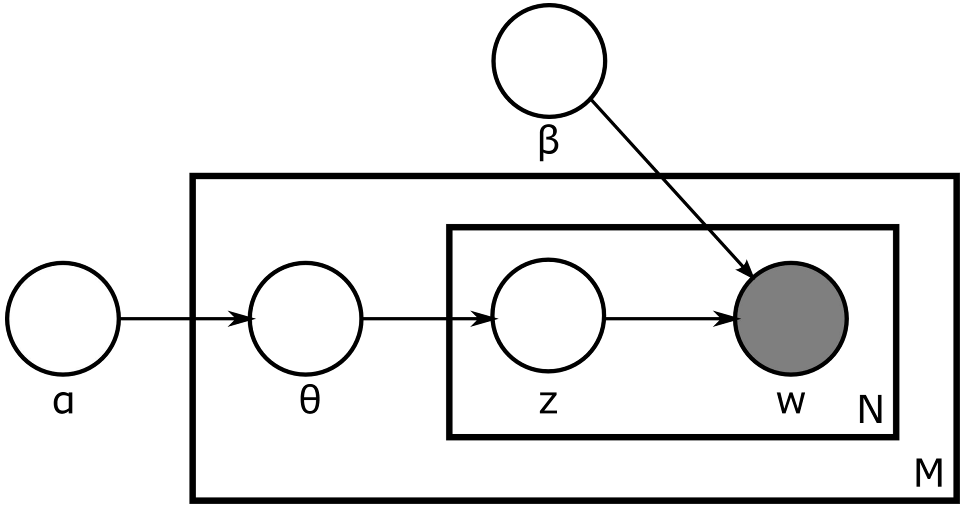
- The result is a bag-of-words document, since the order of the words doesn’t matter
- E.g “fussy cold dog sunny yesterday”

The Maths Behind LDA



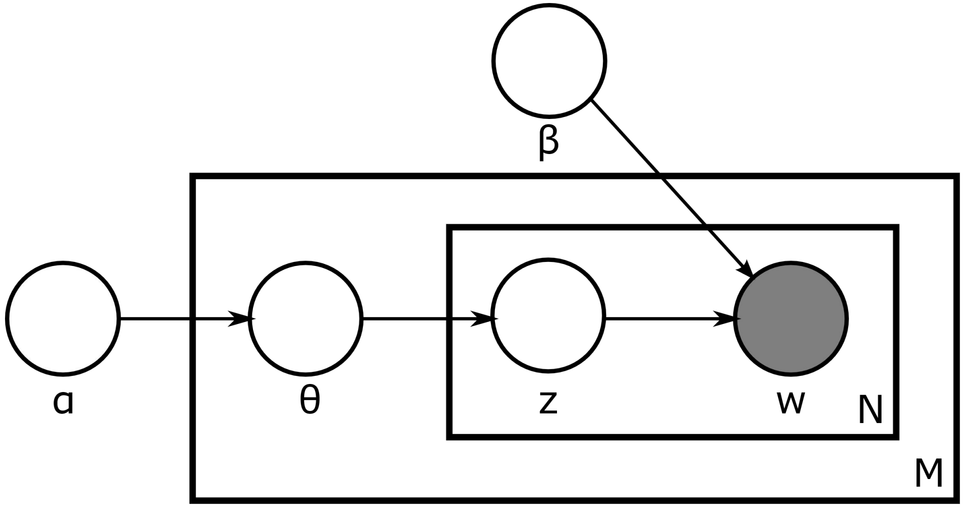
- α represents the variety of topics distributed across a document. A higher value means more variety
- β represents the distribution of words across a topic
- θ represents the distribution of topics across a document. Each document has a corresponding θ

The Maths Behind LDA



- z and w are both word-level parameters, i.e. they are sampled once for each word
- z is the topic associated with the word
- w is the word itself

The Maths Behind LDA



- This is a probabilistic graphical model (PGM)
- The model essentially assigns words to a topic based on a probability distribution of the topic across a document and the distribution of words across a topic

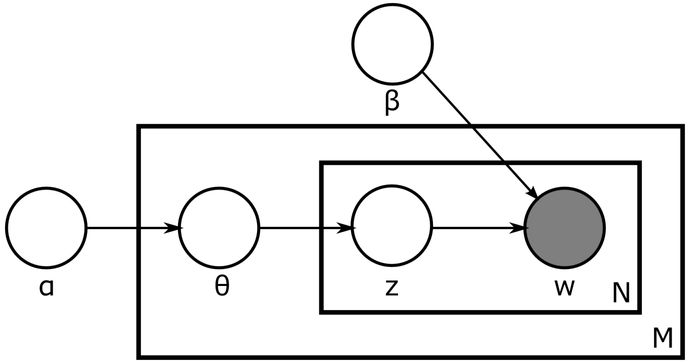
Inference, Parameter Estimation and Smoothing

Inference

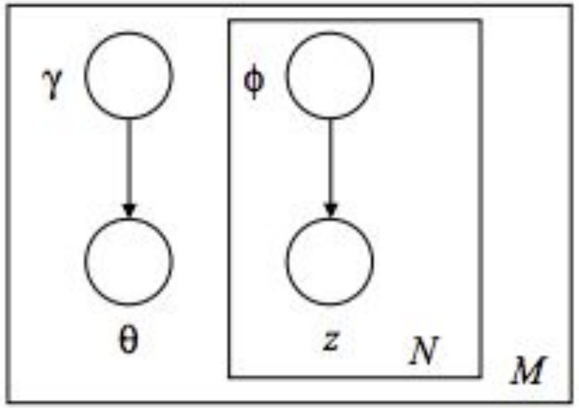
- θ and z are hidden variables in this model and cannot be calculated directly in part because of the awkward relationship between θ and β
- To get around this, a variety of inference algorithms can be used to approximate the posterior distribution of the hidden variables, i.e. p(θ, z | w, α, β)
- Some of these algorithms –
  - Laplace approximation
  - Variational approximation
  - Markov chain Monte Carlo
- The paper discuss a "simple" variational algorithm

Variational Inference

BEFORE



AFTER



Variational Inference

- γ is the Dirichlet parameter that represents concentration for the distribution
- Φ is the variational parameter. There is one such parameter for every word
- Essentially, for every word in every document we optimize Φ using γ and β iteratively until there is no change, i.e. convergence is reached

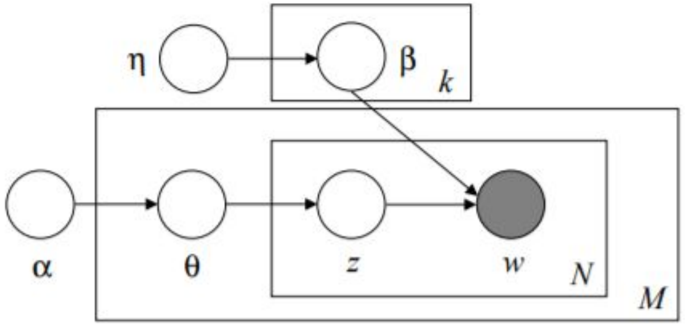
Variational Inference – What it means in Practice

- We decide how many topics we want in our model, say k = 3
- Initially, each word is equally likely to be assigned to a topic, i.e. the probabilities of the topics are equal
- First we iteratively optimize this allocation of words to topics
- Then we use this more accurate distribution to improve the distribution of topics across a document

Parameter Estimation

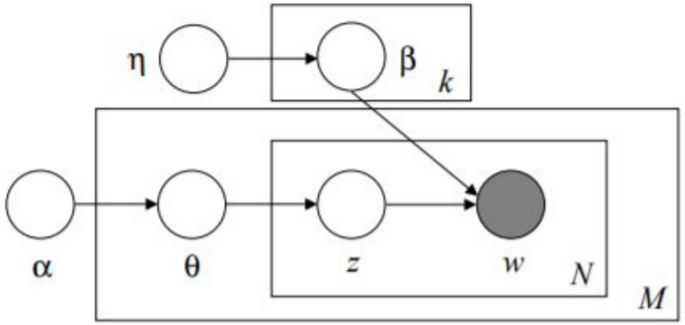
- The aim is to obtain the best possible values of α and β
- The best values of α and β maximize the log likelihood of the data i.e. $l(\alpha,\beta) = \sum^{M}_{d=1} \log p(w_d \mid \alpha,\beta)$
- The paper uses a variational expectation maximization technique which builds off the variational inference used prior to this
- We get $p(w_d \mid \alpha,\beta)$ from the inference method chosen which also gives us the lower bound for the log likelihood. This is the expectation step
- We then maximise this lower bound with respect to α and β
- These two steps are repeated until convergence

Smoothing



- Problem – How do we deal with words not in the training corpus?
- Solution - Assign a basic non-zero probability to all words in the vocabulary

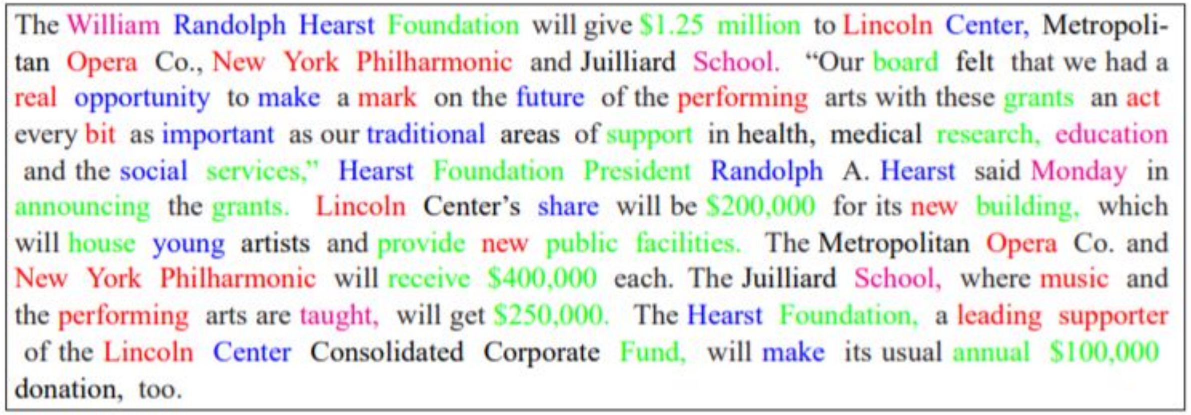
Smoothing



- β is now a k x V random matrix, i.e. each topic contains all the words in the vocabulary
- We use variational inference once again to approximate probabilities for each element in β i.e. the probability of each word being in that topic
- Functionally n replaces β in the parameter estimation process

Application and Results

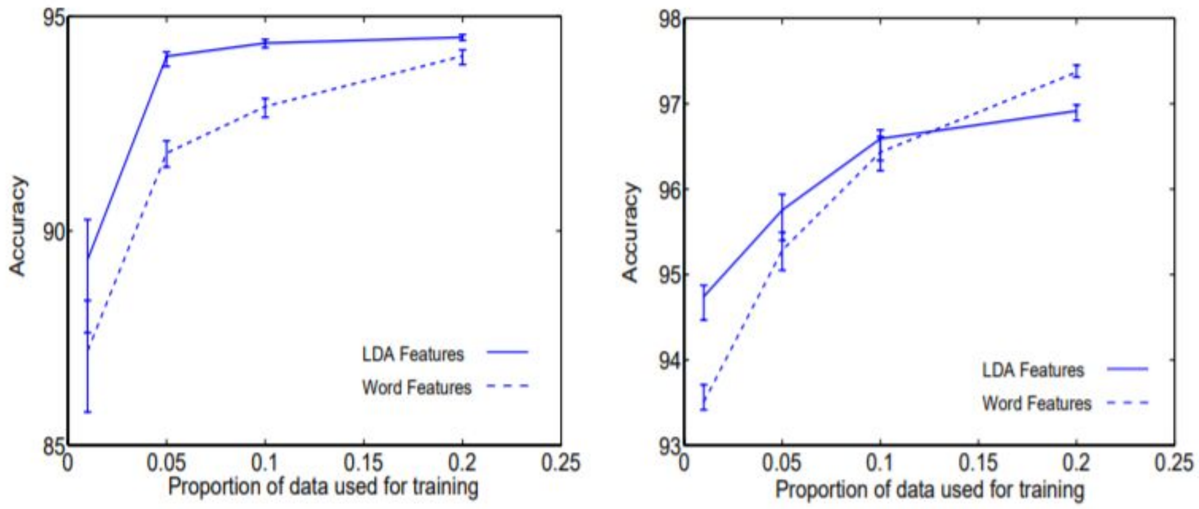
“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI



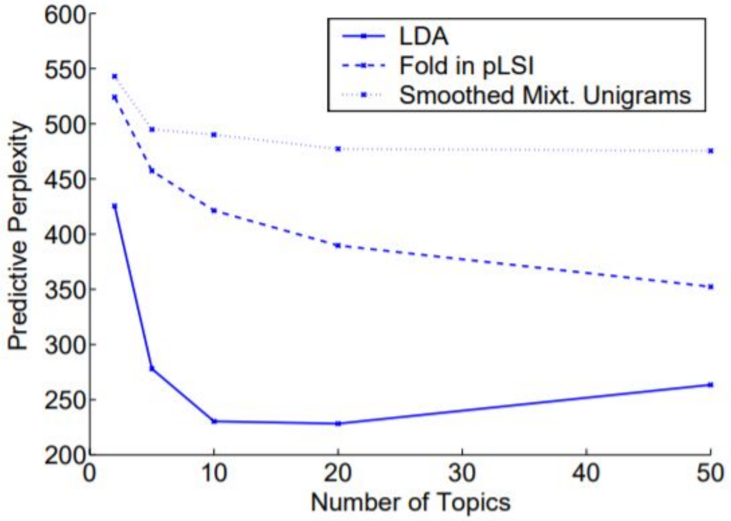
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Document Classification

LDA can be used as a classification algorithm



Collaborative Filtering



- User -> Document
- Movie -> Words
- Evaluated based on the likelihood assigned to the held-out movie

Conclusion

Concluding Points

- LDA can be viewed as a dimensionality reduction technique, but its underlying structure and semantics make sense for the data it models
- LDA can’t be calculated exactly, but a large number of approximation inference methods exist
- LDA has a modular nature and can therefore be extended and scaled up
- It can be embedded in more complex models

Discussion Points

- Positives of LDA
  - Its modularity
  - Logical progression based on its predecessors
- Not –So-Goods of LDA
  - Doesn’t factor in metadata or prior data (mitigated by the fact that LDA can be adapted)
  - Concerns over using perplexity as an evaluation parameter