Introduction to Latent Dirichlet Allocation (LDA)

Swapnil Hingmire

Indexing Text Documents

- Important problems in Information Retrieval (IR)
  - Given some keywords find relevant documents
  - Given a document find its similar documents
  - Identify major themes underlying a document corpus
  - Classify documents according to these themes

Indexing Text Documents

- Important problems in Information Retrieval (IR)
  - Given some keywords find relevant documents
  - Given a document find its similar documents
  - Identify major themes underlying a document corpus
  - Classify documents according to these themes
- Text as Data

Latent Dirichlet Allocation (LDA)

- LDA is a probabilistic generative model:
  Assumes a procedure to generate a document using simple probabilistic rules:

  (1) Choose a distribution over T topics
  (2) For each word position in the document,
  choose a topic randomly
  choose a word from the topic's distribution over the vocabulary

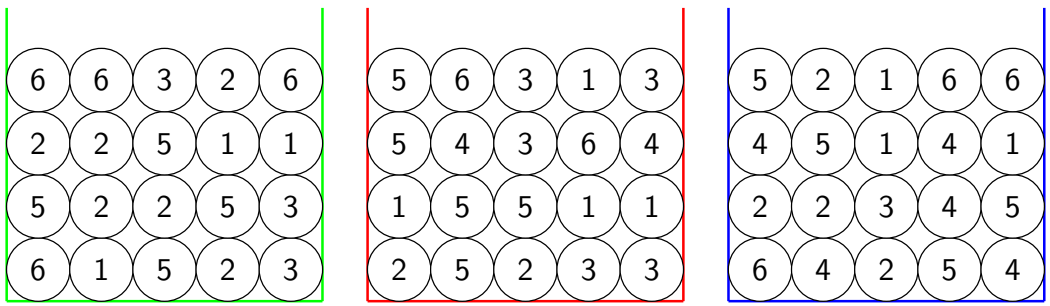
LDA : Pólya Urn Interpretation

(1) Let T be the number of topics
(T is **known** and **fixed** while generating documents).

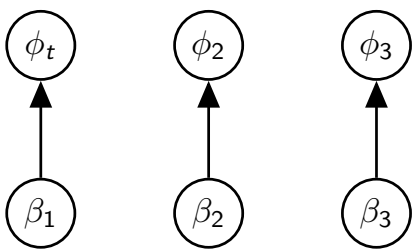
(2) A topic urn:
- It contains N_V types of balls of the same size
(N_V is the vocabulary size of the corpus D)
- Topics are assumed to be **fixed** while generating documents
and to be **inferred** at the time of inference.

(3) A document urn:
- It contains balls of T different colours,
such that each colour represents a topic.
- Initially, α_t balls of colour c_t are added to the urn

Generative Process of LDA

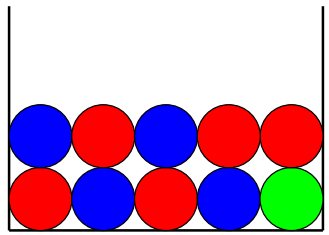


Topic Urn (ϕ): 1 Topic Urn (ϕ): 2 Topic Urn (ϕ): 3



1. Select word probabilities for each topic

Generative Process of LDA

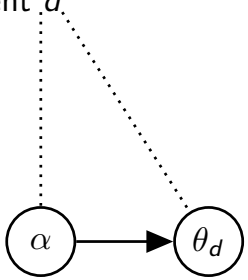


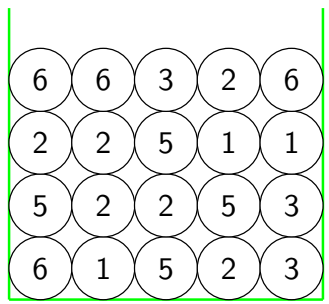
Document-Topic Urn (θ)

$z_{d,n}$ =

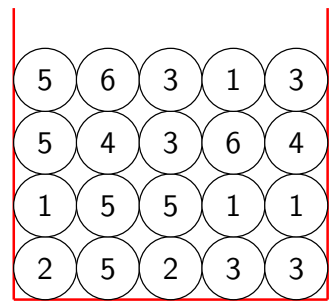
$w_{d,n}$ =

2. Select topics probabilities for document d

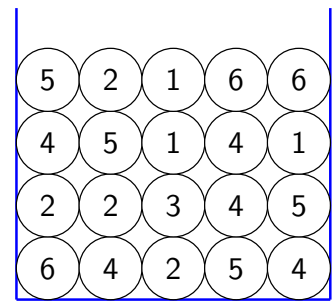




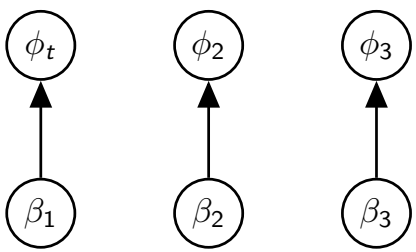
Topic Urn (ϕ): 1



Topic Urn (ϕ): 2

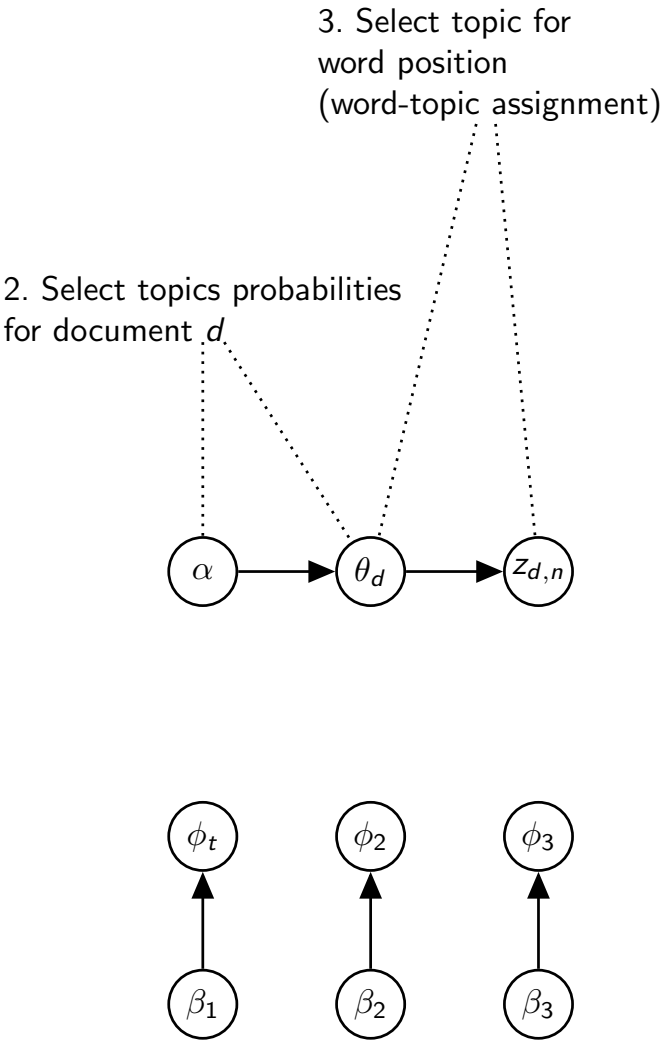


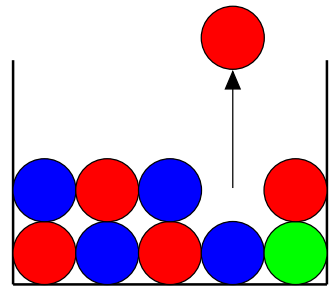
Topic Urn (ϕ): 3



1. Select word probabilities for each topic

Generative Process of LDA

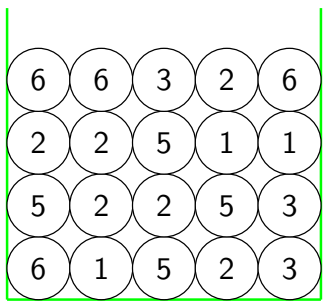




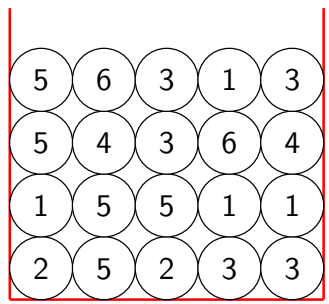
Document-Topic Urn (θ)

$z_{d,n}$ =

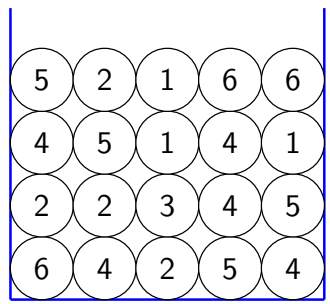
$w_{d,n}$ =



Topic Urn (ϕ): 1

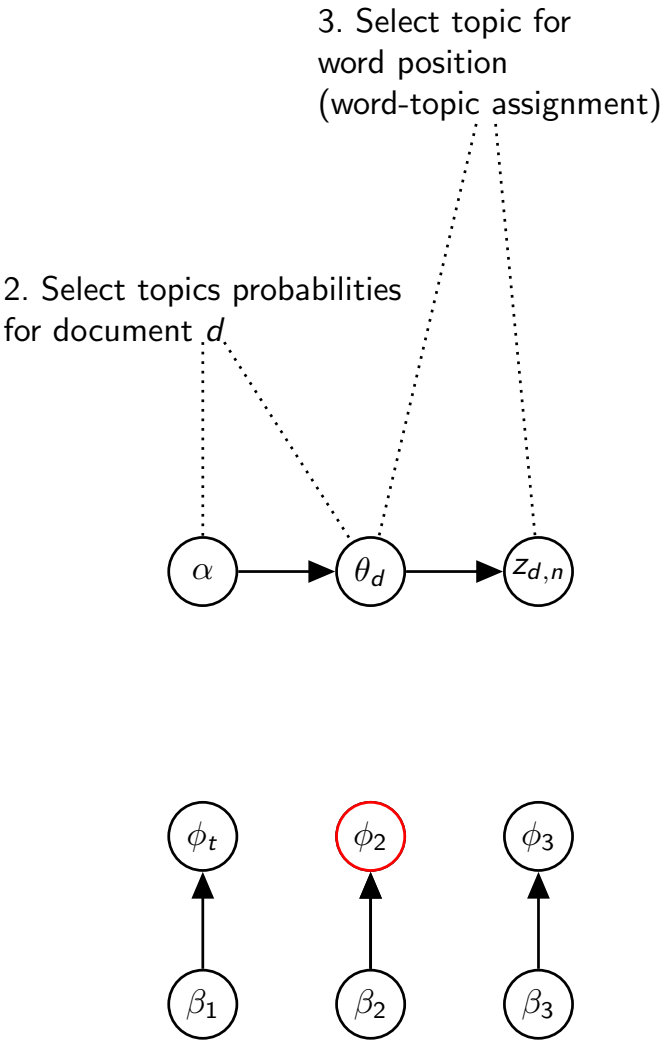


Topic Urn (ϕ): 2

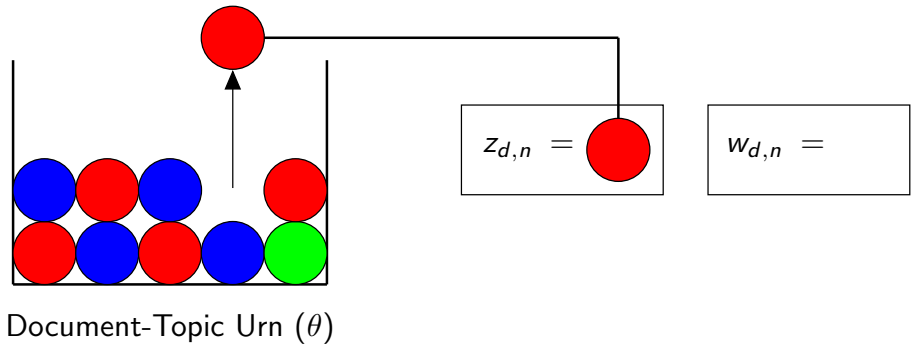


Topic Urn (ϕ): 3

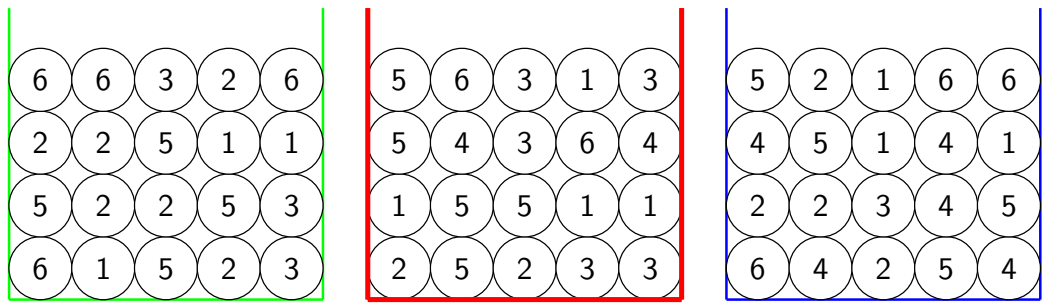
Generative Process of LDA



1. Select word probabilities for each topic

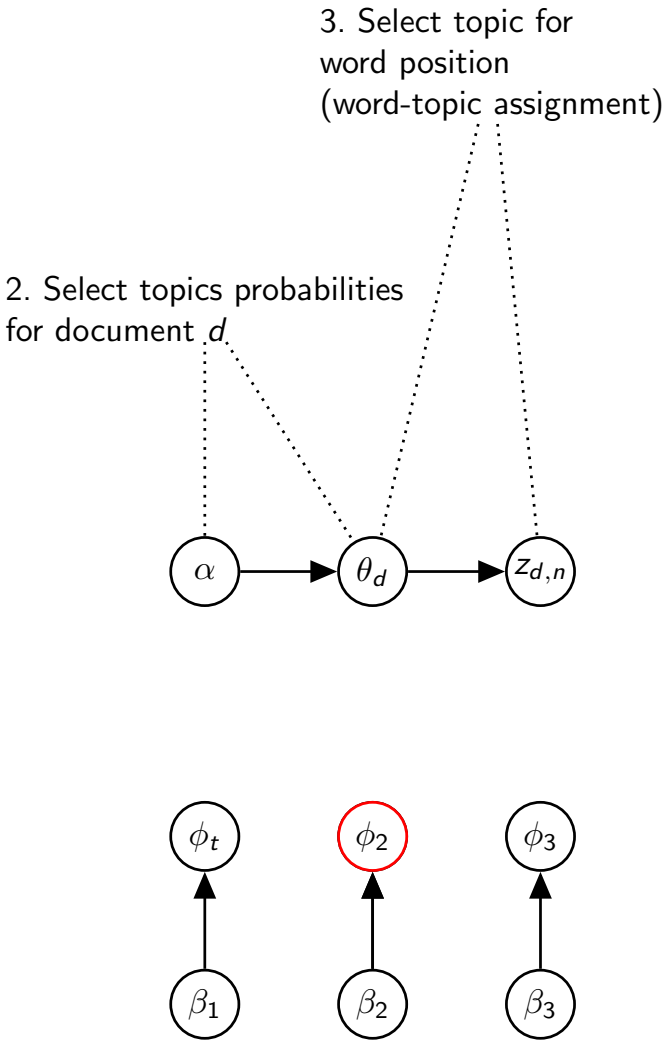


Document-Topic Urn (θ)

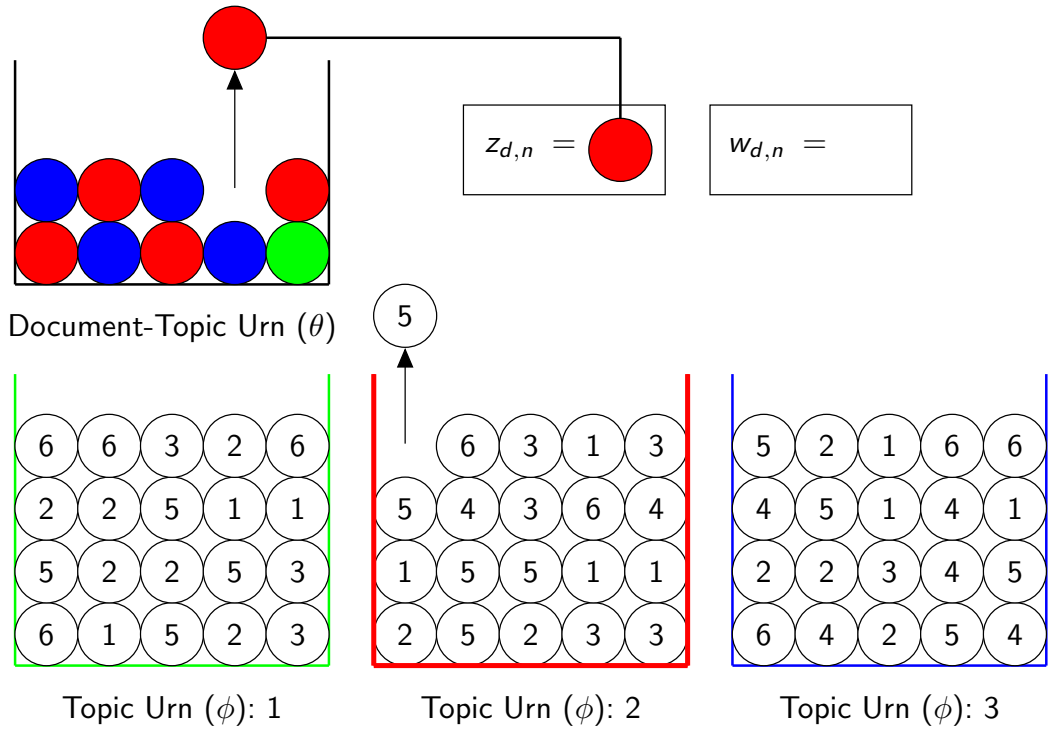


Topic Urn (ϕ): 1 Topic Urn (ϕ): 2 Topic Urn (ϕ): 3

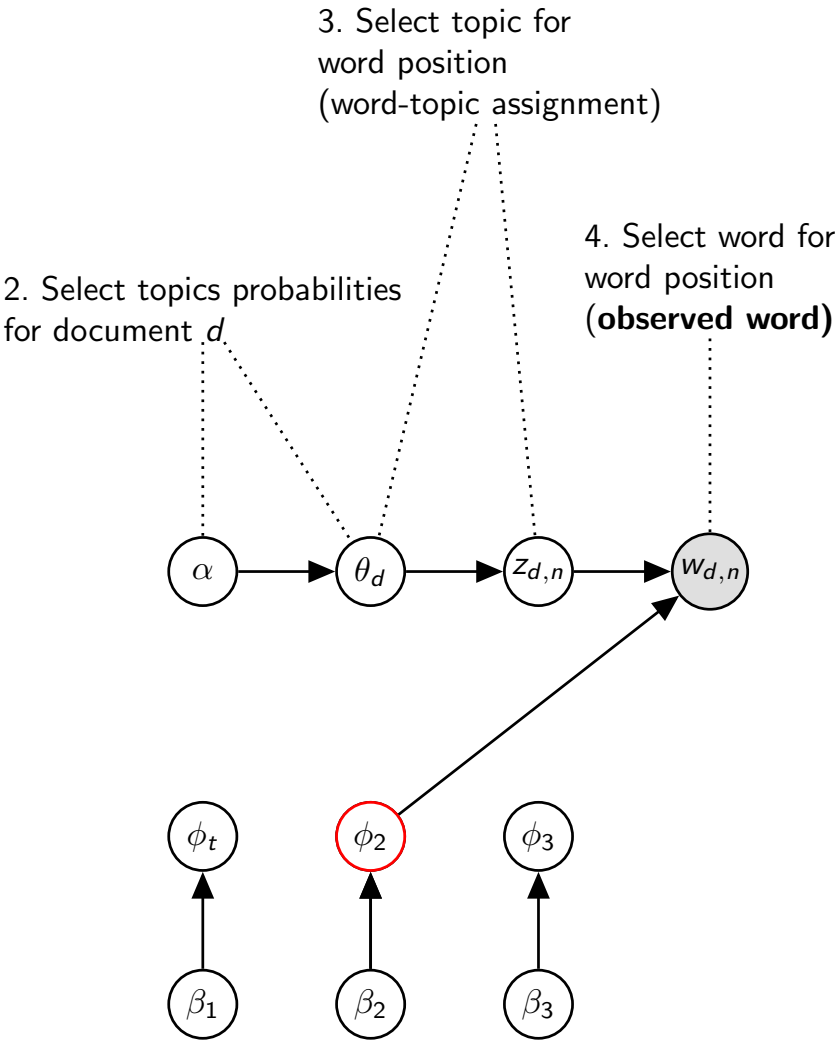
Generative Process of LDA

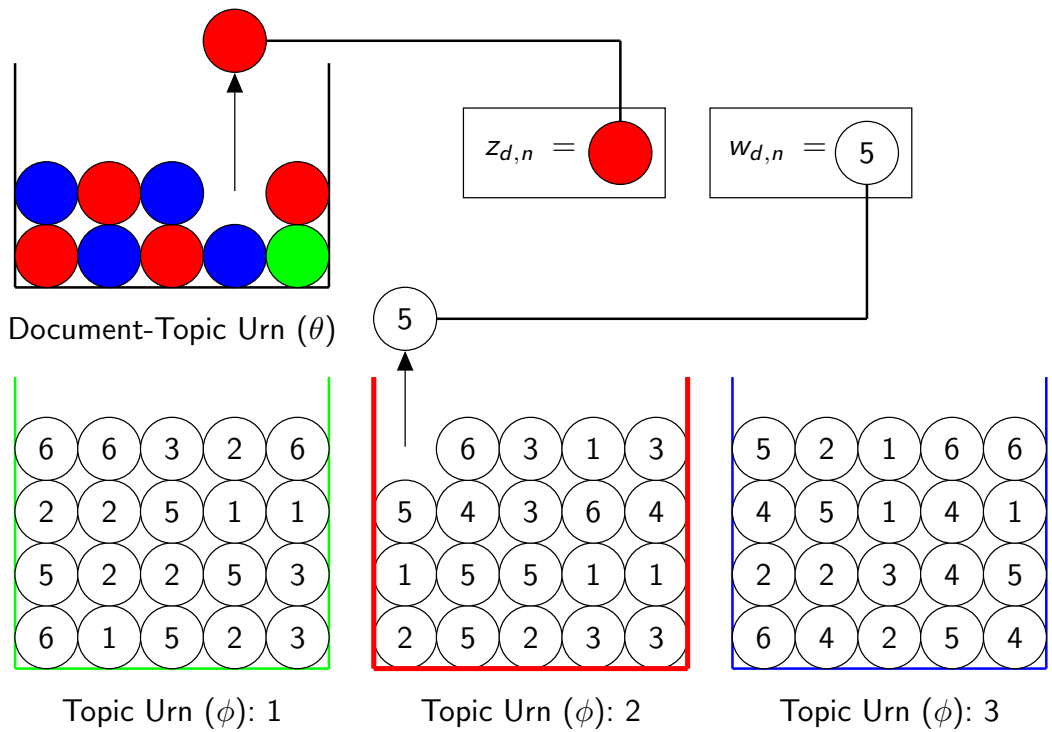


1. Select word probabilities for each topic



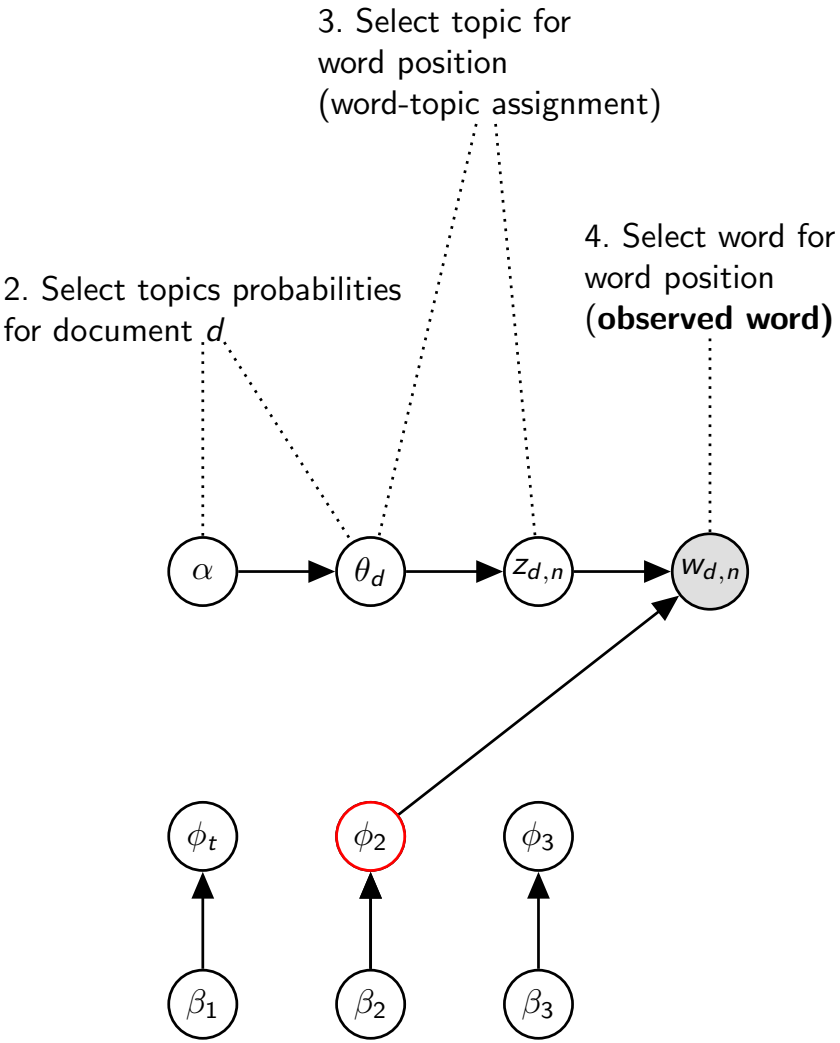
Generative Process of LDA

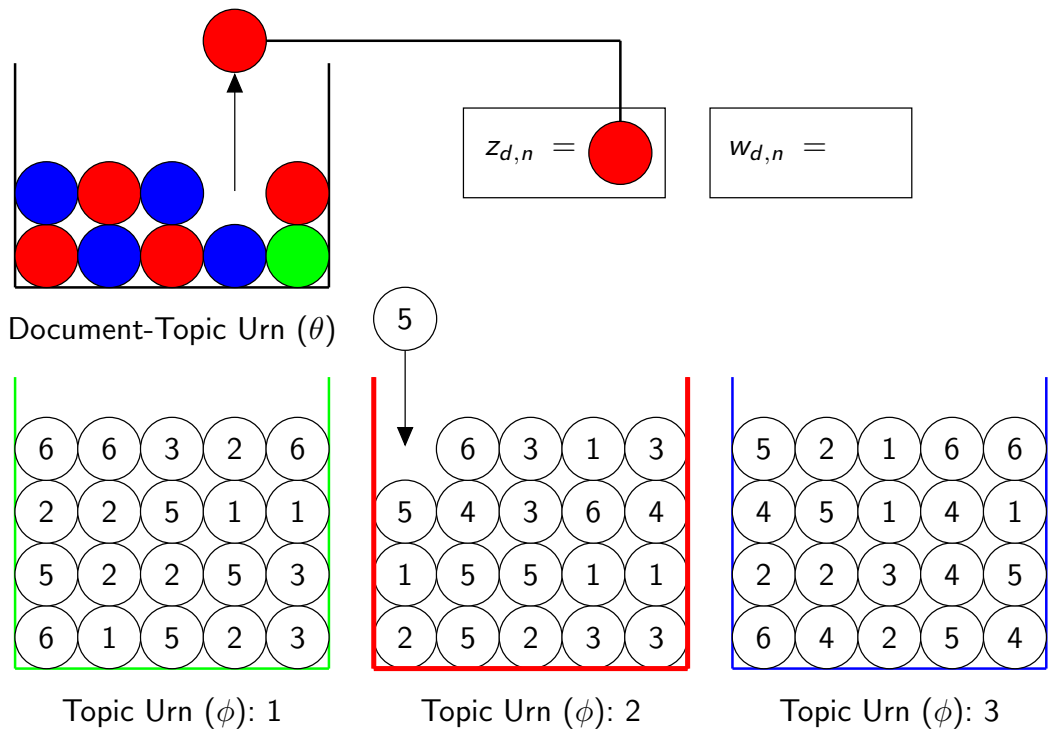




1. Select word probabilities for each topic

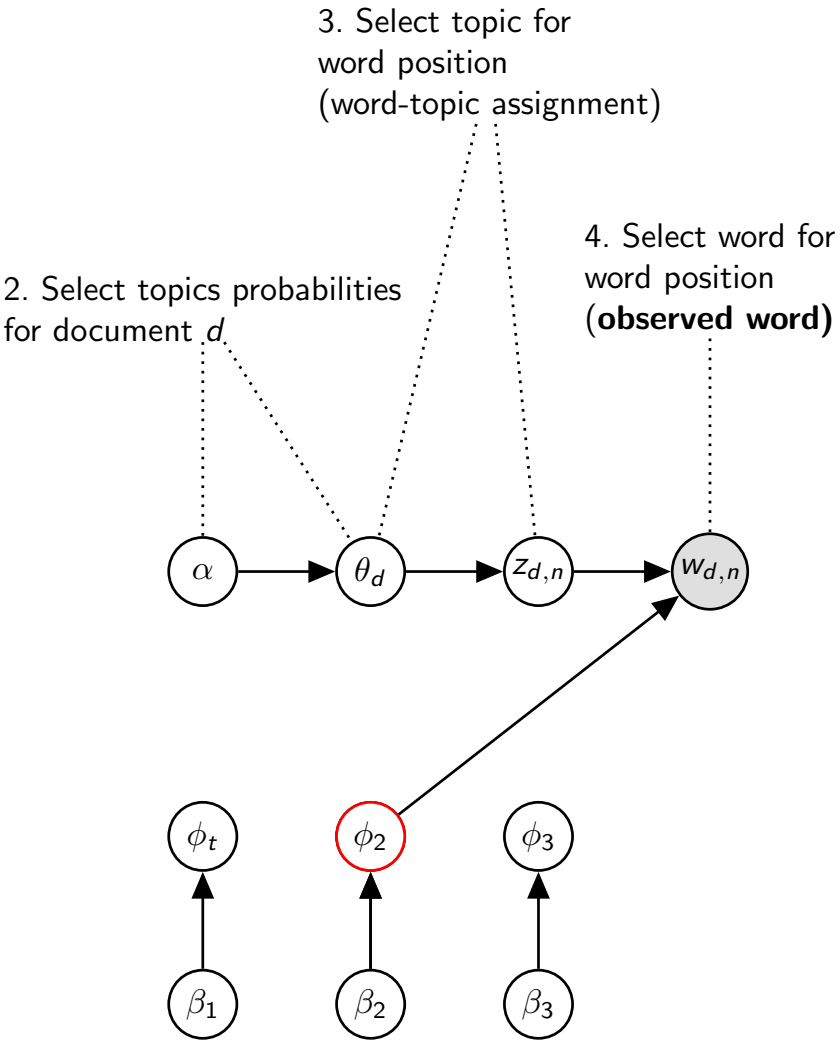
Generative Process of LDA



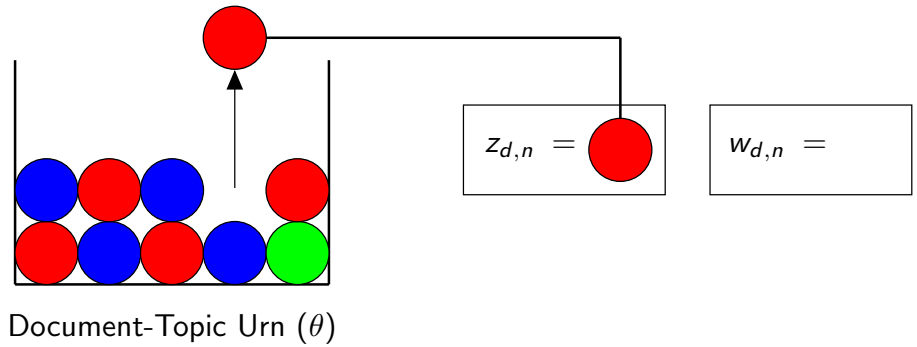


1. Select word probabilities for each topic

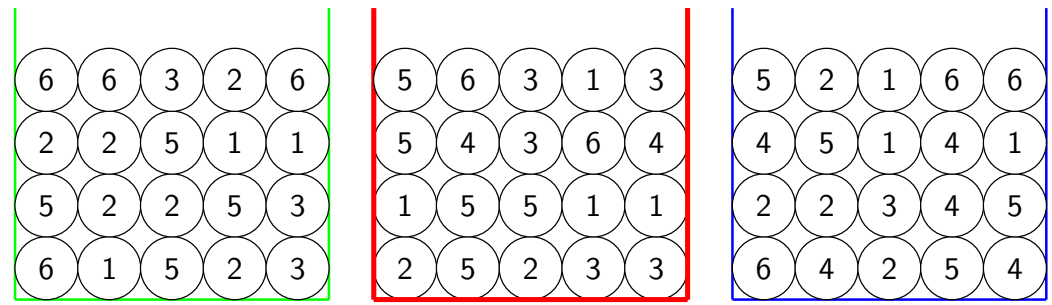
Generative Process of LDA



1. Select word probabilities for each topic

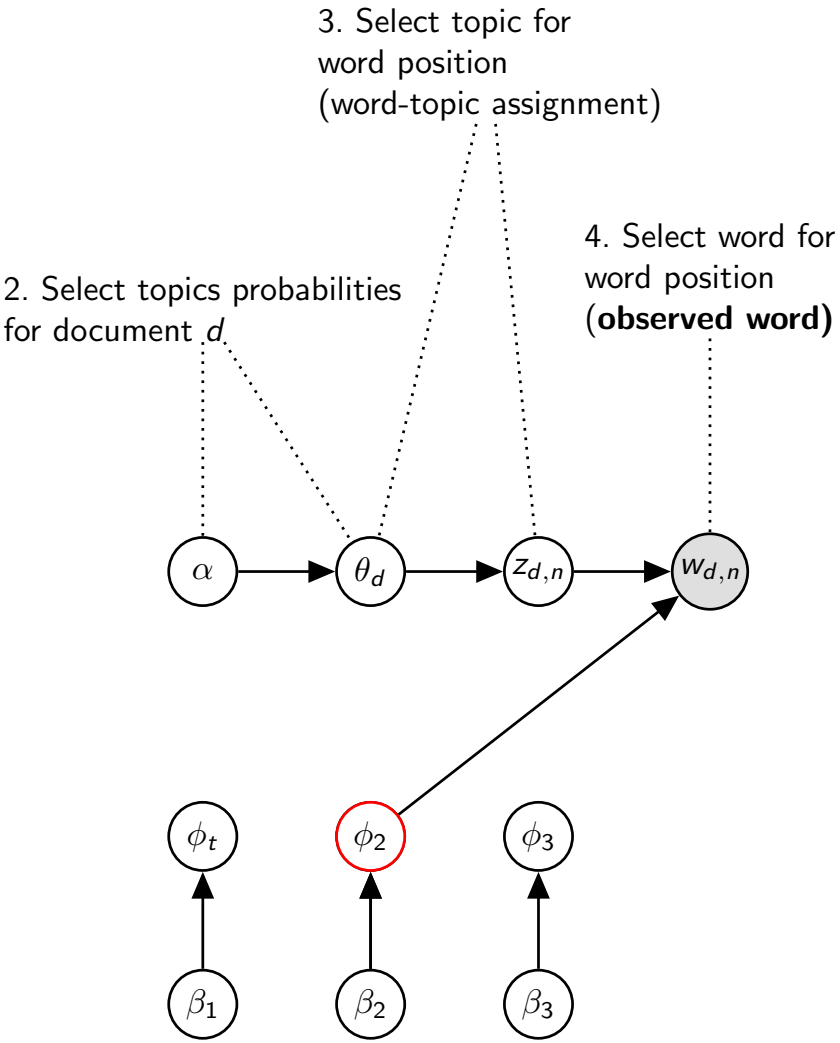


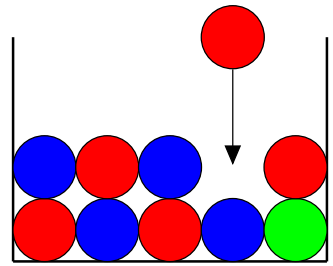
Document-Topic Urn (θ)



Topic Urn (ϕ): 1 Topic Urn (ϕ): 2 Topic Urn (ϕ): 3

Generative Process of LDA

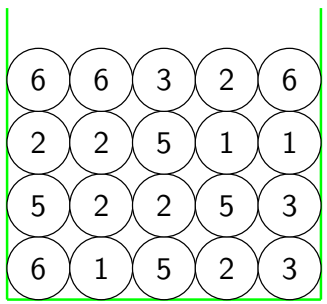




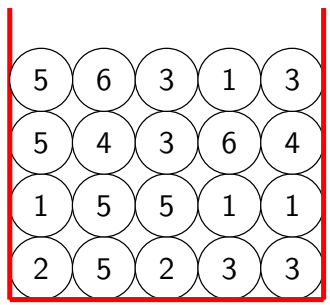
Document-Topic Urn (θ)

$z_{d,n}$ =

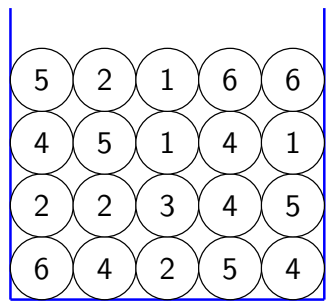
$w_{d,n}$ =



Topic Urn (ϕ): 1



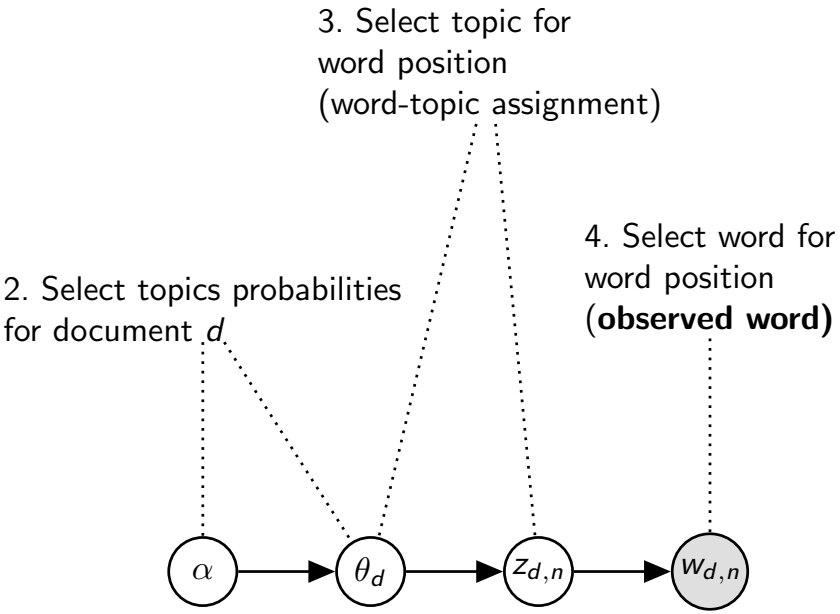
Topic Urn (ϕ): 2

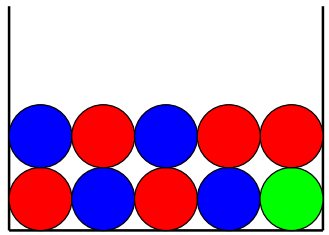


Topic Urn (ϕ): 3

1. Select word probabilities for each topic

Generative Process of LDA

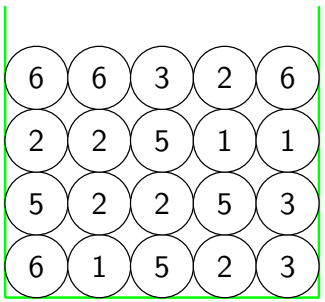




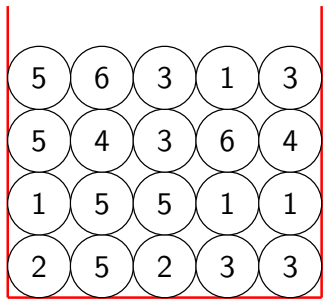
Document-Topic Urn (θ)

$z_{d,n}$ =

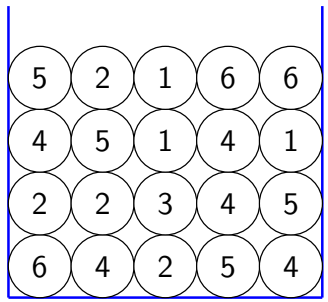
$w_{d,n}$ =



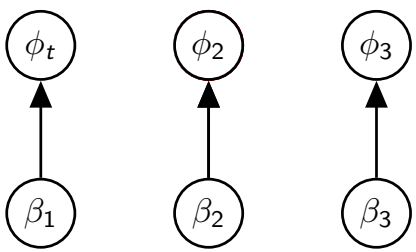
Topic Urn (ϕ): 1



Topic Urn (ϕ): 2

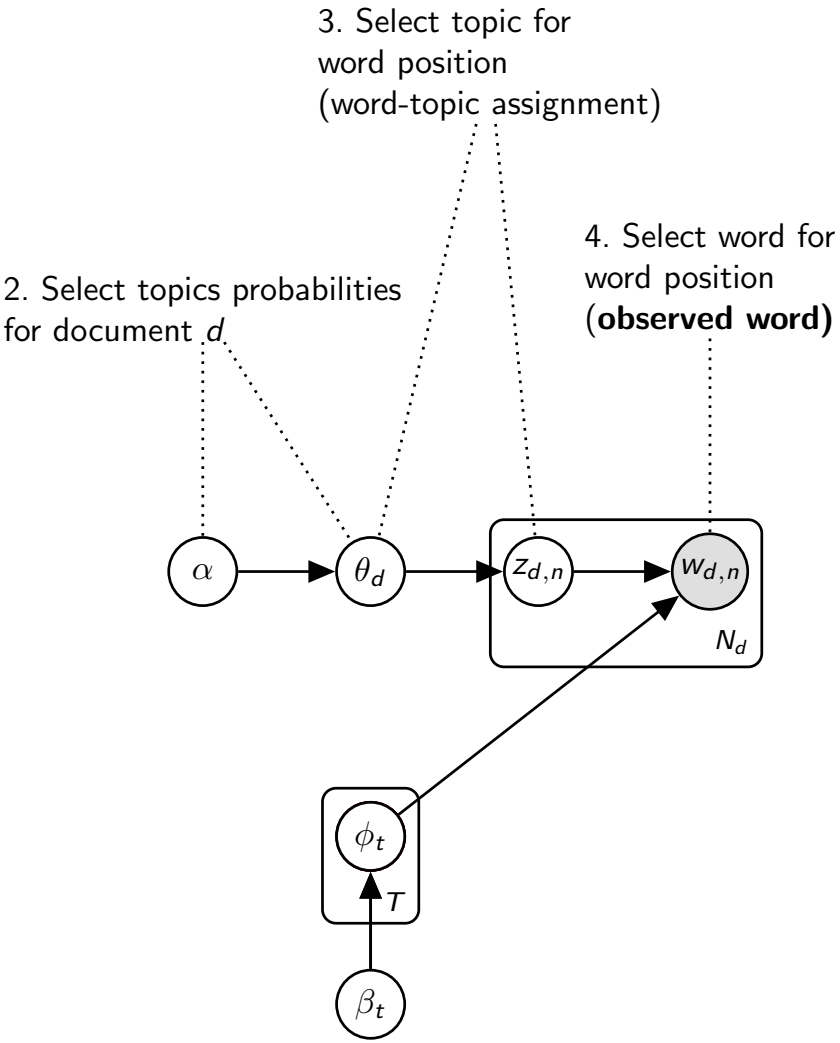


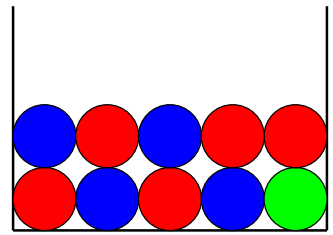
Topic Urn (ϕ): 3



1. Select word probabilities for each topic

Generative Process of LDA

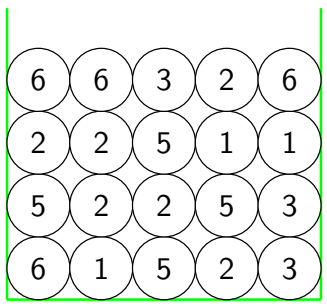




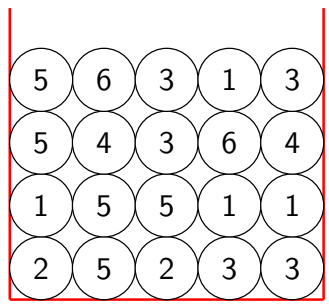
Document-Topic Urn (θ)

$z_{d,n}$ =

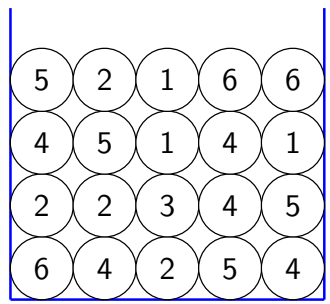
$w_{d,n}$ =



Topic Urn (ϕ): 1



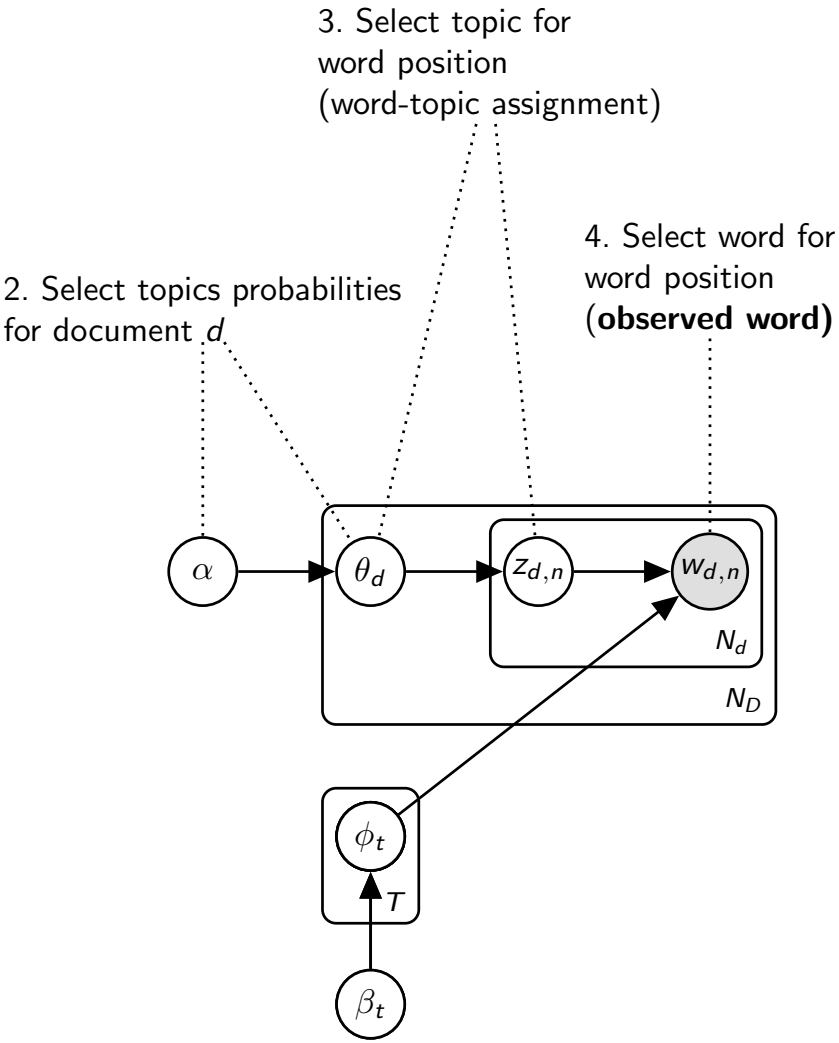
Topic Urn (ϕ): 2

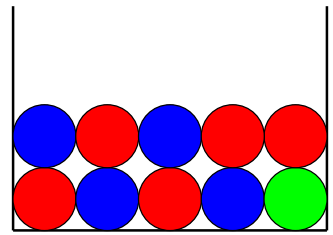


Topic Urn (ϕ): 3

1. Select word probabilities for each topic

Generative Process of LDA

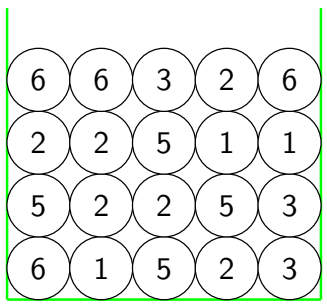




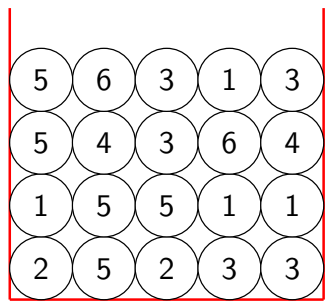
Document-Topic Urn (θ)

$z_{d,n}$ =

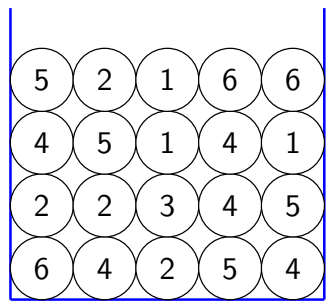
$w_{d,n}$ =



Topic Urn (ϕ): 1



Topic Urn (ϕ): 2

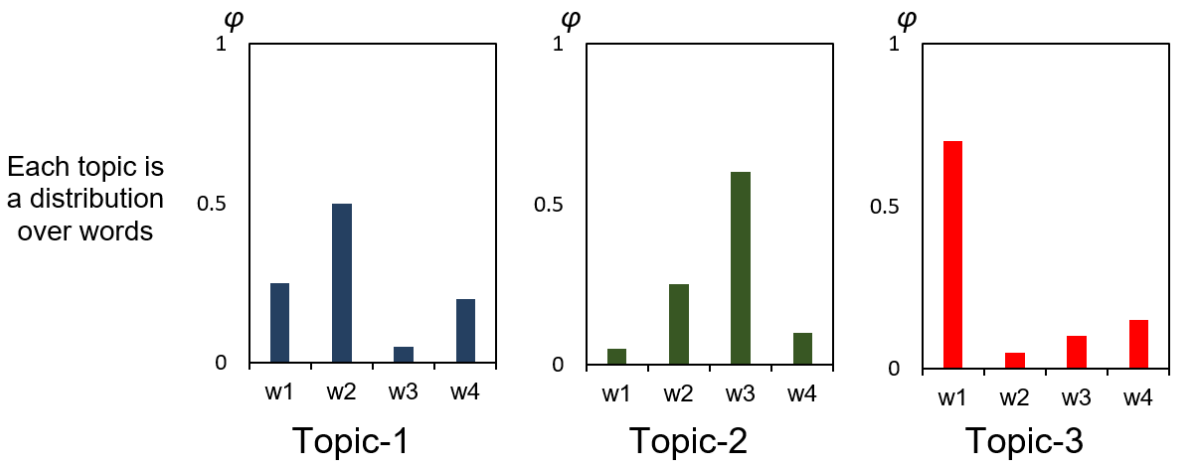


Topic Urn (ϕ): 3

1. Select word probabilities for each topic

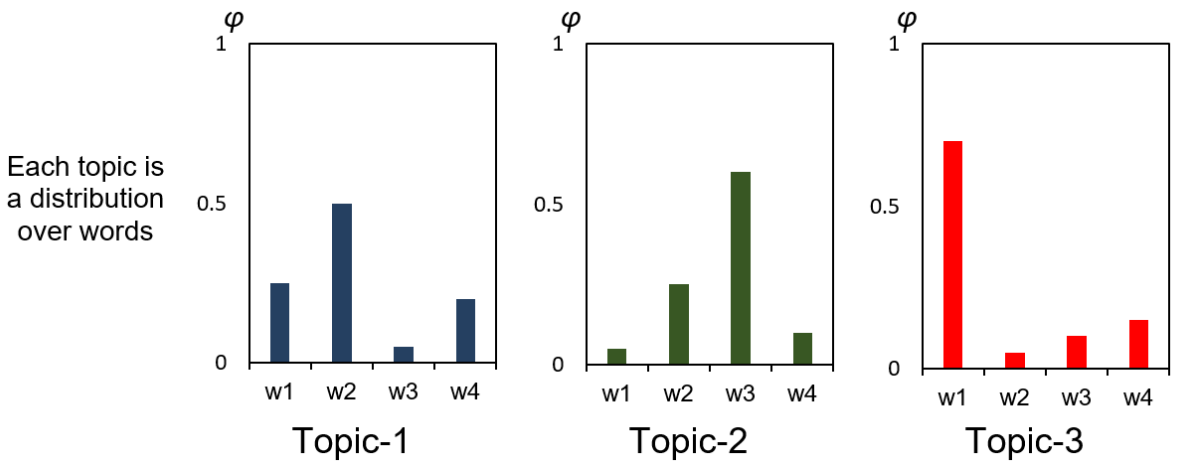
LDA : Key Assumptions

- Words that frequently co-occur with each other are related to the same subject.
  Call such clusters of co-occurring words "topics (or concepts)".

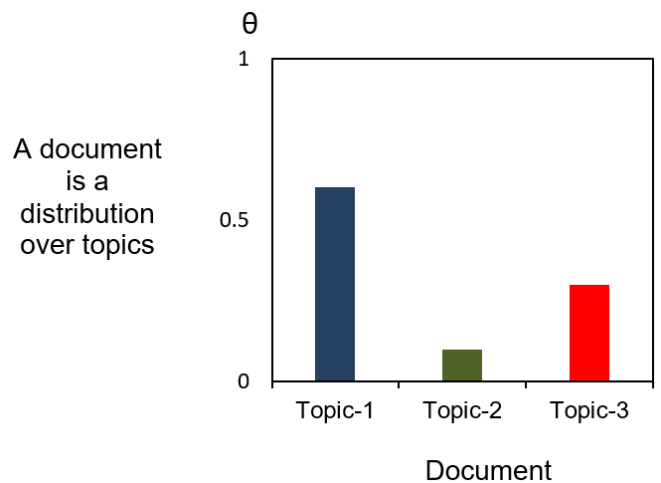


LDA : Key Assumptions

- Words that frequently co-occur with each other are related to the same subject.
  Call such clusters of co-occurring words "topics (or concepts)".



- Each document in the corpus exhibits the topics to varying degrees.



Overview of LDA

Overview of LDA

A set of topics on medical-space newsgroup corpus:

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
food	science	space	space
doctor	health	nasa	launch
day	research	earth	nasa
pain	information	orbit	technology
read	medical	mission	moon
disease	water	spacecraft	program
treatment	cancer	lunar	station
evidence	theory	solar	flight
blood	hiv	shuttle	commercial

Overview of LDA

A set of topics on medical-space newsgroup corpus:

Topic 1	Topic 2	Topic 3	Topic 4
food	science	space	space
doctor	health	nasa	launch
day	research	earth	nasa
pain	information	orbit	technology
read	medical	mission	moon
disease	water	spacecraft	program
treatment	cancer	lunar	station
evidence	theory	solar	flight
blood	hiv	shuttle	commercial

Overview of LDA

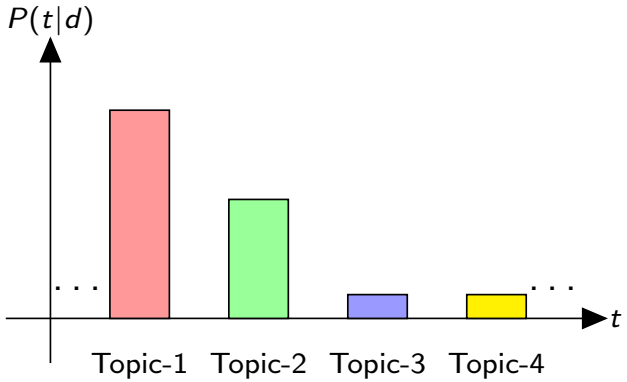
A set of topics on medical-space newsgroup corpus:

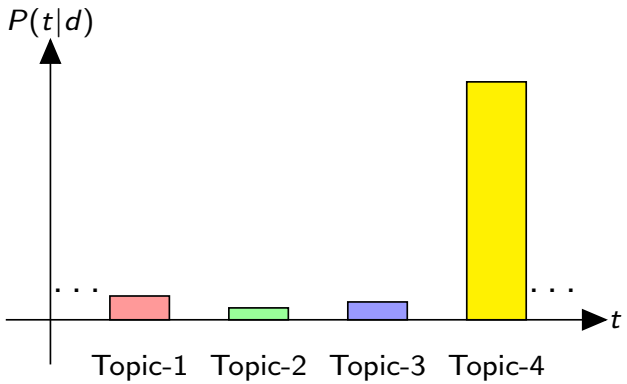
Topic 1	Topic 2	Topic 3	Topic 4
food	science	space	space
doctor	health	nasa	launch
day	research	earth	nasa
pain	information	orbit	technology
read	medical	mission	moon
disease	water	spacecraft	program
treatment	cancer	lunar	station
evidence	theory	solar	flight
blood	hiv	shuttle	commercial
medical	**medical**	**space**	**space**

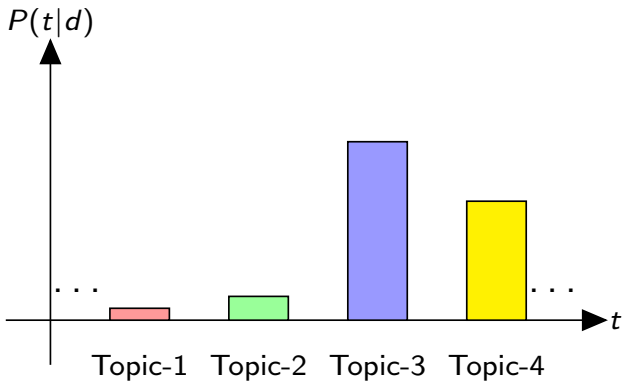
Overview of LDA

A set of topics on medical-space newsgroup corpus:

Topic 1	Topic 2	Topic 3	Topic 4
food	science	space	space
doctor	health	nasa	launch
day	research	earth	nasa
pain	information	orbit	technology
read	medical	mission	moon
disease	water	spacecraft	program
treatment	cancer	lunar	station
evidence	theory	solar	flight
blood	hiv	shuttle	commercial
medical	**medical**	**space**	**space**



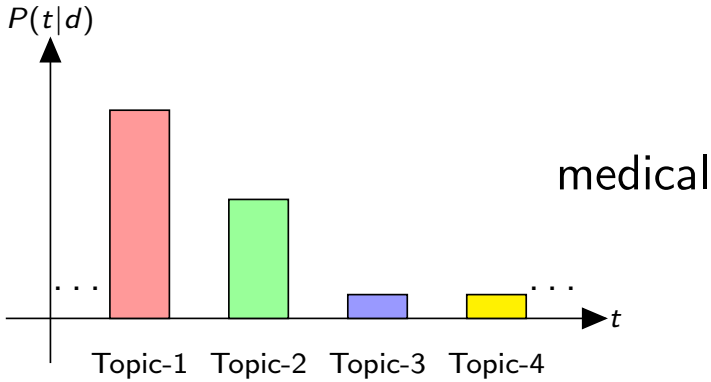


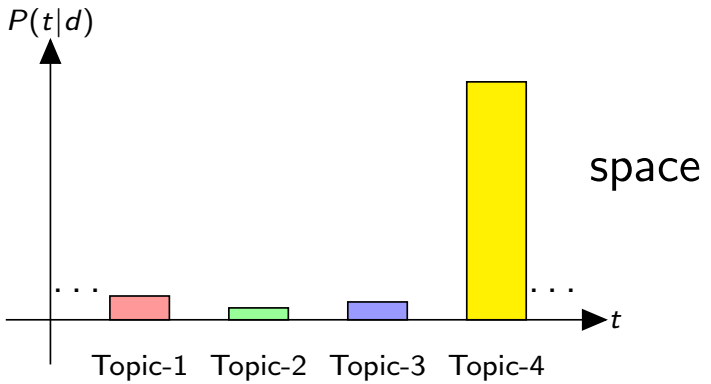


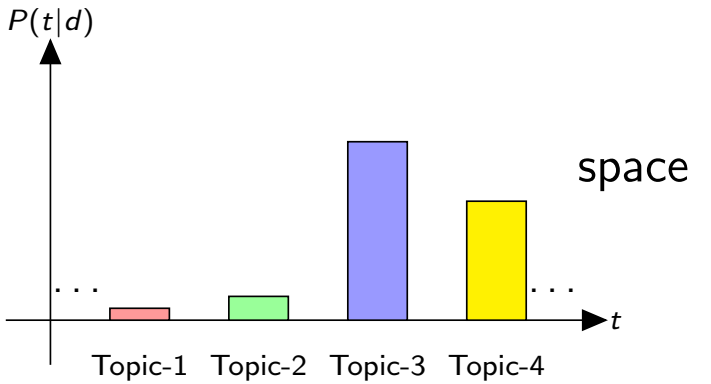
Overview of LDA

A set of topics on medical-space newsgroup corpus:

Topic 1	Topic 2	Topic 3	Topic 4
food	science	space	space
doctor	health	nasa	launch
day	research	earth	nasa
pain	information	orbit	technology
read	medical	mission	moon
disease	water	spacecraft	program
treatment	cancer	lunar	station
evidence	theory	solar	flight
blood	hiv	shuttle	commercial
medical	**medical**	**space**	**space**







Case Study: Community Earth System Model (CESM) discussion forum

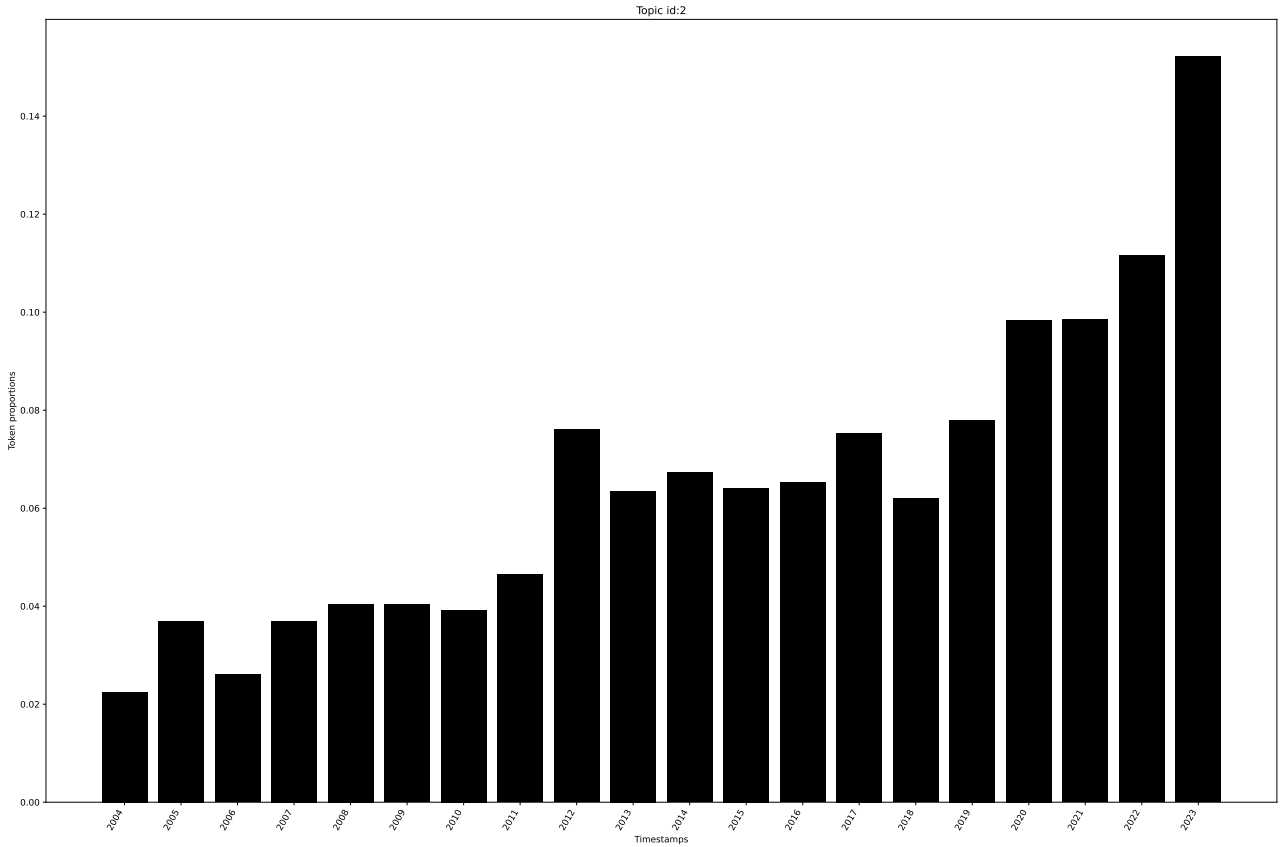
Community Earth System Model (CESM)

- Global climate model
- Facilitates simulations of Earth's climate states
- CESM Discussion Forums: Forums ($\sim$7000 posts since 2004)

Example LDA Topics

Id	Most probable words of a topic	Manually assigned label
0	component svn download line calling input clm compset case missing run directory check_input_data setting clean shr_strdata_print externals failed git	Version Management
1	setenv nthreads memory npes invalid netcdf init initialize join ids seq_comm_printcomms name echo comp argument explicit seq_comm_joincomm	Parallel computing
2	run clm files case forcing simulation compset year cmip restart xmlchange want set initial years output ssp change start land atmospheric said spinup b.e	CMIP and SSP
3	configure include gmake dlinux directory compiler pio checking mpif dfortranunderscore build test	Installing and setting up CESM
5	unknown reference undefined cesm.exe netcdf text line function lib indices ccsm.exe routine increasing...will source netcdf_mod_nf main forrtl image	Errors while running models
7	ocean ice grid sst pop cice sea files land output change values set domain forcing read som compset variables	Ocean modeling
12	code call variable subroutine module add end line variables procedure parameter number scam mean source history integer bug write	Source code related changes

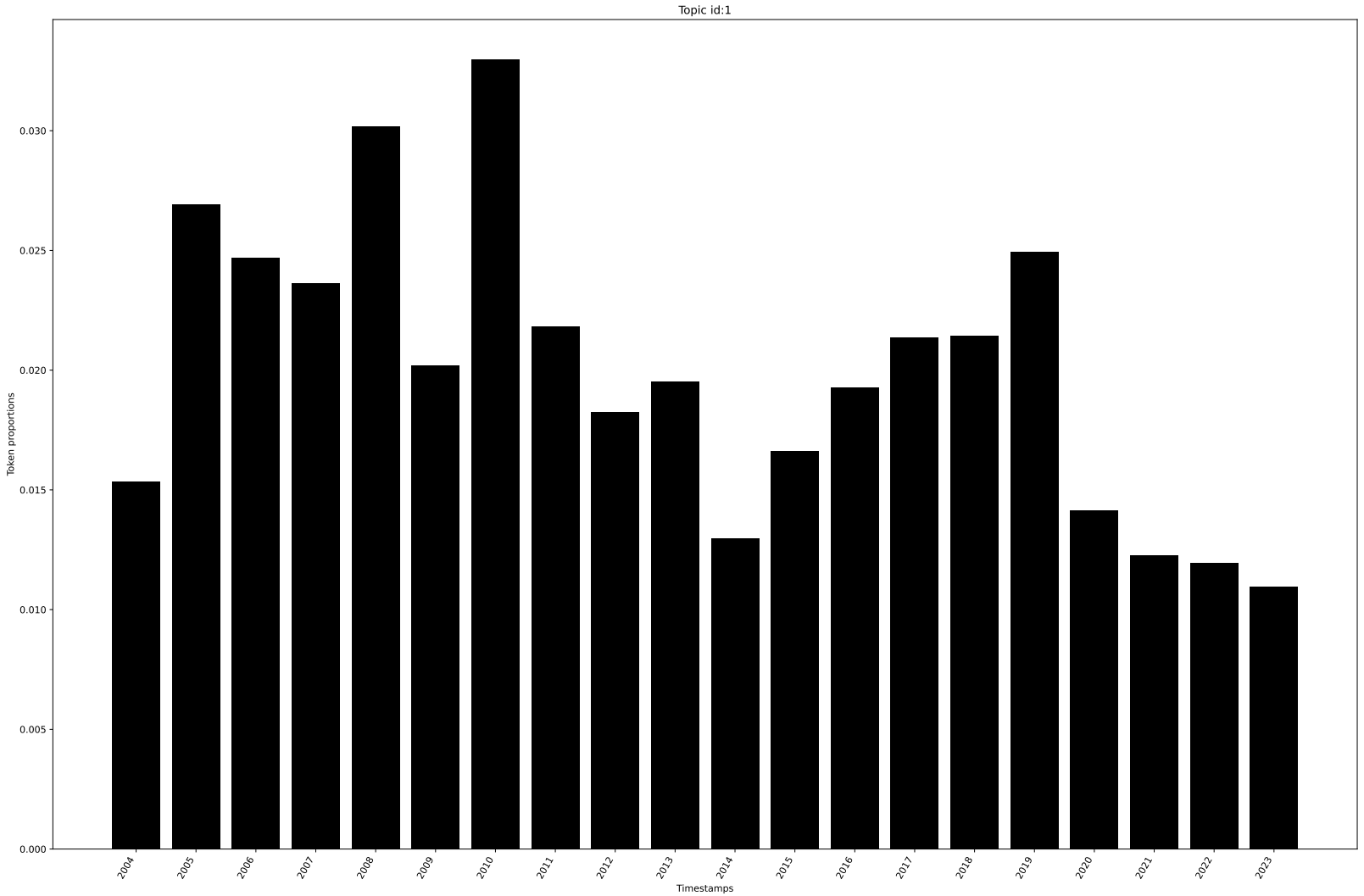
Topic 2 : CMIP and SSP



run clm files case forcing simulation compset year cmip[1] restart xmlchange want set initial years output ssp[2] change start land atmospheric said spinup b.e

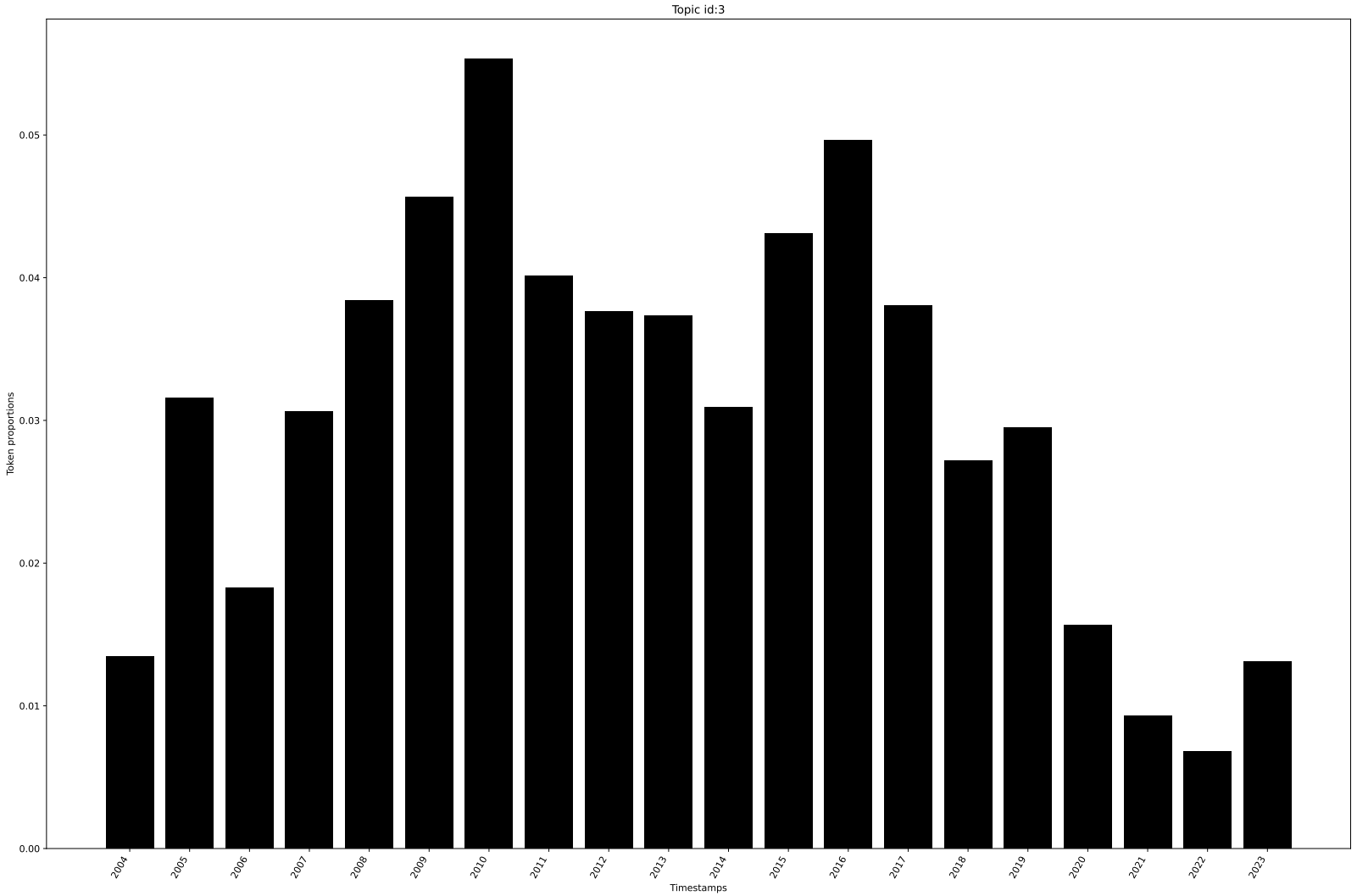
[1] https://en.wikipedia.org/wiki/Coupled_Model_Intercomparison_Project
[2] https://en.wikipedia.org/wiki/Shared_Socioeconomic_Pathways

Topic 1 : Parallel computing



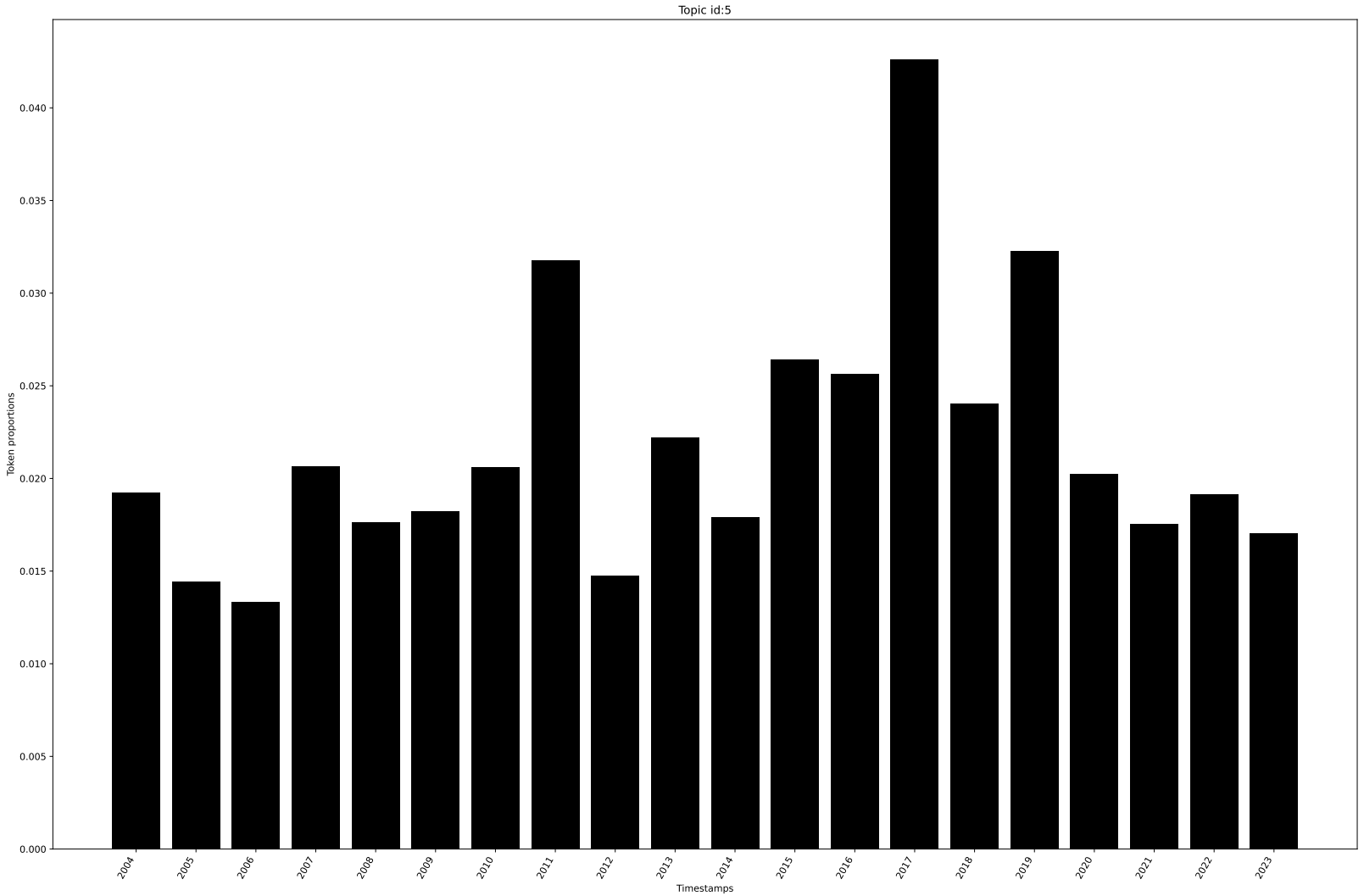
setenv nthreads memory npes invalid netcdf init initialize join ids seq_comm_printcomms name echo comp argument explicit seq_comm_joincomm

Topic 3 : Installing and setting up CESM



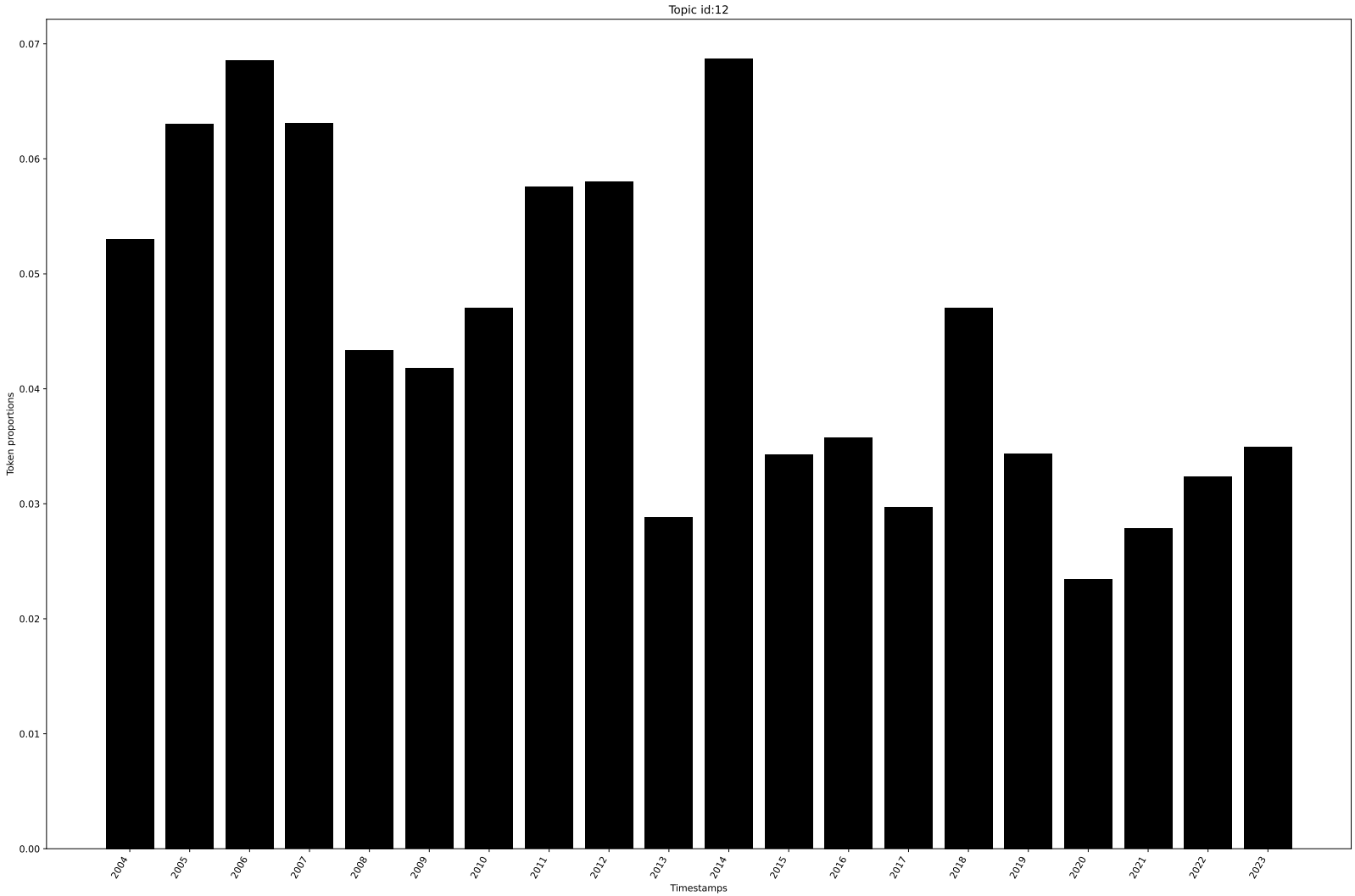
configure include gmake dlinux directory compiler pio checking mpif dfortranunderscore build test

Topic 5 : Errors while running models



unknown reference undefined cesm.exe netcdf text line function lib indices ccsm.exe routine increasing will source netcdf_mod_nf main forrtl image

Topic 12 : Source code related changes



component svn download line calling input clm compset case missing run directory check_input_data setting clean shr_strdata_print files externals failed server protocol git list cime build sandbox create_newcase description inputdata cam

Word Frequencies and Topics

Topic ID	Most probable words in the topic
0	and cam model discusscesm version waccm not data ccsm use available
2	data not forcing error model date atm and xmlchange clm files
8	clm and surface not land data model soil ctsm your pft
12	error and warning model process message not nan problem running called
13	ice ocean pop and grid cice sea not model forcing land
16	and data files restart initial sst model cam cmip compset
19	and not surface radiation temperature flux solar heat values variables model

- Topics are dominated by stop-words such as:
  model, *data*, *and*, *not*, *files*, etc.

Word Frequencies and Topics

- Due to a high frequency of stop-words, they are likely to dominate topics
- **it will also be true for domain-specific stop-words**
  (e.g. words like *model*, *data*, *files* in CESM corpus)
  should not be a stop-word?
- However, we expect topics which are **different** from each other and **specific** to a certain concept or theme
- Stop-word removal is an important text-processing step

Hands-on

- An implementation of LDA in Javascript + D3
  https://mimno.infosci.cornell.edu/jsLDA/index.html

- Thank you for your attention!