TOPIC MODELING WITH LDA

CSX4210/ INX4210

Natural Language Processing and Social Interaction

WHAT IS TOPIC MODELING?

A process to deduce the hidden topics / thematic structure from the document (or a collection to documents).

TOPIC MODELING TECHNIQUES

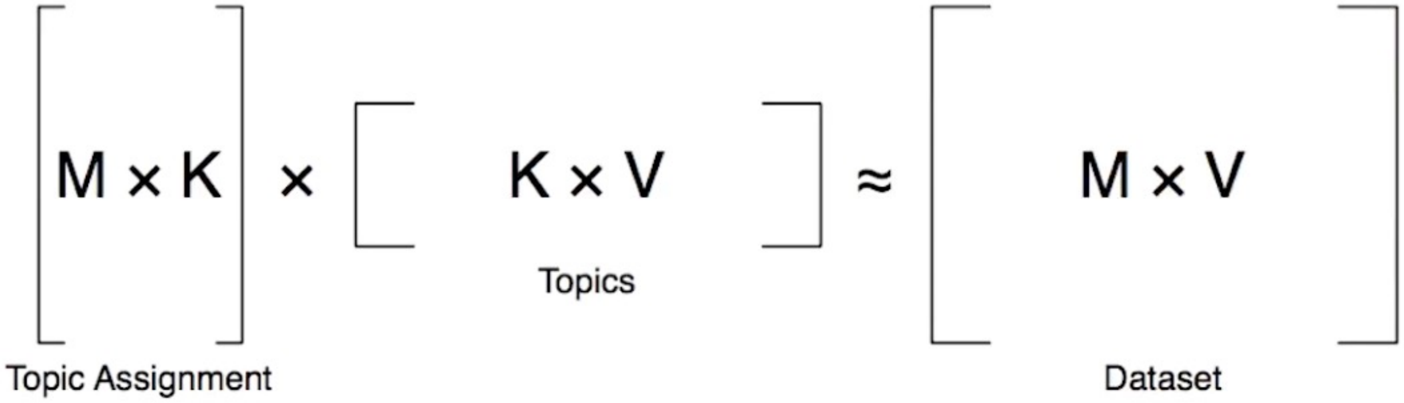
Non-Negative Matrix Factorization (NMF)

Latent Dirichlet Allocation (LDA)

Probabilistic Latent Semantic Indexing (pLSI)

Correlated Topic Model (CTM)

MATRIX FACTORIZATION



LATENT DIRICHLET ALLOCATION

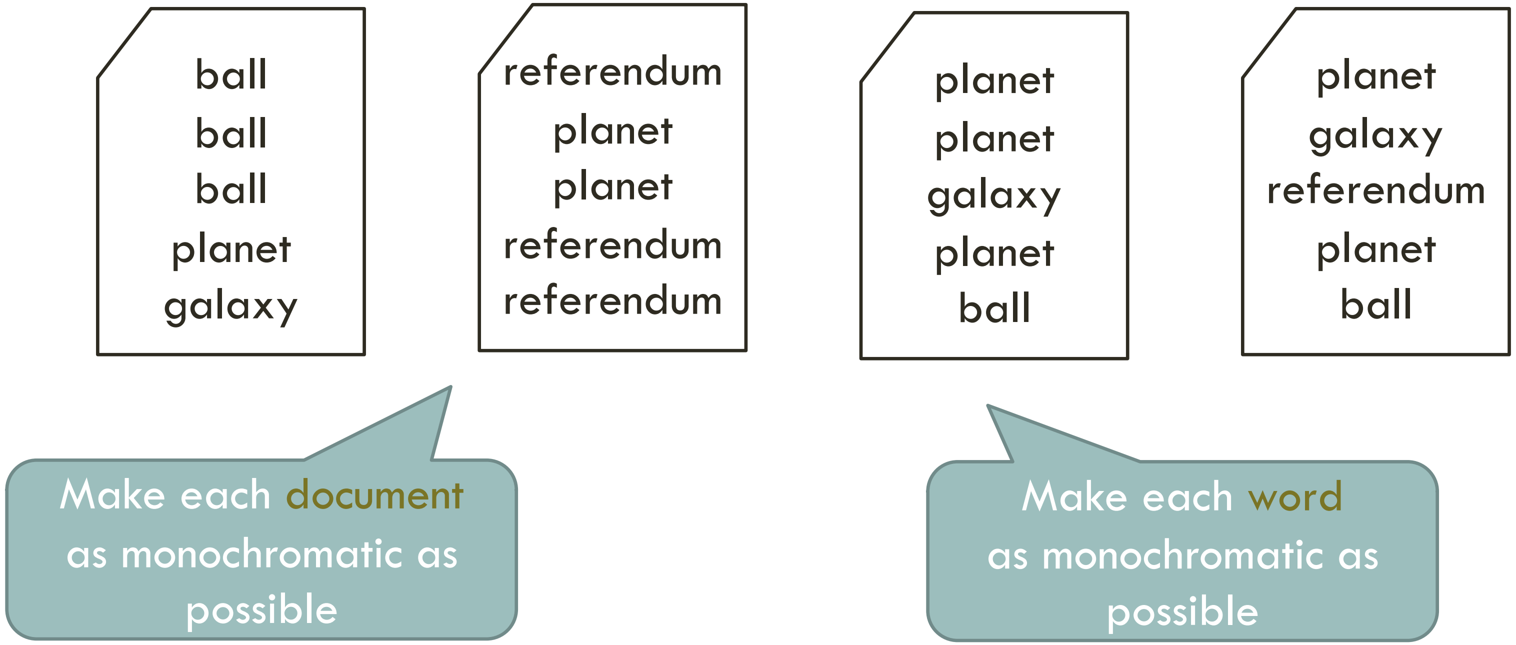
Probabilistic Topic Modeling

Unsupervised Learning

THE GOALS

Number of Components

Suppose the number of topics is 3, we want to color each word in one of the 3 colors i.e. Red, Green, and Blue.



ball
ball
ball
planet
galaxy

referendum
planet
planet
referendum
referendum

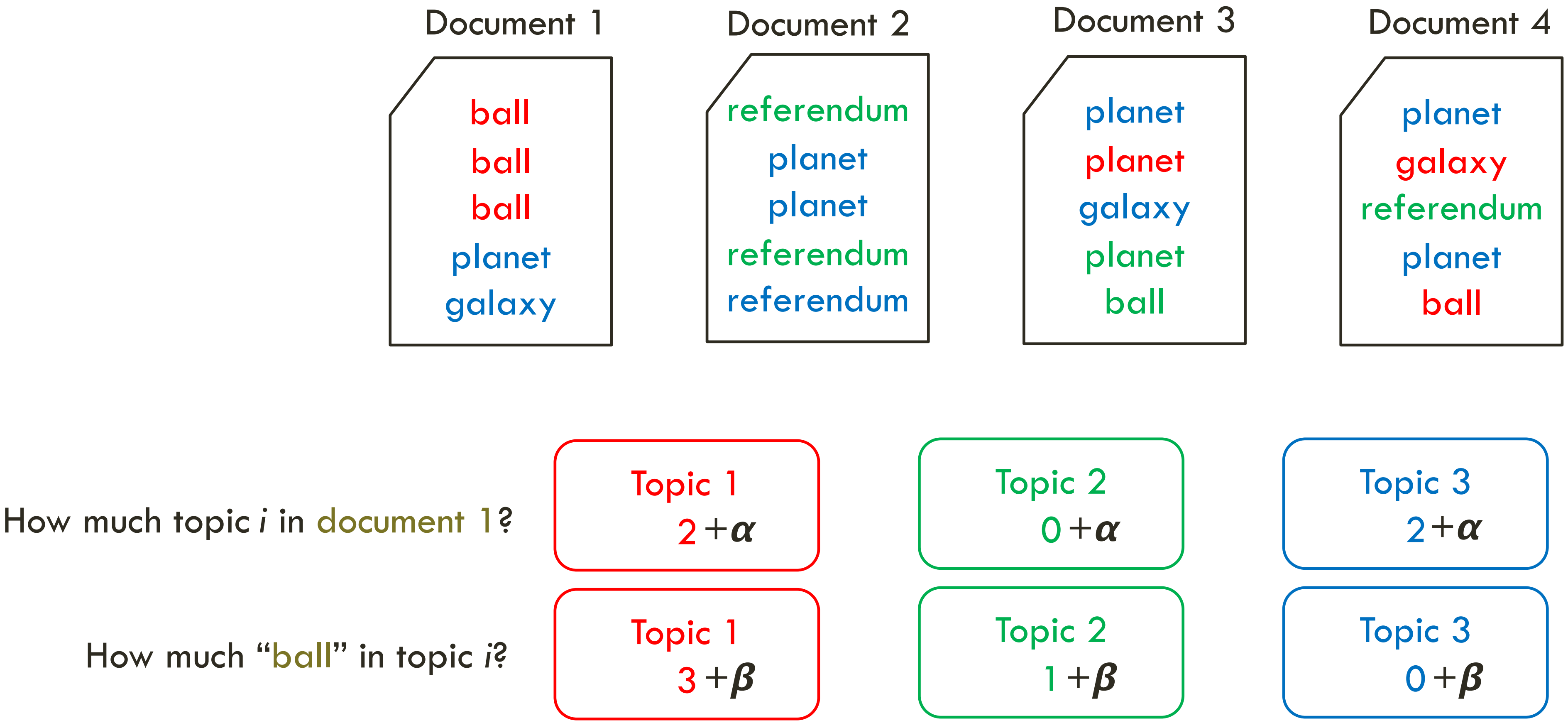
planet
planet
galaxy
planet
ball

planet
galaxy
referendum
planet
ball

Make each document as monochromatic as possible

Make each word as monochromatic as possible

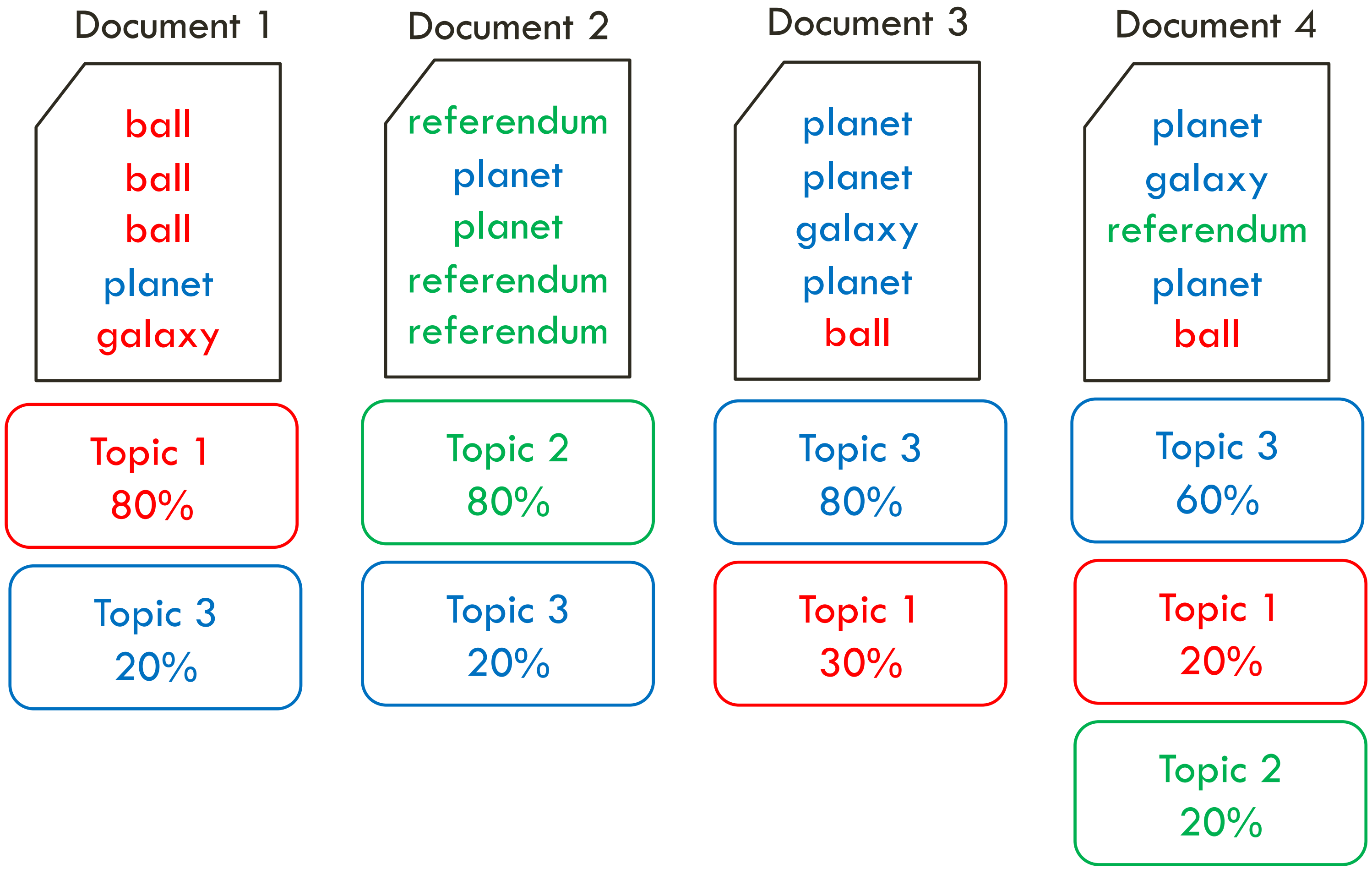
TOPIC-WORD ASSIGNMENT



Document 1	Document 2	Document 3	Document 4
ball	referendum	planet	planet
ball	planet	planet	galaxy
ball	planet	galaxy	referendum
planet	referendum	planet	planet
galaxy	referendum	ball	ball

How much topic *i* in document 1?	Topic 1 $2 + \boldsymbol{\alpha}$	Topic 2 $0 + \boldsymbol{\alpha}$	Topic 3 $2 + \boldsymbol{\alpha}$
How much "ball" in topic *i*?	Topic 1 $3 + \boldsymbol{\beta}$	Topic 2 $1 + \boldsymbol{\beta}$	Topic 3 $0 + \boldsymbol{\beta}$

DOCUMENT-TOPIC ASSIGNMENT



Document 1

ball
ball
ball
planet
galaxy

Topic 1
80%

Topic 3
20%

Document 2

referendum
planet
planet
referendum
referendum

Topic 2
80%

Topic 3
20%

Document 3

planet
planet
galaxy
planet
ball

Topic 3
80%

Topic 1
30%

Document 4

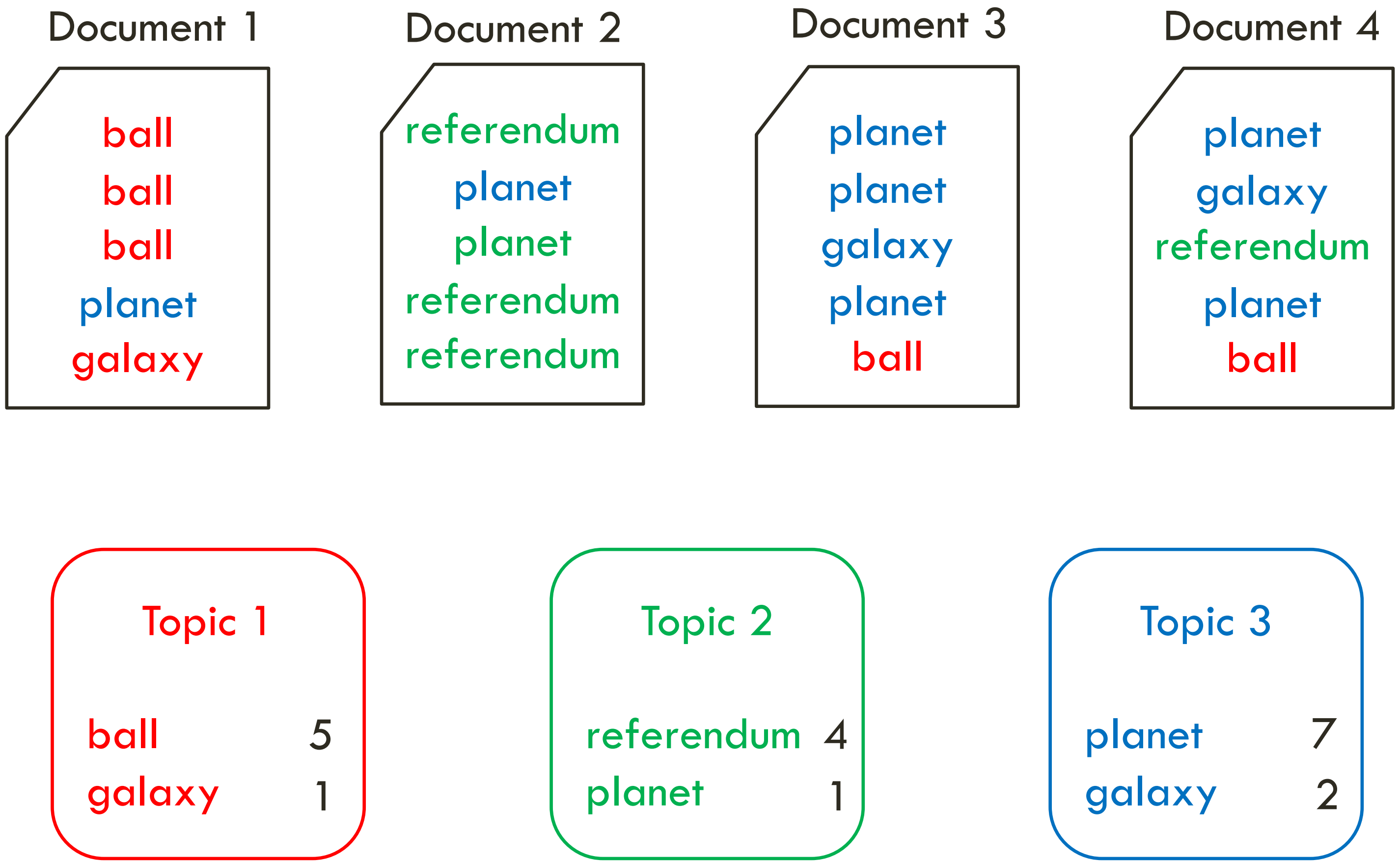
planet
galaxy
referendum
planet
ball

Topic 3
60%

Topic 1
20%

Topic 2
20%

TOPIC-WORD ASSIGNMENT



Document 1

ball
ball
ball
planet
galaxy

Document 2

referendum
planet
planet
referendum
referendum

Document 3

planet
planet
galaxy
planet
ball

Document 4

planet
galaxy
referendum
planet
ball

Topic 1

ball	5
galaxy	1

Topic 2

referendum	4
planet	1

Topic 3

planet	7
galaxy	2

TECHNICAL SIDE

KEY CONCEPTS

A distribution of distribution

Documents-to-Topics

Topics-to-Words

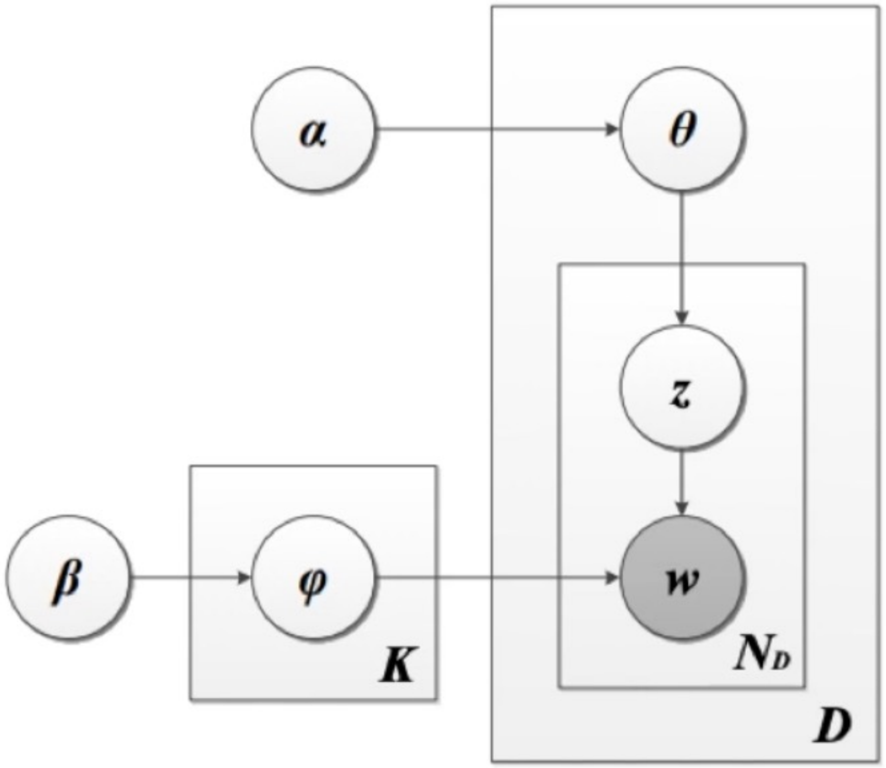
PROBABILITY OF A DOCUMENT

Dirichlet Distribution (triangle)

- Alpha = 1, uniform
- Alpha < 1, towards corner
- Alpha > 1, towards center

Multinomial Distribution

LDA BLUEPRINT



Alpha: Document-to-Topics

Beta: Topic-to-Words

Theta: Picking topics

Phi: Picking words

Z: list of topics

W: list of words

TOPIC MODEL VISUALIZATION

pyLDAvis

BERTopic

TopicWizard

Termite Plot

EVALUATION

Log Likelihood

- Held-out data

Perplexity

Interpretability

- Rely on human
- Model precision
  - Word Intrusion: Find the words that don’t belong to the topics.
  - Topic Intrusion:: Topic log odds (TLO)
  - Topic Coherence