دانشگاه بوعلی سینا
Bu-Ali Sina University

Topic Modeling

Amin Nazari

Clustering

Problem description

- Given:

  A data set of N data items which are d-dimensional data feature vectors.

- Task:

  Determine a natural, useful partitioning of the data set into a number of clusters (k) and noise.

Major Types of Clustering Algorithms

- Partitioning:

  Partition the database into k clusters which are represented by representative objects of them

- Hierarchical:

  Decompose the database into several levels of partitioning which are represented by dendrogram

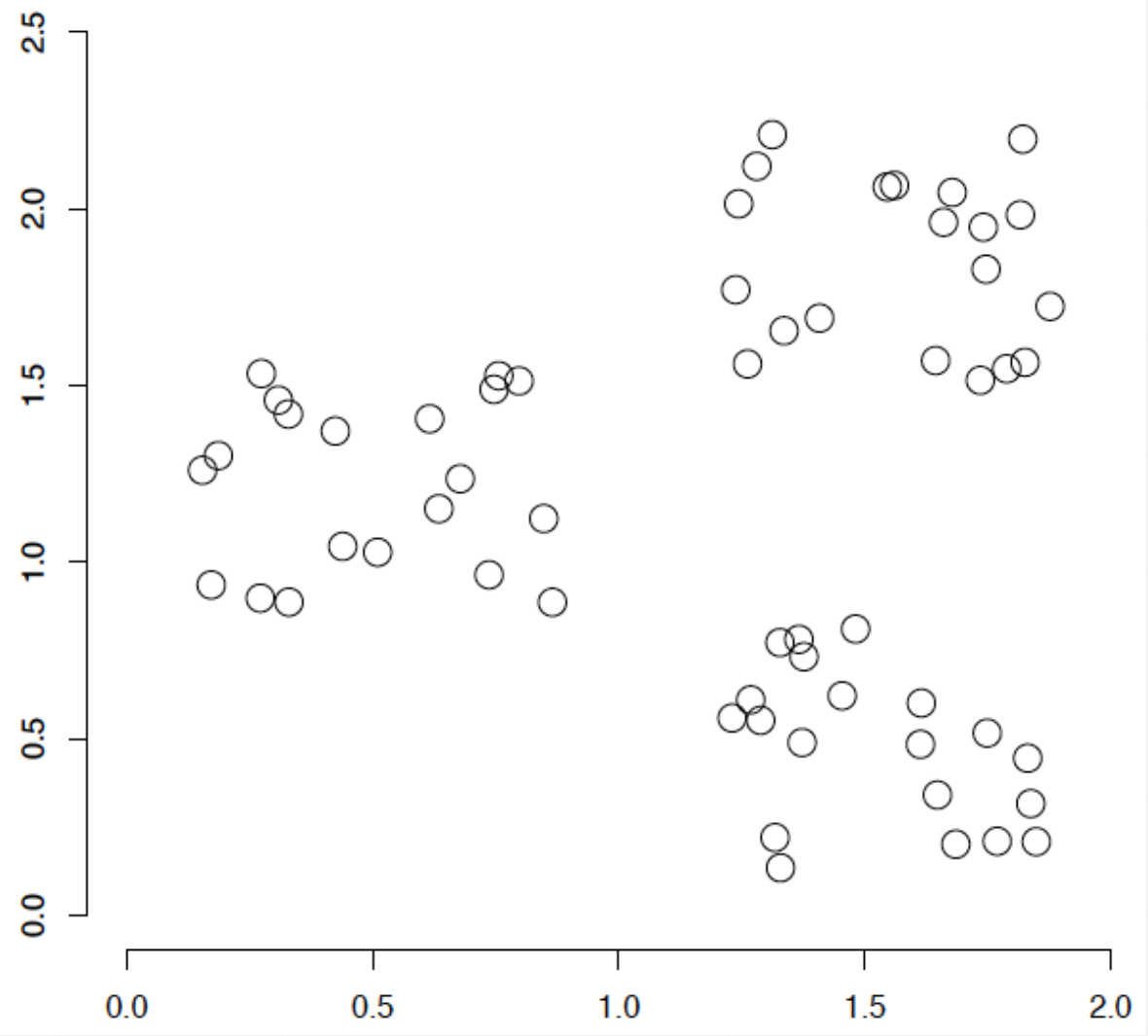
Other kinds of Clustering Algorithms

- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

What is K-Means Clustering?

- An unsupervised learning algorithm
- Groups data into K distinct clusters
- Each cluster is represented by the centroid
- Goal: Minimize intra-cluster variance

A data set with clear cluster structure



- How would you design an algorithm for finding the three clusters in this case?

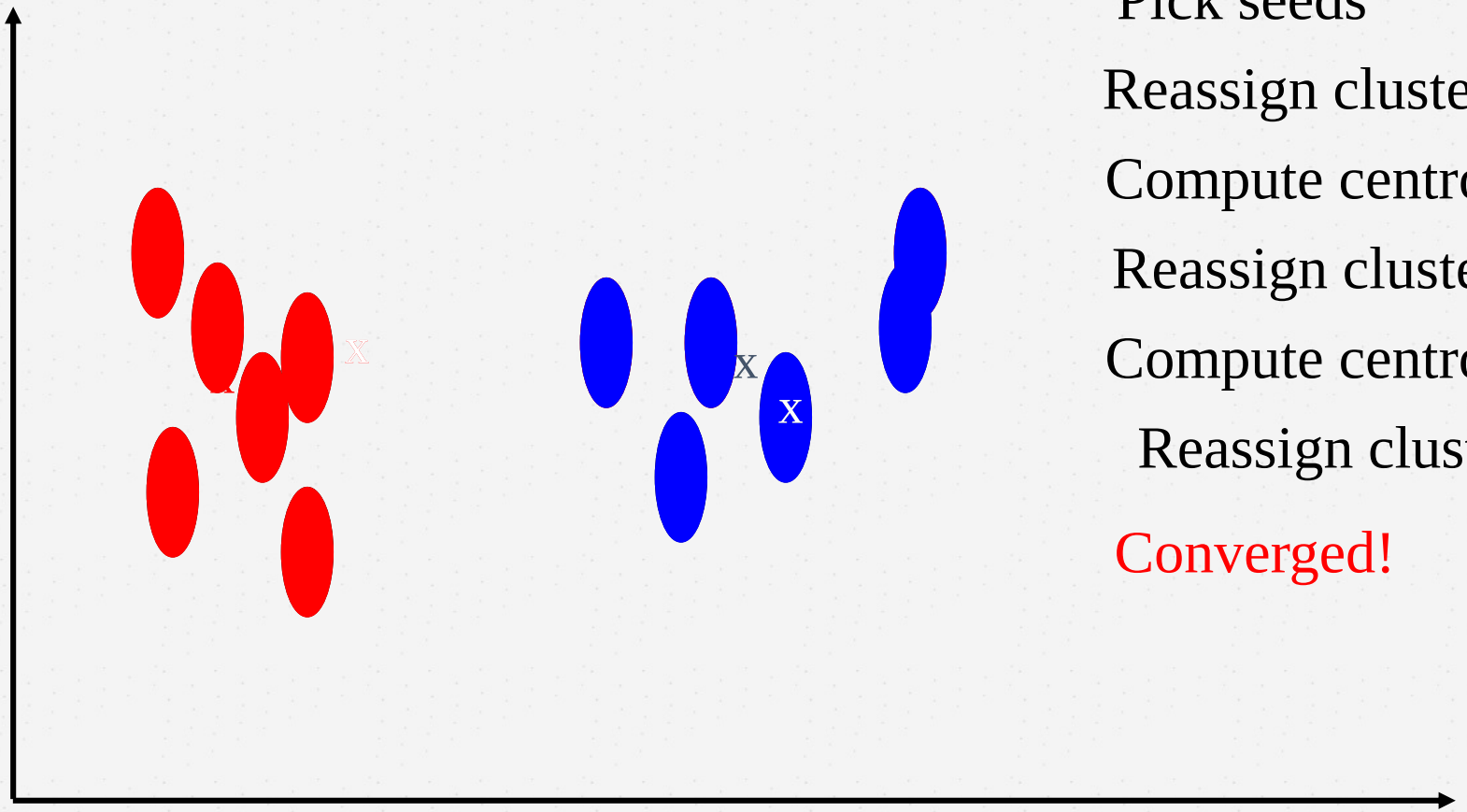
Intuition Behind K-Means

- Similar data points are grouped together
- Each point belongs to the cluster with the nearest centroid
- Centroids are updated iteratively to optimize clustering

K-Means Algorithm Steps

1. Initialize K centroids randomly
2. Assign each data point to the nearest centroid
3. Update each centroid as the mean of the points assigned to it
4. Repeat steps 2–3 until convergence

K Means Example
(*K*=2)



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

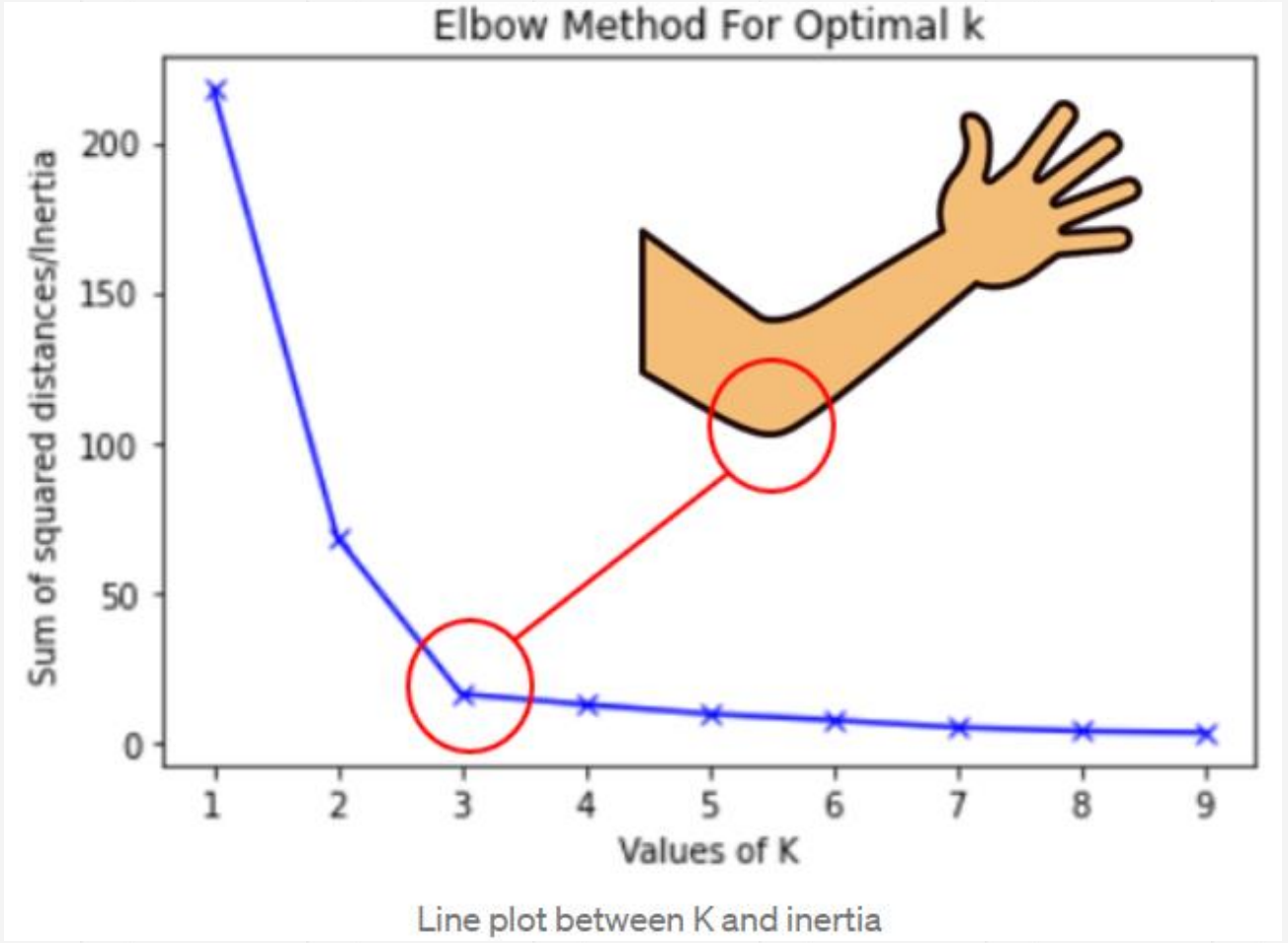
Compute centroids

Reassign clusters

Converged!

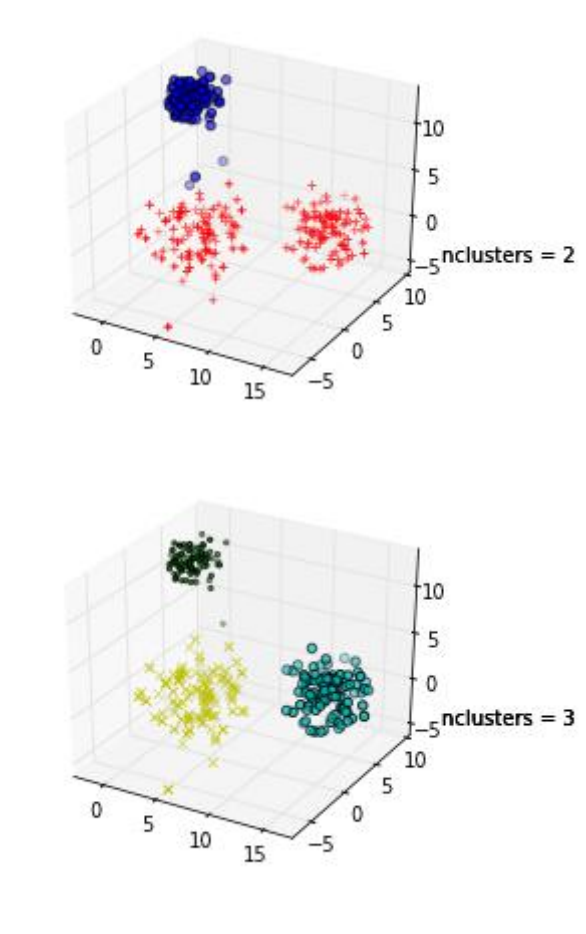
Choosing the Right K

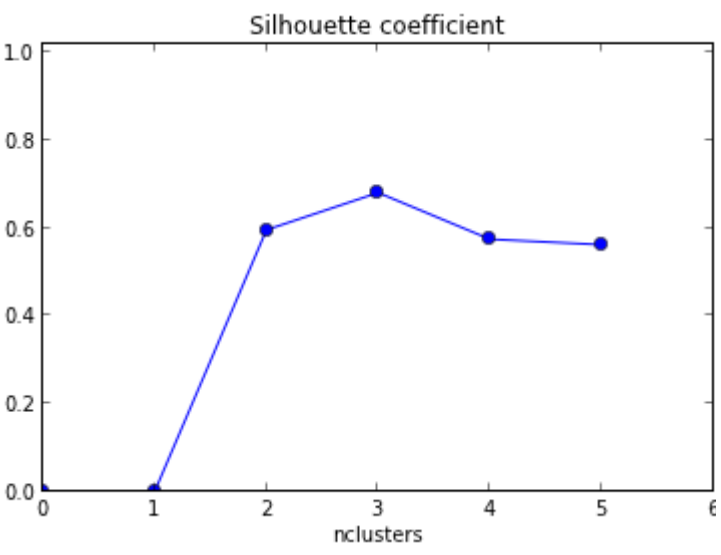
- Elbow Method: Plot SSE vs. K
- Silhouette Score: Similarity within clusters vs. others
- Use domain knowledge or business context

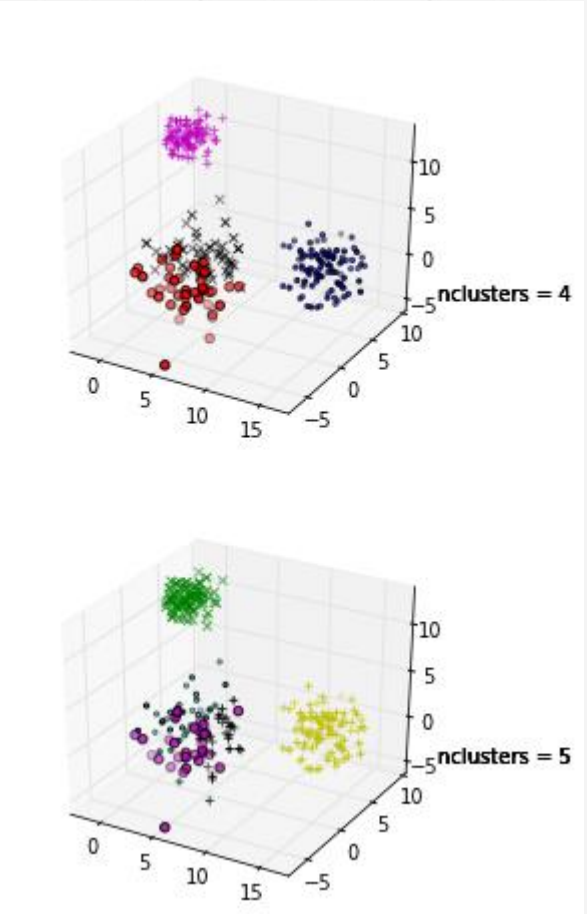


Line plot between K and inertia

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$







Pros and Cons

Pros:

- Simple and fast
- Scales to large datasets
- Works well when clusters are spherical and balanced

Cons

- Need to specify K
- Sensitive to initial centroids
- Struggles with non-convex shapes

What is Topic Modeling?

- Topic modeling is a machine learning technique that automatically analyzes text data to determine cluster words for a set of documents.
- Since topic modeling doesn’t require training, it’s a quick and easy way to start analyzing your data.

Topic modeling vs. topic classification

- topic modeling algorithms churn out collections of expressions and words that it thinks are related, leaving you to figure out what these relations mean, while topic classification delivers neatly packaged topics, with labels such as Price, and Features, eliminating any guesswork.

Topic modeling Algorithms

- Latent Semantic Indexing(1990)
- Probabilistic Latent Semantic Indexing(1999)
- Latent Semantic Analysis (LSA)(2005)
- Latent Dirichlet Allocation (LDA)(2003)

What is Topic Modeling?

- - Unsupervised learning to discover hidden themes in text
- - Each document = mixture of topics
- - Each topic = distribution over words

Matrix Factorization Overview

- - Decomposes a matrix into smaller matrices
- - In topic modeling:
- Document-Term Matrix ≈ Document-Topic × Topic-Term

Document-Term Matrix (DTM)

- - Rows: Documents
- - Columns: Terms (words)
- - Values: Frequency or TF-IDF
- - Typically sparse

Applying NMF to Topic Modeling

- - NMF: Non-negative Matrix Factorization
- - $V \approx W \times H$
- • W: Document-topic matrix (Random)
- • H: Topic-term matrix(Random)

$$H \leftarrow H \times \frac{W^\top V}{W^\top W H}$$

$$W \leftarrow W \times \frac{V H^\top}{W H H^\top}$$

- - Non-negativity improves interpretability

Interpreting the Results

- - W: topic distribution for documents
- - H: word distribution for topics
- - Enables:
  - Topic labeling
  - Document classification
  - Topic visualization

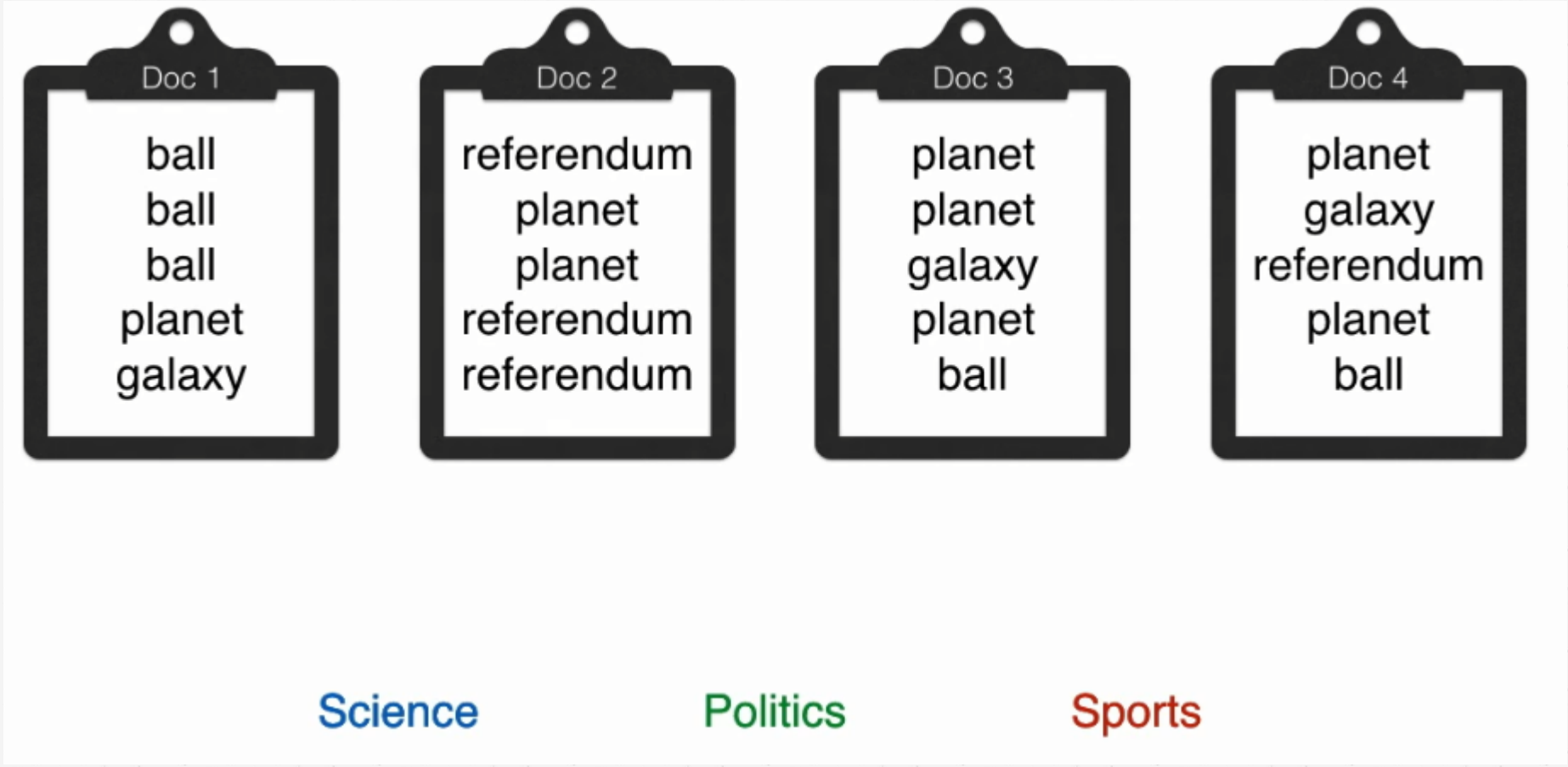
Example

- - 5 documents × 10 terms DTM
- - NMF discovers 2 topics
- - Each topic represented by top terms

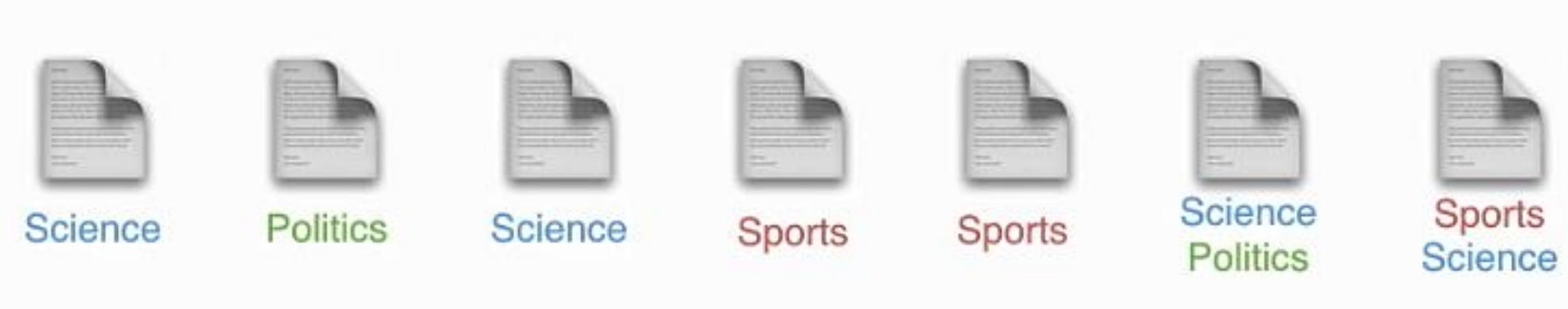
Preprocessing Steps

- - Tokenization
- - Lowercasing, stop word removal
- - Stemming or lemmatization
- - Convert to TF or TF-IDF matrix

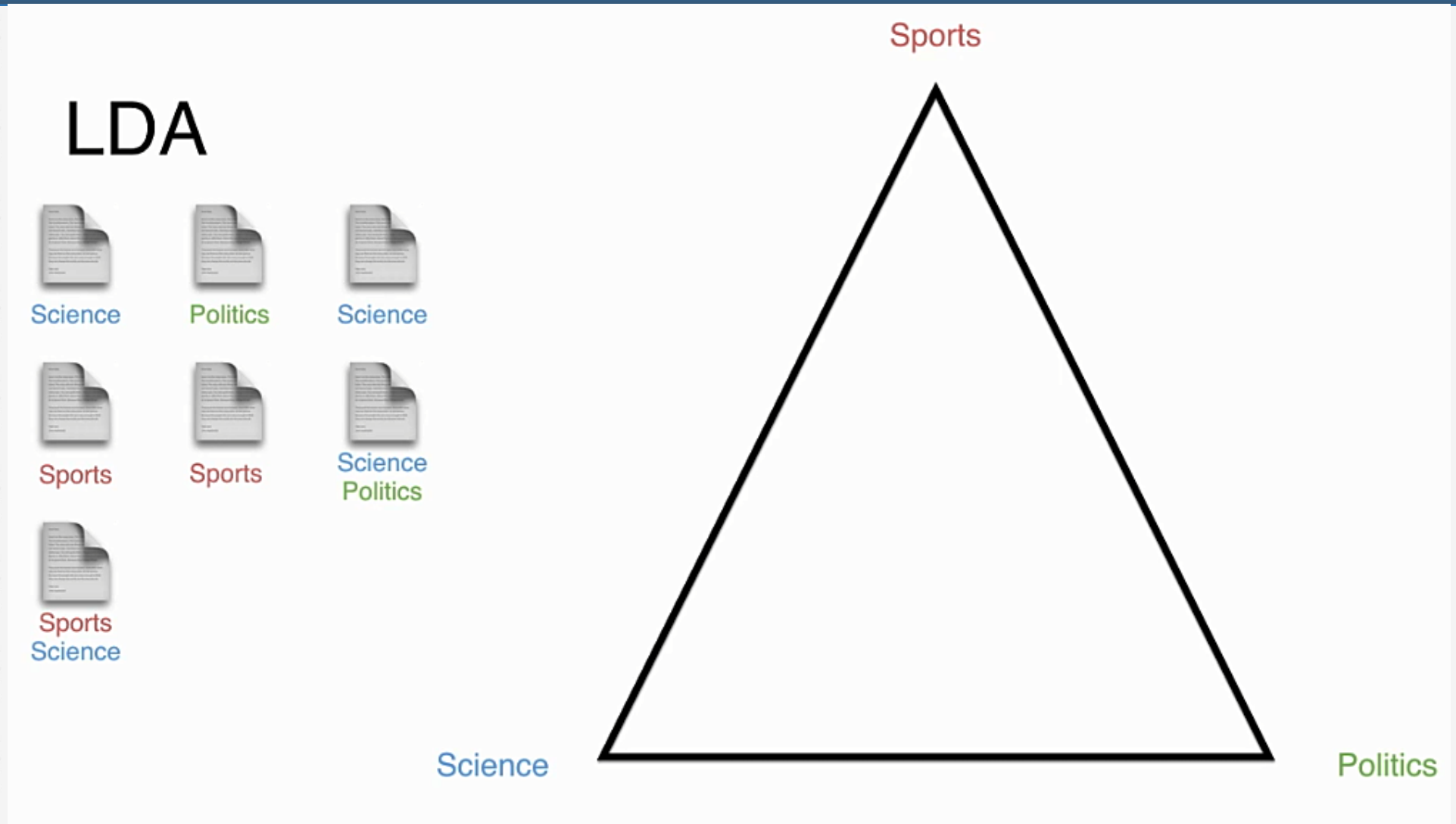
Topic modeling Algorithms



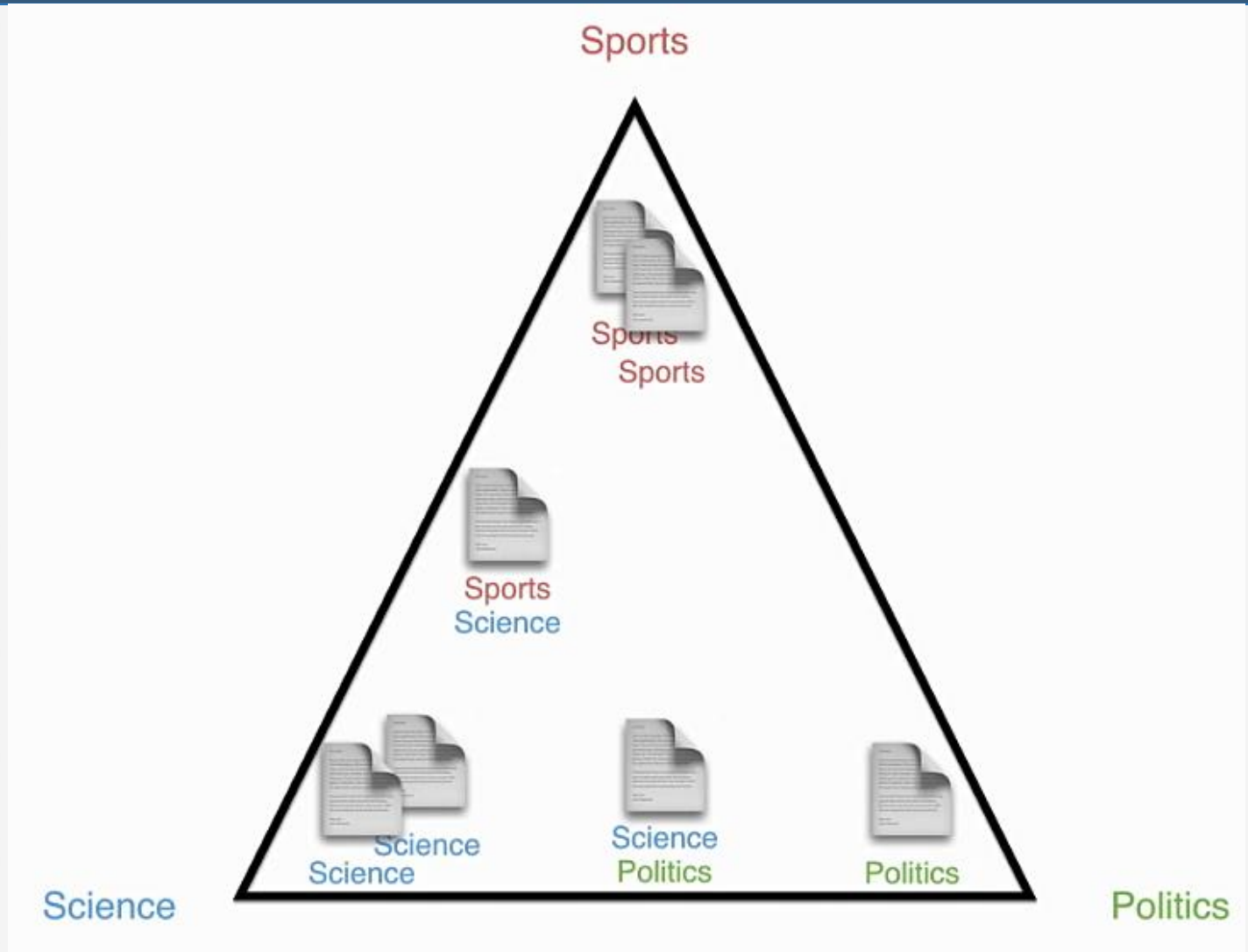
Topic modeling Algorithms



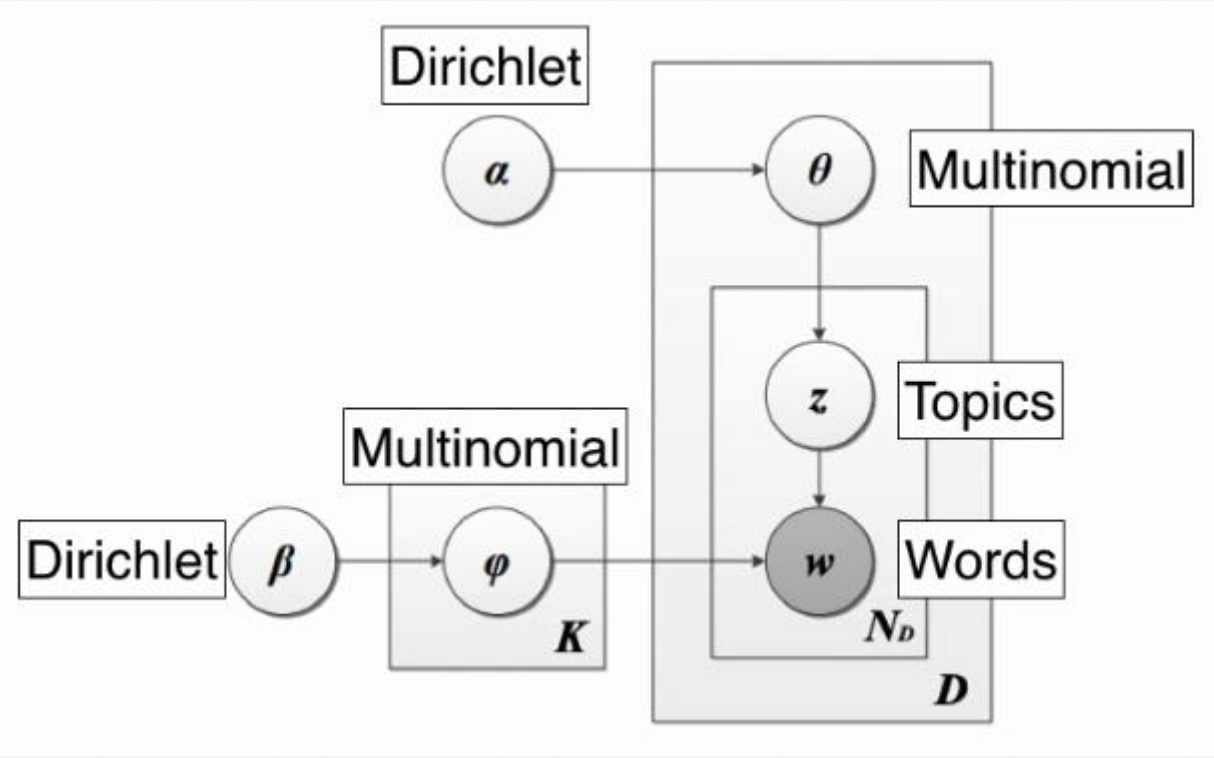
Topic modeling Algorithms



Topic modeling Algorithms



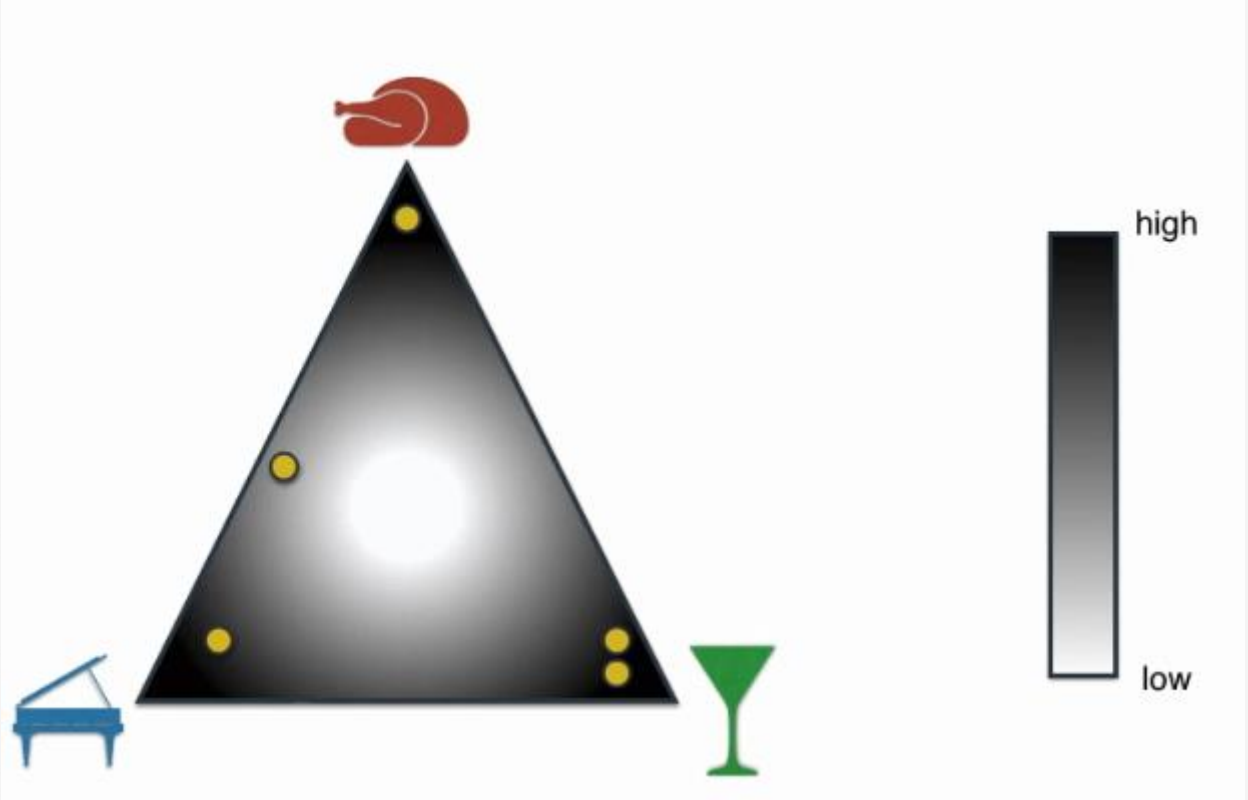
Topic modeling Algorithms



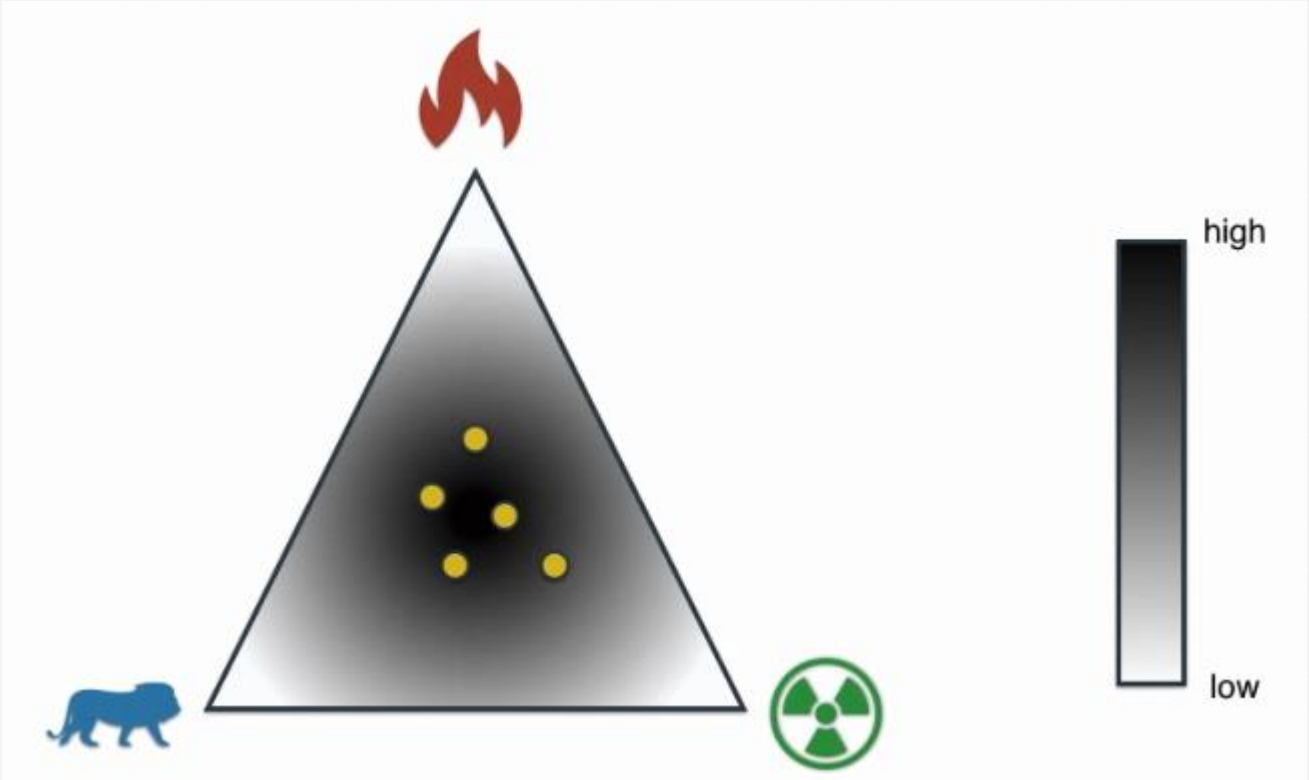
Topic modeling Algorithms

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

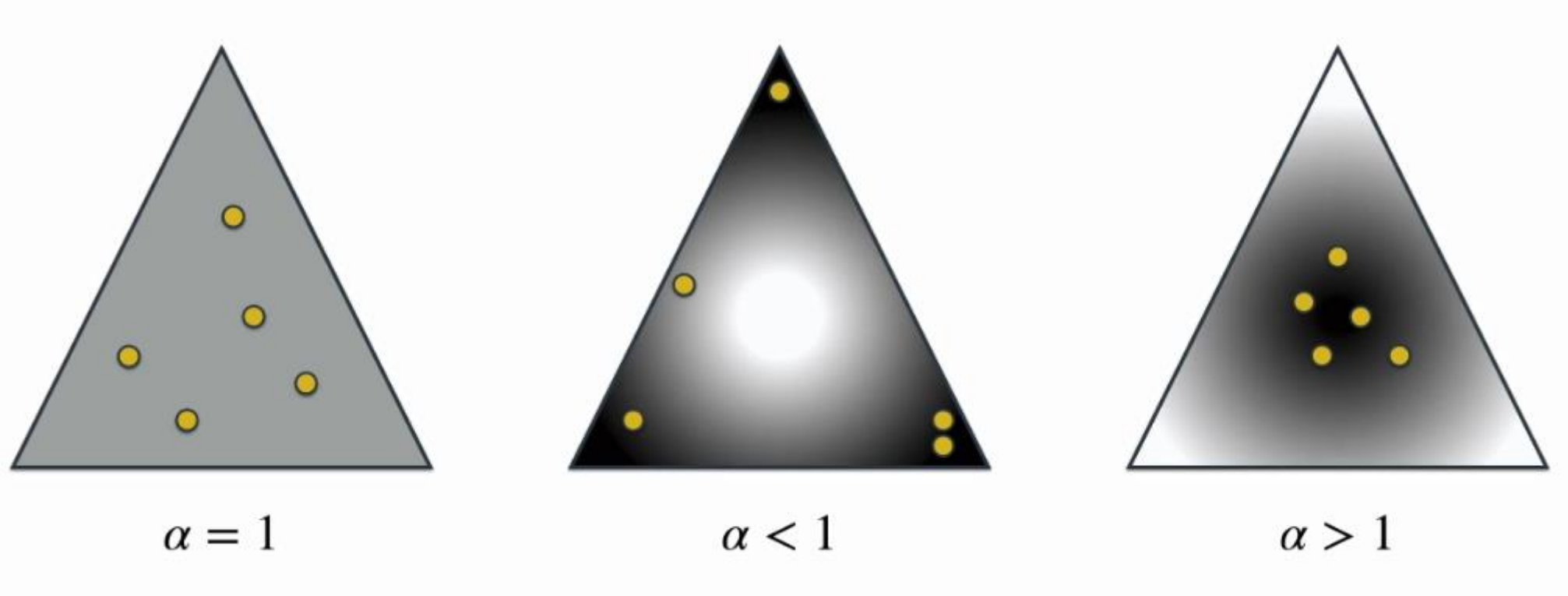
Topic modeling Algorithms

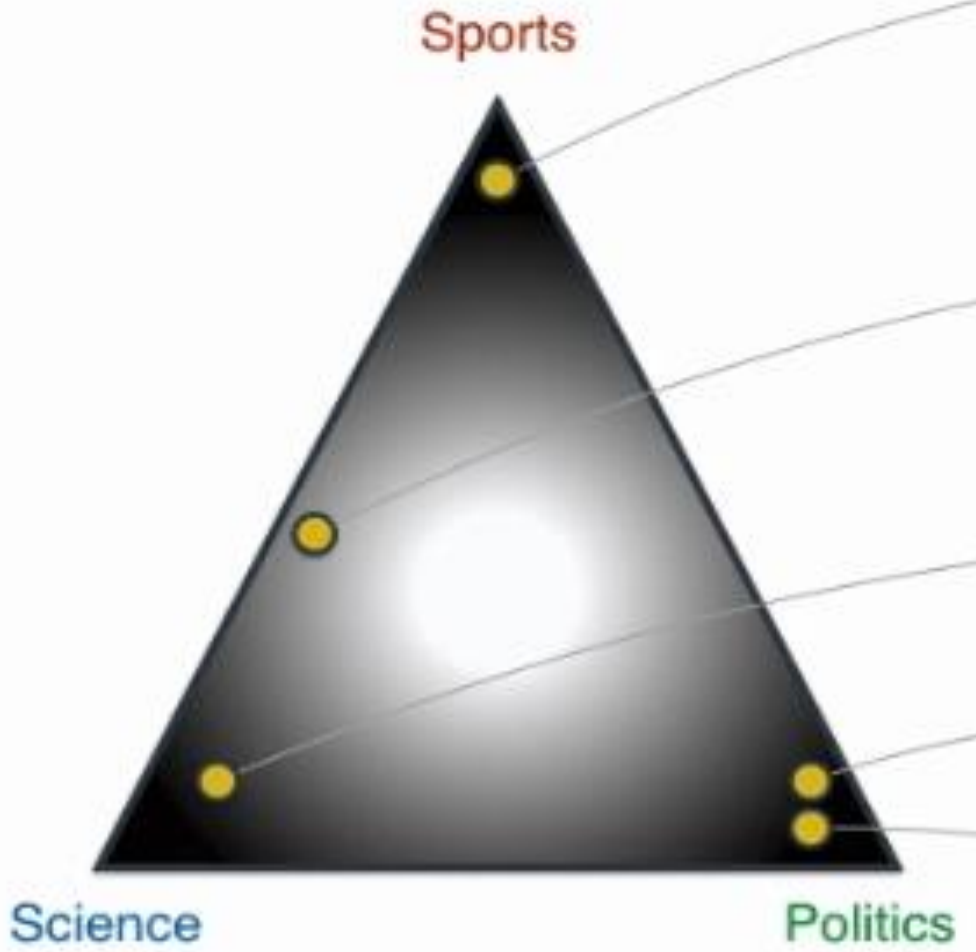


Topic modeling Algorithms



Topic modeling Algorithms





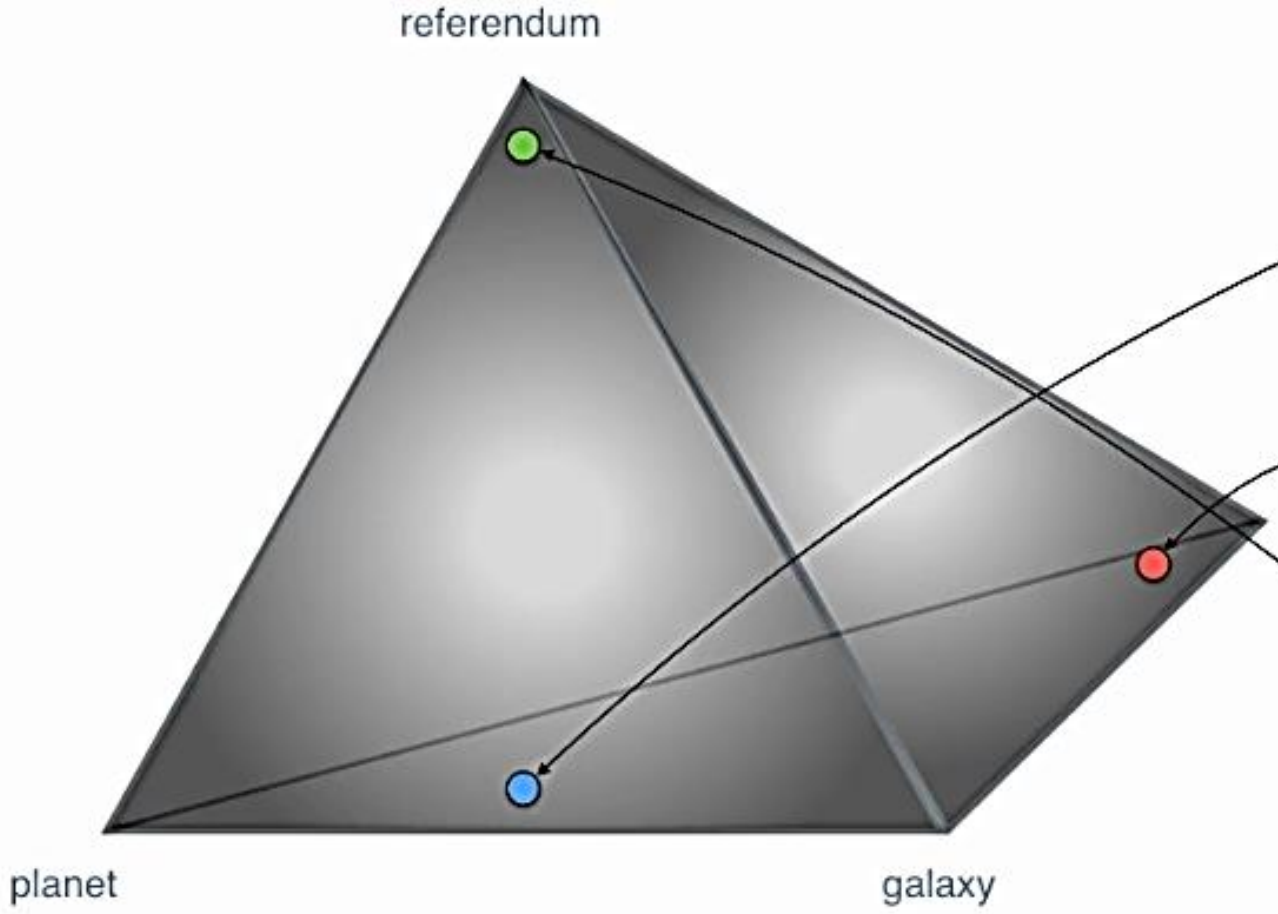
science	politics	sports
0.07	0.03	0.9

science	politics	sports
0.45	0.05	0.5

science	politics	sports
0.8	0.1	0.1

science	politics	sports
0.05	0.8	0.15

science	politics	sports
0.05	0.9	0.05



science

Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

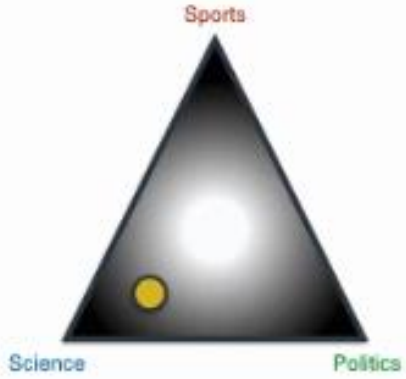
sports

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

politics

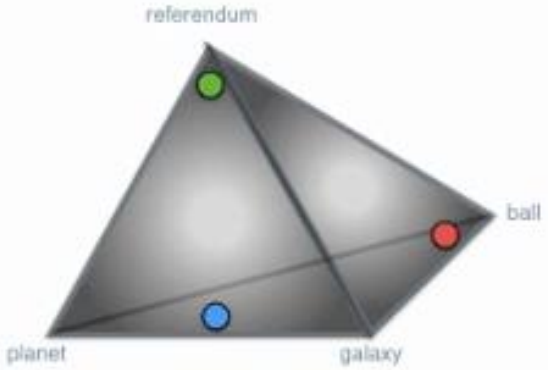
Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

$$\prod_{j=1}^{M} P(\theta_j; \alpha)$$



science	politics	sports
0.7	0.1	0.2

$$\prod_{i=1}^{K} P(\varphi_i; \beta)$$

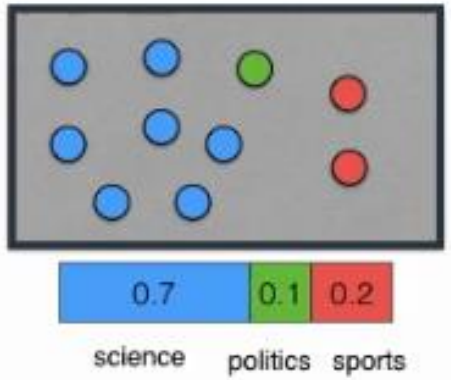


Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

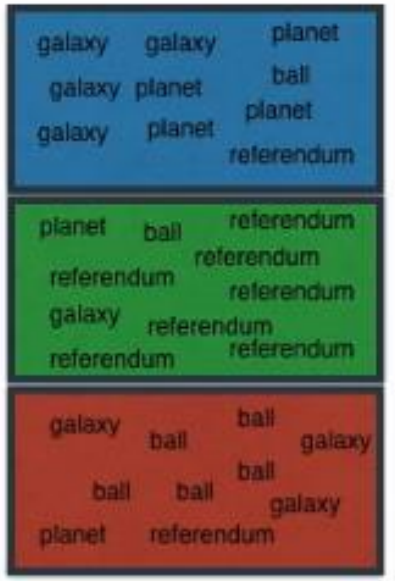
Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

$$\prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)$$

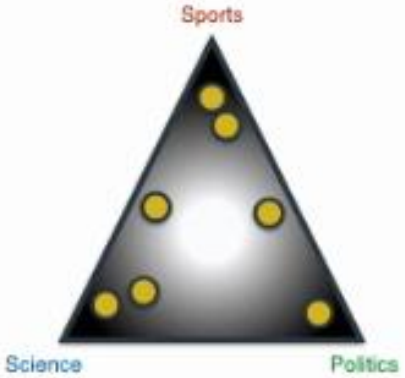


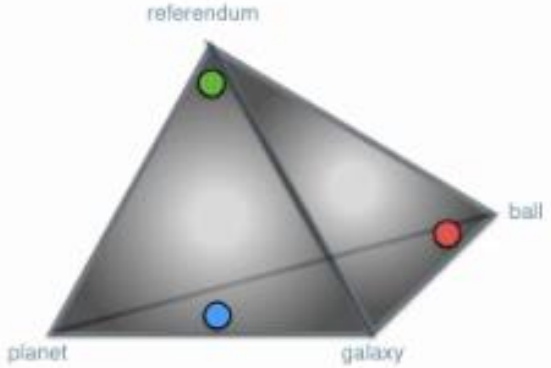
$$P(W_{j,t} \mid \varphi_{Z_{j,t}})$$



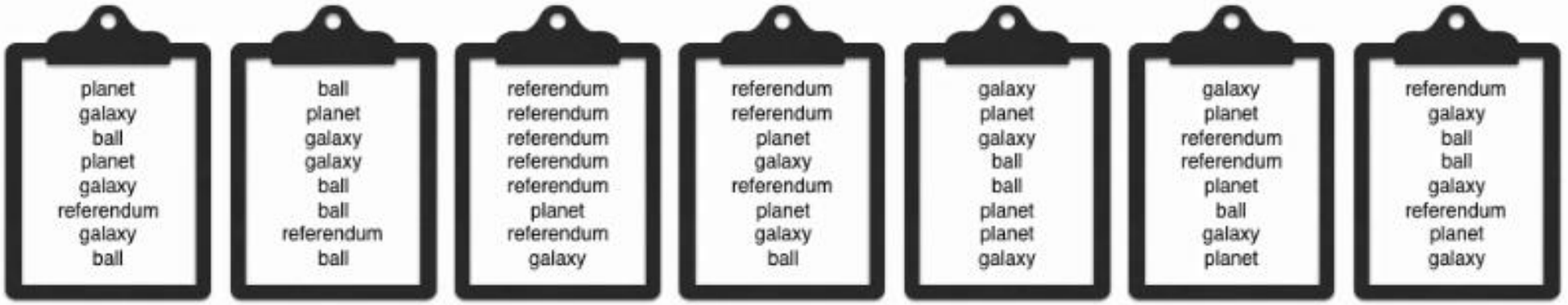
Topics	Words
science	planet
science	galaxy
sports	ball
science	planet
science	galaxy
politics	referendum
sports	galaxy
sports	ball
science	referendum

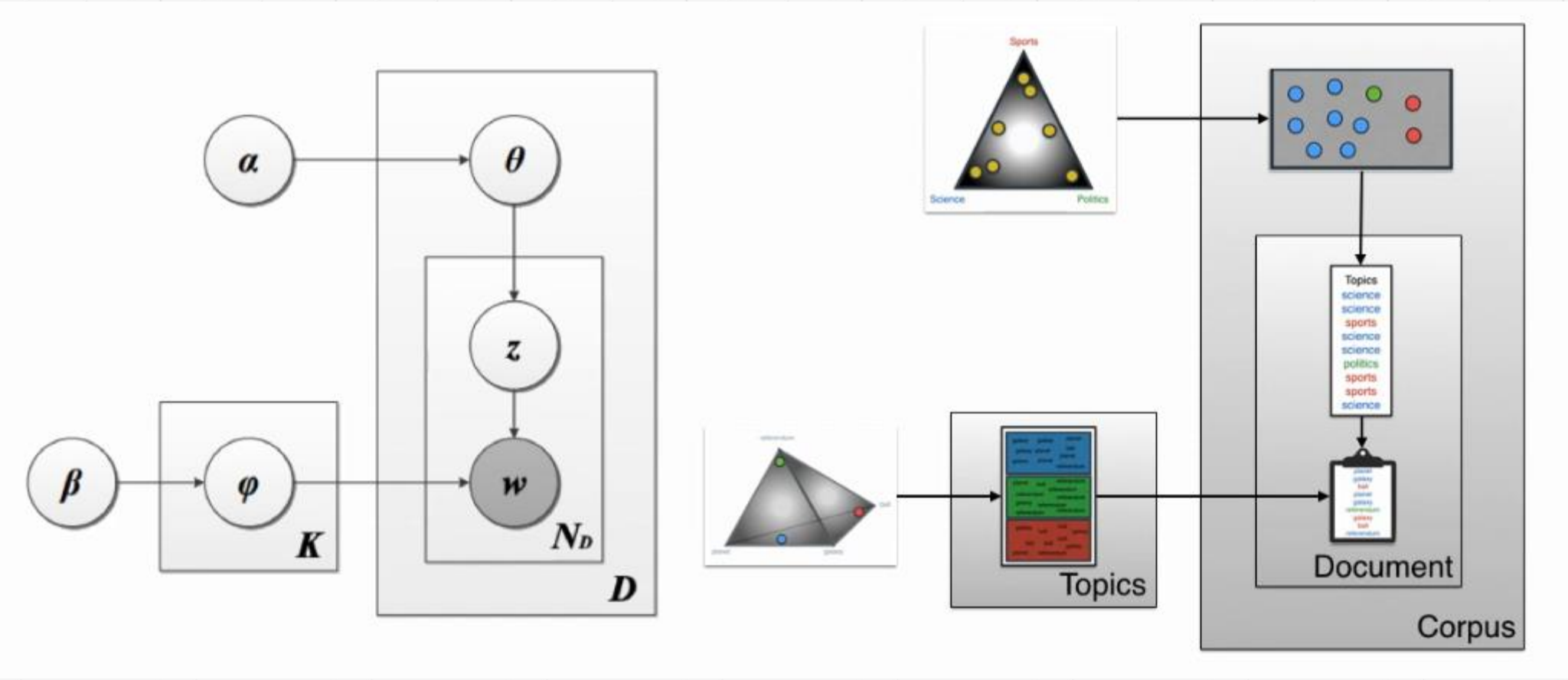
$$\prod_{j=1}^{M} P(\theta_j;\alpha) \qquad \prod_{i=1}^{K} P(\varphi_i;\beta) \qquad \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \qquad P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

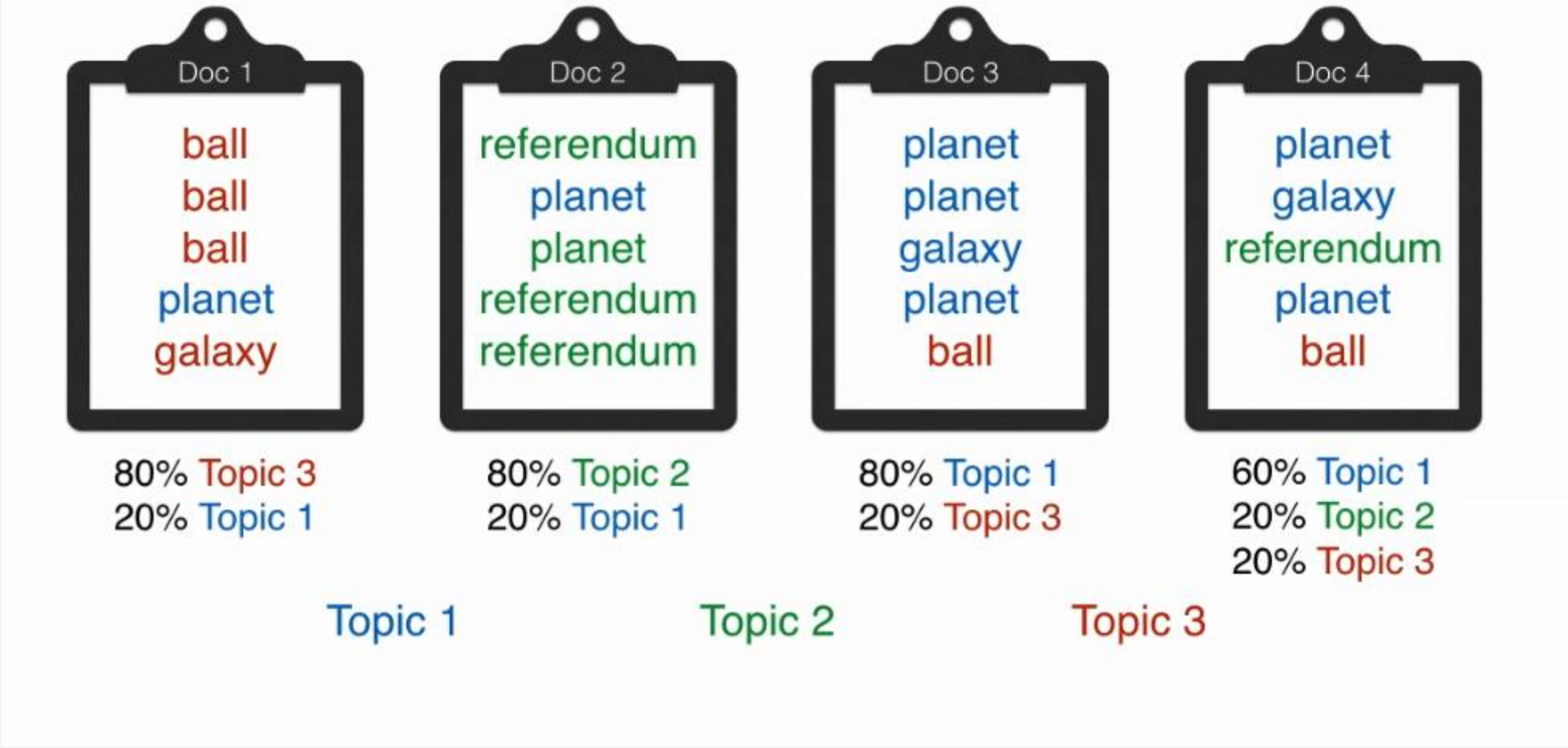




P(same articles) = low




α
θ
z
w
β
φ
K
N_D
D
Sports
Science
Politics
Topics
science
science
sports
science
science
politics
sports
sports
science
Topics
Document
Corpus


Doc 1
ball
ball
ball
planet
galaxy
Doc 2
referendum
planet
planet
referendum
referendum
Doc 3
planet
planet
galaxy
planet
ball
Doc 4
planet
galaxy
referendum
planet
ball
80% Topic 3
20% Topic 1
80% Topic 2
20% Topic 1
80% Topic 1
20% Topic 3
60% Topic 1
20% Topic 2
20% Topic 3
Topic 1
Topic 2
Topic 3