Nonparametric Topic Modeling using Chinese Restaurant Franchise with Buddy Customers

Shoaib Jameel

Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong.

Co-authors: **Wai Lam** (Advisor) and **Lidong Bing**

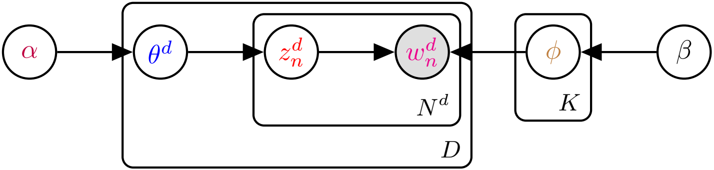
Overview

Probabilistic Topic Models

What is...?

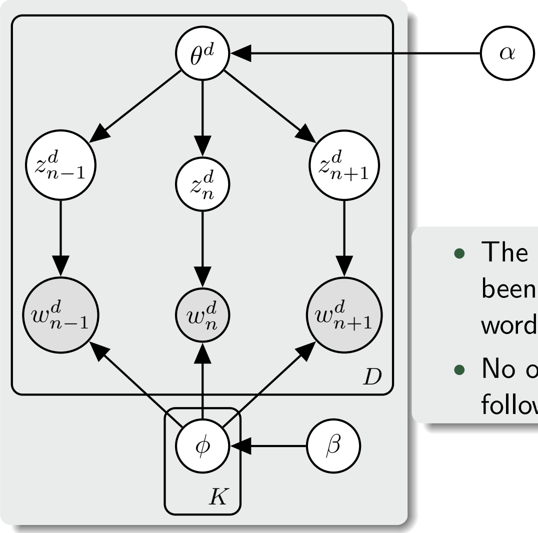
- **Graphical Model**: A graph representing the conditional dependence between different random variables.
- **Topic Model**: Models that discover latent semantic structure in the document collection.
- **Topic**: Probability distribution over words in the vocabulary.
- **Parametric Topic Models**: Topic models that have a fixed dimensional functional form. The dimensionality or the number of topics is specified by humans.
- **Nonparametric Topic Models**: Topic models which automatically uncover the number of dimensions or number of topics based on the data characteristic.

Latent Dirichlet Allocation (LDA) [Blei et al., 2003]



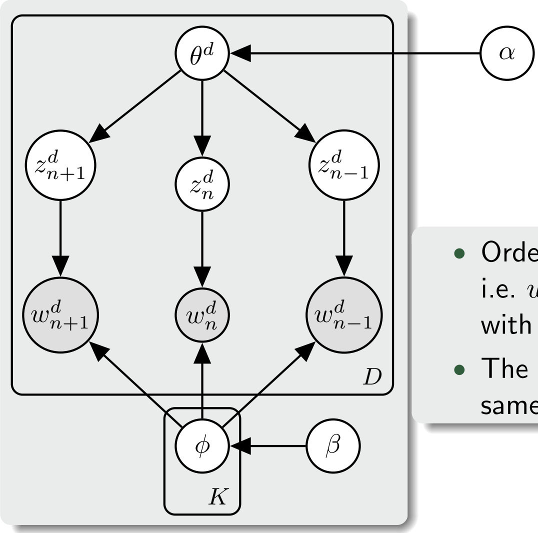
Document prior from Dirichlet	Document specific topic distribution	Topic assignment	Observed word	Word topic assignment matrix	Word prior from Dirichlet

Latent Dirichlet Allocation (LDA) - Alternate View



- The graphical model has been expanded to show words in plate notation.
- No order of words is followed.

Latent Dirichlet Allocation (LDA) - Alternate View



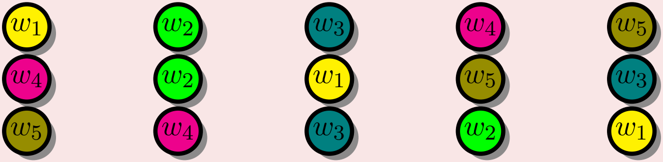
- Order of words is changed i.e. w_{n-1}^d is exchanged with w_{n+1}^d.
- The model is still the same.

Assumptions used in Topic Models

Two assumptions which many probabilistic topic models currently rely upon:

- **Bag-of-words** - Order of words is not taken into account.
- **Number of Topics** - User pre-defines this discrete value. As a result, the model restricts itself only to these latent topics.

Bag-of-words Illustration



$P(w_1, w_2, w_3, w_4, w_5) = P(w_4, w_2, w_1, w_5, w_3) = P(w_5, w_4, w_3, w_2, w_1)$

Bag-of-Words Assumption

Limitations

- Document's semantic structure is not taken into account.
- Models cannot be used in applications such as speech recognition, text compression, etc.

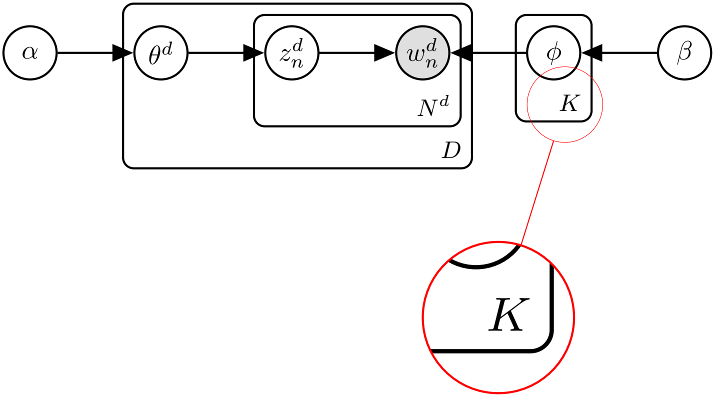
First Order Markov Chain

Illustration of word order.

Mathematical Depiction

$P(w_1, w_2, w_3, w_4, w_5) \neq P(w_2, w_1, w_5, w_4, w_3)$

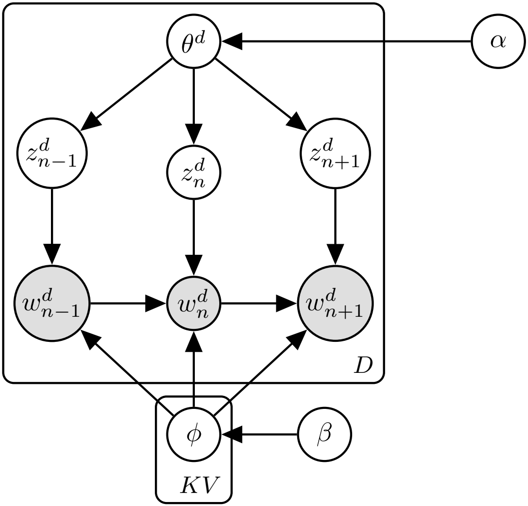
Number of Topics (K)



Limitation

User has to specify the value of K. Fixed dimensional parameter space.

Bigram Topic Model (BTM) [Wallach, 2006]

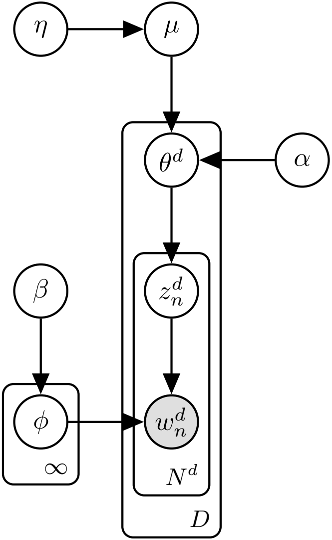


Model-based Learning to Find K

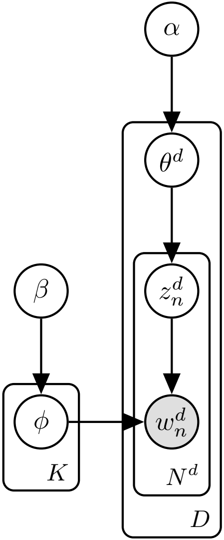
Bayesian Nonparametrics in Topic Models

1. Nonparametric Bayesian topic models are models on an **∞-dimensional** parameter space
2. Nonparametric topic models do not assume a restricted functional form
3. They do have parameters. Such models allow **complexity to grow** with the data
4. **Hierarchical Dirichlet Processes (HDP)** model is a nonparametric topic model which **automatically discovers** the number of topics K

Hierarchical Dirichlet Processes (HDP) [Teh et al., 2006]

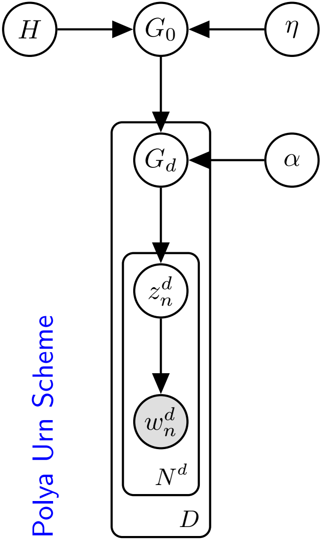


HDP

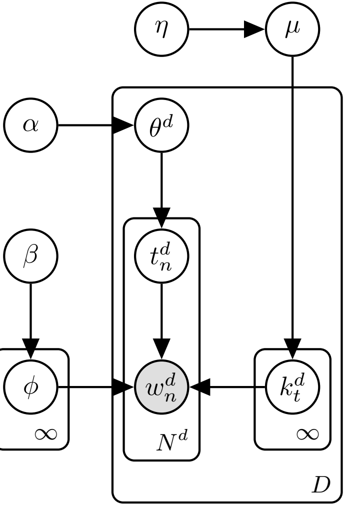


LDA

Hierarchical Dirichlet Processes (HDP)



Polya Urn Scheme



Chinese Restaurant Franchise Scheme

Chinese Restaurant Franchise (CRF)

- A metaphor used to describe the HDP
- Introduces the sharing property among clusters
  - Allows multiple restaurants to share common menu, which has a set of dishes
  - A restaurant has infinite tables, each table has only one dish
- Hierarchical paradigm to model clustering of data

Restaurant := Document   Customer := Word

Dish := Topic

Chinese Restaurant Process (CRP)

- Metaphor to describe the Dirichlet Process

Dirichlet Process

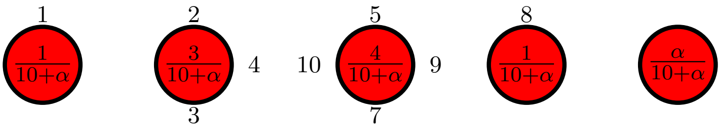
Dirichlet Process (DP) is a distribution over distributions

Mechanism Behind CRP

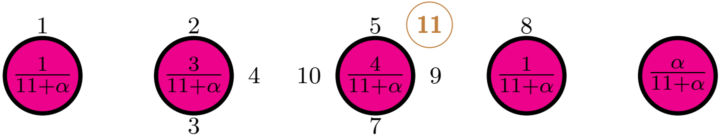
1. Imagine a restaurant with an infinite number of tables
2. First customer sits at the first table
3. Customer i does the following:
    1. Sits at previously occupied table
    2. Sits on a new table

Chinese Restaurant Process (CRP)

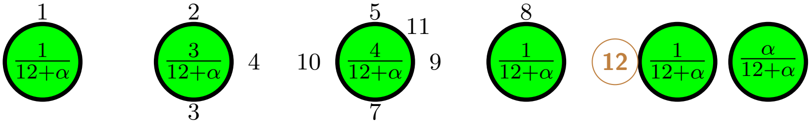
Starting table configuration



New customer comes in



New customer, new table



Chinese Restaurant Franchise (CRF)

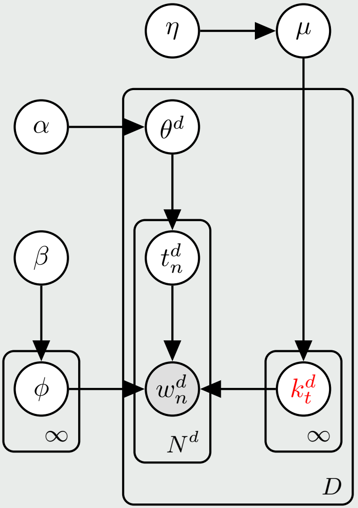
- Two-stage Chinese Restaurant Process (CRP)
  1. Customers choose tables in each restaurant
  2. Dishes are assigned to tables among all restaurants

Mechanism Behind CRF

- A franchise where restaurants share a menu with infinite number of dishes
- First customer of each table orders a dish which is shared by all future customers sitting at that table. Different restaurants can serve the same dish.
- When a new customer comes in the restaurant, the customer either,
  1. Sits at previously occupied table
  2. Sits on a new table

Chinese Restaurant Franchise (CRF)

HDP Model in CRF Representation



Limitations

Order of words is not maintained

Symbol Definitions

- k_t^d - Global dish to table assignment variable. Symbolizes sharing between clusters.
- Customers sit a tables. Table assignments indicate topic assignment.

Chinese Restaurant Franchise with Buddy Customers

Overview

- We propose a new **non-exchangeable** metaphor in Bayesian nonparametrics
- We introduce a notion of **Buddies** (friends) in the CRF metaphor
- We introduce a notion of **reserved** and **unreserved** tables
- Customers enter restaurant following **word order** in the document
- Buddies are assigned based on the global word co-occurrence statistics

Chinese Restaurant Franchise with Buddy Customers

Mechanism

1. Some of the customers have **pre-planned** their visit so that they can spend time together with their **good old buddies**
2. These buddies have already **reserved** their tables beforehand
3. Customers wait in the queue outside the restaurant in the same order as that of the words in a document
4. There might be **loners** as in CRF
5. Every customer carries with herself a **table**, a **buddy** and **word order** assignments

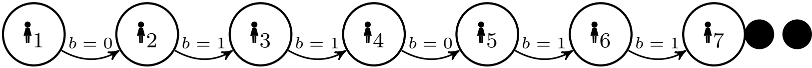
Buddy Allocation

How often two words commonly occur in sequence?

1. Is based on the global co-occurrence information
2. If two friends commonly hang-out most of the time, they are buddies
3. In language modeling parlance, we are finding the probability for bi-gram formation
4. We introduce a buddy assignment variable b in the model to find buddies
5. Buddy assignment variable is a binary random variable

Restaurant Level Depiction with Buddies

Customer configuration (a text document)



Buddy Customers Assignment

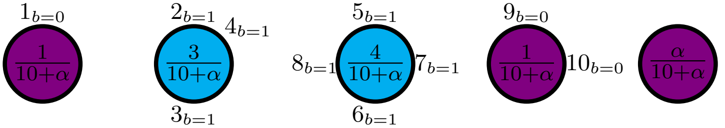
Loners: $_1$, $_9$, $_{10}$

Buddy Group 1: $_2$, $_3$, $_4$

Buddy Group 2: $_5$, $_6$, $_7$, $_8$

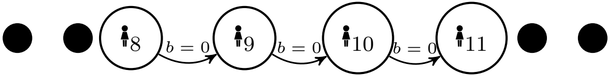
Restaurant Level Depiction with Buddies

Starting table configuration



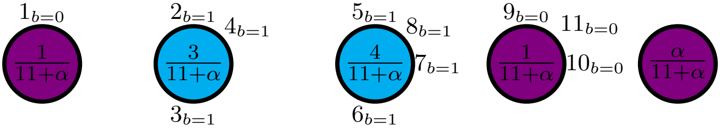
Unreserved Table Reserved Table

New customer ($_{11}$) comes in with $b_{(10,11)} = 0$

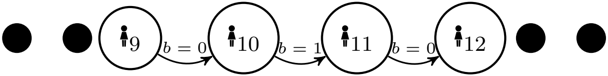


Restaurant Level Depiction with Buddies

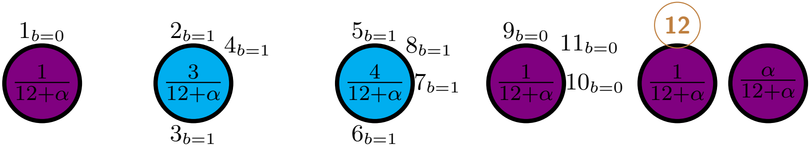
Updated Seating Arrangement



New customer (12) comes in with $b_{(11,12)} = 0$



New customer, new table



Experiments and Results

Evaluation Focus

1. Quantitative Analysis
   - Generalization ability on unseen data an
2. Qualitative Analysis
   - Top-k topic words

Datasets

1. AQUAINT-1 (TREC HARD track) - 1,033,461
2. NIPS Papers - 1,830
3. OHSUMED (Medical) - 233,448
4. Reuters - 806,791

Dataset size

It can be seen that some of the datasets are fairly large.

Comparative Methods

Parametric Topic Models

1. Latent Dirichlet Allocation (LDA) [Blei et al., 2003]
2. Bigram Topic Model (BTM) [Wallach, 2006]
3. LDA-Collocation Model (LDA-COL) [Griffiths et al., 2007]
4. Topical N-gram (TNG) [Wang et al., 2007]
5. N-gram Topic Segmentation Model (NTSeg) [Jameel and Lam, 2013b]

Nonparametric Topic Models

1. Hierarchical Dirichlet Processes (HDP) [Teh et al., 2006]
2. N-gram HDP (NHDP) [Jameel and Lam, 2013a]

Tuning process was used to determine the number of topics automatically in parametric topic models.

Results

Model	Perplexity			
	AQUAINT-1	NIPS	OHSUMED	Reuters
LDA	4599.48	834.45	2305.32	3490.12
BTM	4578.57	833.75	2229.96	3411.98
LDACOL	4501.44	831.45	2398.22	3298.76
TNG	4423.76	828.32	2315.72	3108.43
NTSeg	4400.76	811.32	2295.72	3112.43
HDP	4322.32	825.43	2240.23	3192.54
NHDP	4495.32	820.56	2299.45	3102.53
Our	4107.75	766.90	2192.44	3089.44

Note

Lower perplexity value signifies better performance

Qualitative Results

HDP	NDHP		Our	
	Unigrams	N-grams	Unigrams	N-grams
year	test	internet sale	phone	web site
game	computer	search engine	digit	cell phone
music	year	create search engine	computers	high technology
computer	project	internet user	technology	microsoft windows
train	modern	index html	information	computer technology
new	service	state department	web	computer device
team	software	computer software	mail	laptop equipment
church	internet	computer bulletin	user	recognition software
transit	editor	latin america	online	large comfortable keyboard
time	technology	talk real person	network	speech technology

Note

The above words are top 10 high probability words in a topic

The Chinese University of Hong Kong

Qualitative Results

HDP	NDHP		Our	
	Unigrams	N-grams	Unigrams	N-grams
report	year	oil product	oil	oil price
bank	japan	crude oil	trade	gulf war
win	iraq	new oil product	cargo	oil stock
pakistan	oil	january february	high	crude oil
oil	crude	saudi arabia	market	domestic crude
rate	demand	total product	price	iraq ambassador
net	gasoline	crude export	fuel	oil product
french	saudi	gasoline distillation	tonne	indian oil
launch	arabia	thousand barrel	crude	run oil company
qatar	uae	oil import	week	world price

Conclusions

- Automatically determining model complexity improves performance
- Word order improves performance

Why word order helps?

- Captures semantic storyline of document

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent Dirichlet allocation.
JMLR, 3:993–1022.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007).
Topics in semantic representation.
Psychological Review, 114(2):211.

Jameel, S. and Lam, W. (2013a).
A nonparametric n-gram topic model with interpretable latent topics.
In *AIRS*, pages 74–85.

Jameel, S. and Lam, W. (2013b).
An unsupervised topic segmentation model incorporating word order.
In *SIGIR*, pages 472–479.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).
Hierarchical Dirichlet processes.
JASA, 101(476).

Wallach, H. M. (2006).
Topic modeling: beyond bag-of-words.
In *ICML*, pages 977–984.

Wang, X., McCallum, A., and Wei, X. (2007).
Topical N-grams: Phrase and topic discovery, with an application to Information Retrieval.
In *ICDM*, pages 697–702.