Topic Modeling

CPE 393: Text Analytics

Dr. Sansiri Tarnpradab

Department of Computer Engineering
King Mongkut's University of Technology Thonburi

Review

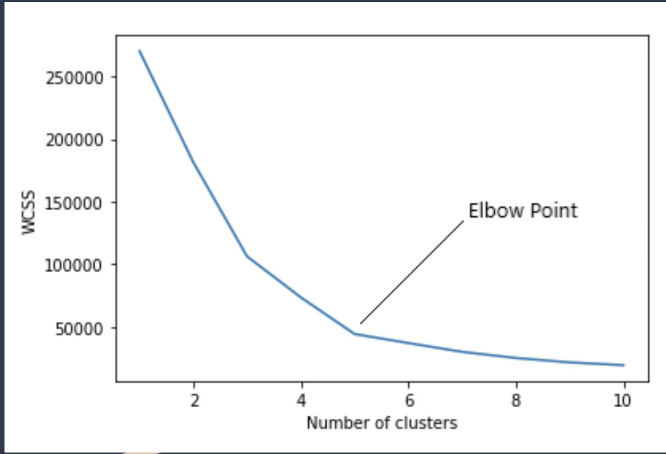
Text Clustering

- **Text Clustering**
  - Similarity
  - Distance functions
  - Quality (inter-cluster, intra-cluster)
  - K-means clustering

- Some **drawbacks** for K-means clustering
  - Dependent on K value
  - Trial-and-error

- **Elbow method**

Elbow Method



- A technique to determine an optimal number of clusters
  - **Plotting** output from different K values
  - **Identifying** the elbow point

- Computation
  - **WCSS** (Within-Cluster Sum of Square)
  - Squared average distance of all the points within a cluster to the cluster centroid
  - As the number of clusters increases, the WCSS value decreases

- Weaknesses
  - May not hold for complex datasets with irregularly shaped or differently sized clusters
  - Sensitive to initial cluster centroids
  - Inefficient for large datasets
  - Only works for K-means clustering

That was *Hard* Clustering...

Intro

Pattern Matching

Text Visualization

Web Scraping

Text Preparation

Text Feature Representation

Text Classification

Text Summarization

Text Clustering

Topic Modeling

TBA

Presentation

Outline

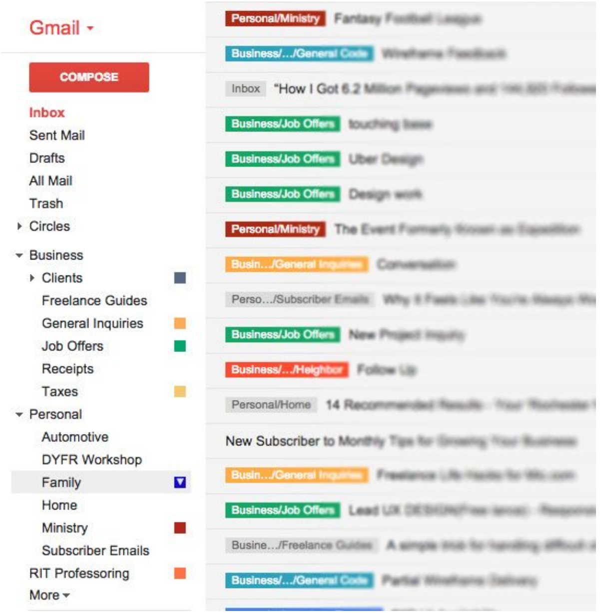
- Intro to Topic Modeling
- **LSA**: Latent Semantic Analysis
- **LDA**: Latent Dirichlet Allocation
- **BERTopic**

Topic Modeling

Why?

Imagine you have **1M** emails to sort through..

- School
- Meeting
- Finance
- Internship



Ref: https://mattolpinski.com/articles/gmail-label-system/

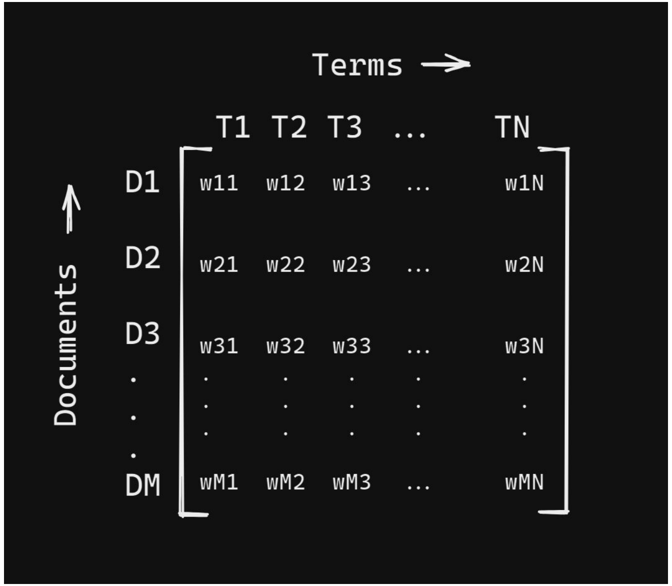
Topic Modeling

In a Nutshell

- **Unsupervised** Learning
- A **technique** for discovering abstract topics in a collection of documents.
- **Goal**:
  - To discover latent topics within a corpus
  - Latent = Hidden
- **Key Elements**:
  - Every document is a mix of topics
  - Every topic is a mix of words
- **Input**
  - Topics & words
  - Document-term matrix
- **Output**
  - Various topics

Revisit

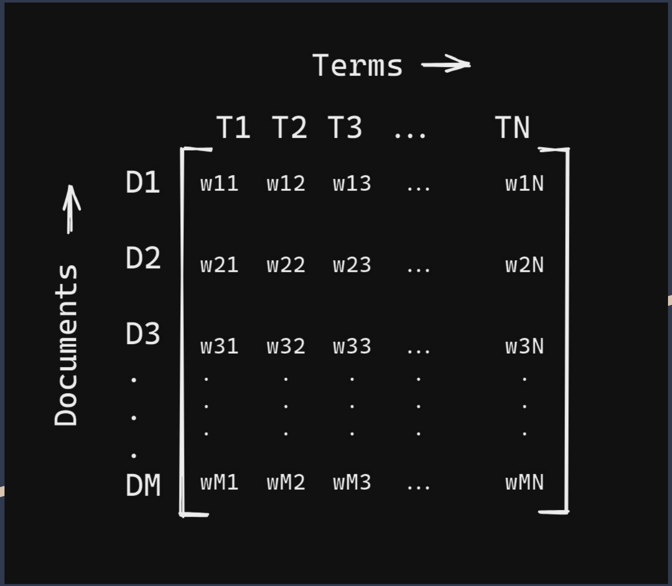
Document-Term Matrix

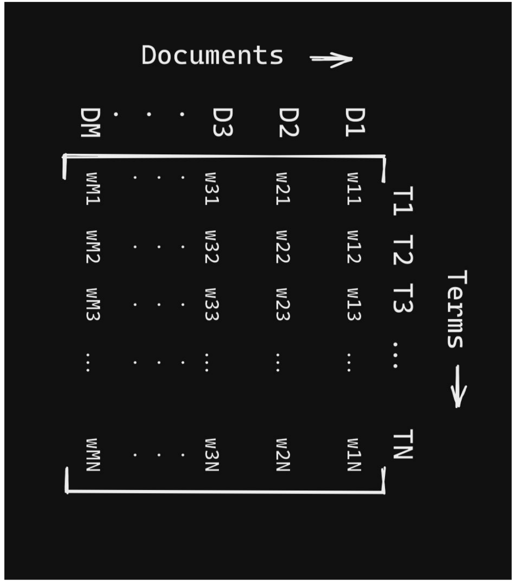


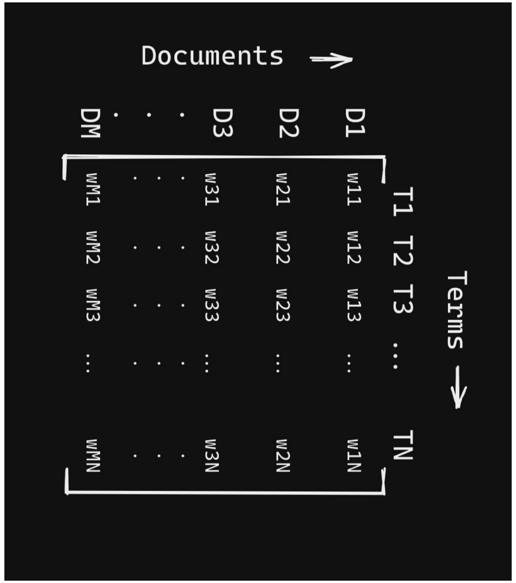
Latent Semantic Analysis (LSA)

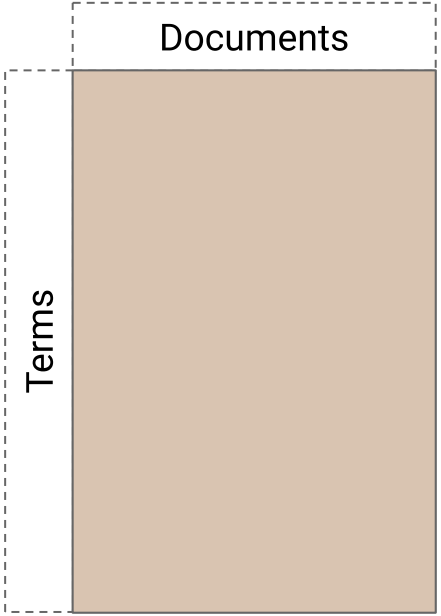
Latent Semantic Analysis: *LSA*

- Introduced in the late 1980s
- Statistical Method
- To analyze terms & documents relationships
- Text data → **Term-Document Matrix**





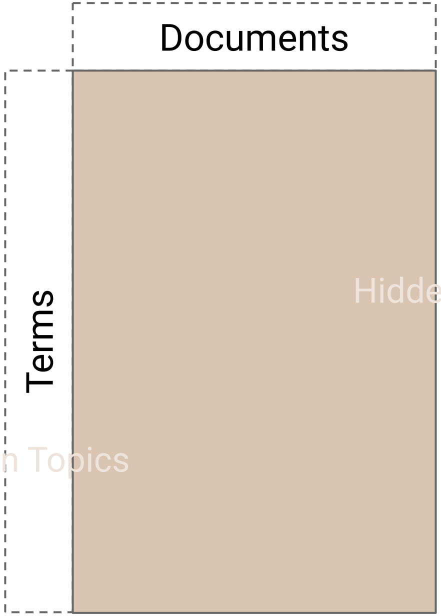




Let’s say..

This matrix presents a total of n documents and m terms.
What’s the dimension of this matrix?

Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics



Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics
Hidden Topics

Matrix:

Dimension Property

In order for matrix multiplication to be defined, the number of columns in the first matrix must be equal to the number of rows in the second matrix.

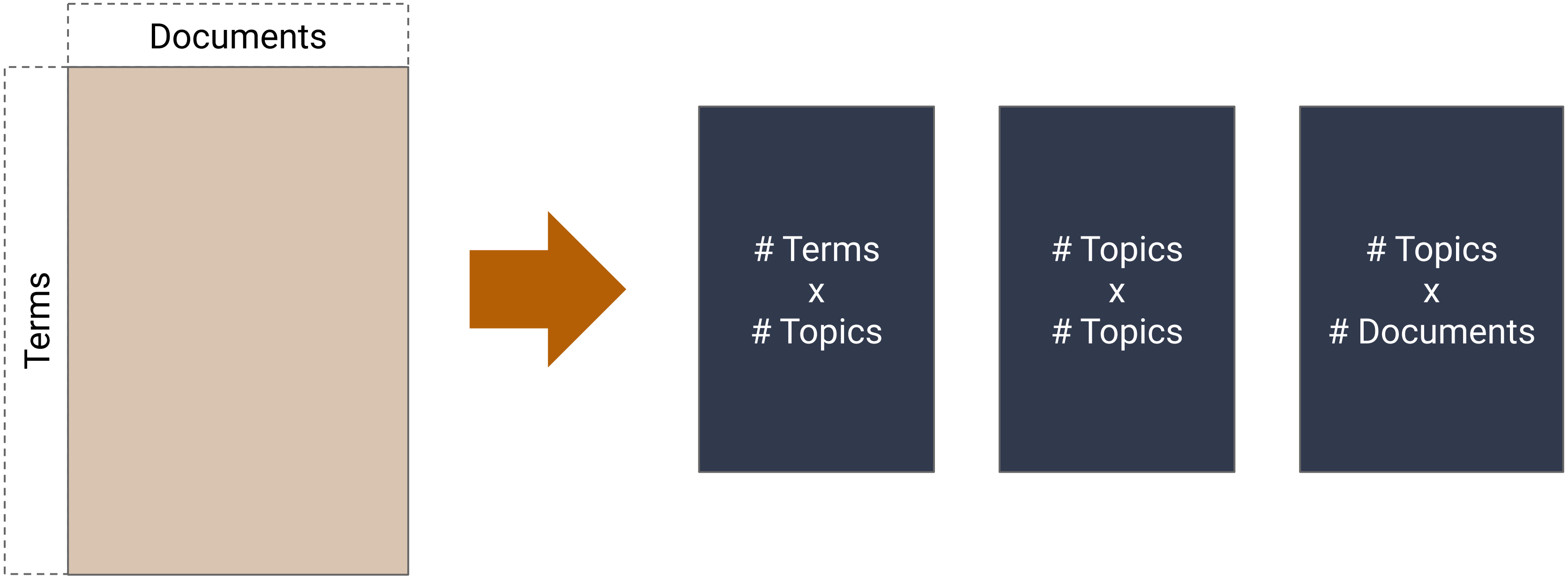
1	3
2	4
2	5

A

1	3	2	2
2	4	5	1

B

What's the dimension of AB?


Documents
Terms
Terms
x
Topics
Topics
x
Topics
Topics
x
Documents

Matrix Factorization

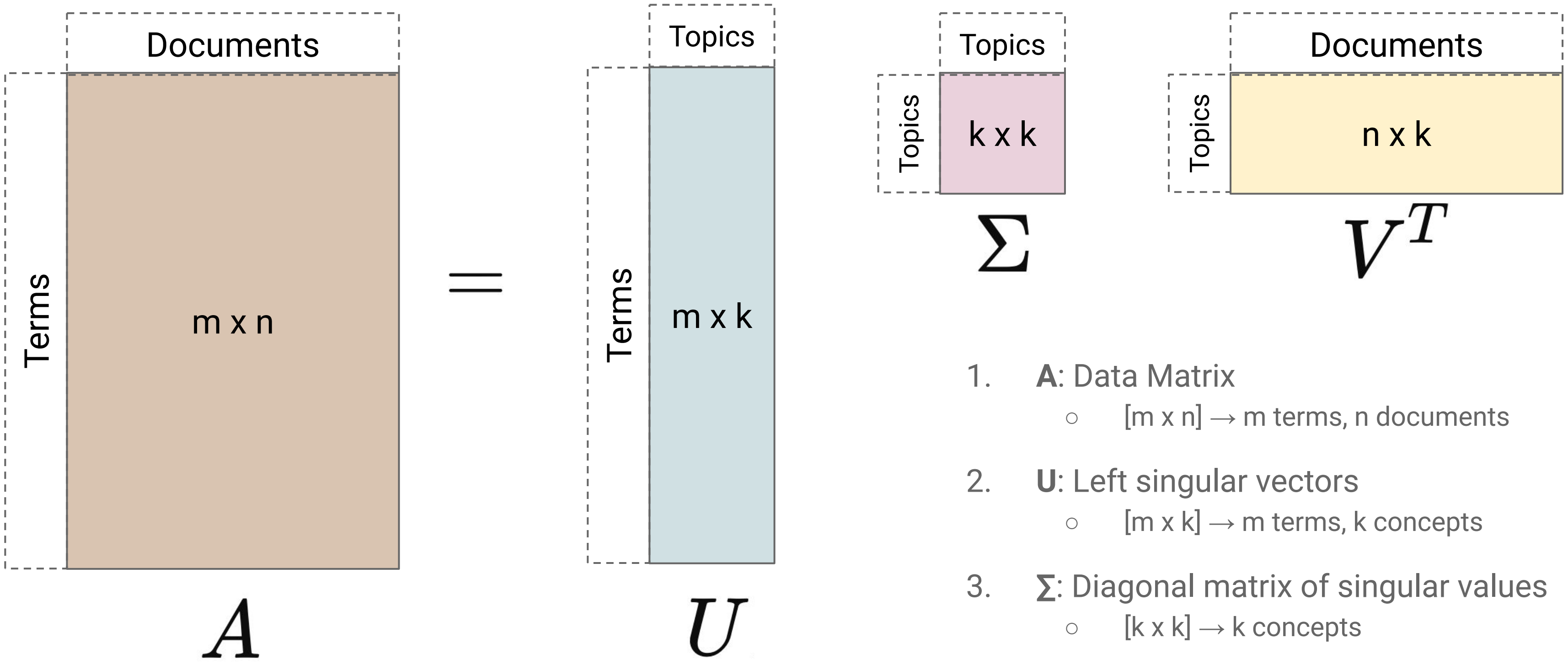
$$A = U\Sigma V^T$$

1. **A**: Data Matrix
   - [m x n] → m terms, n documents
   - Aka Term-Document Matrix

Singular Value Decomposition (SVD) decomposes the term-document matrix A (1) into three matrices (2)-(4):

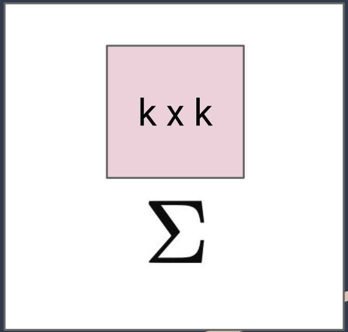
2. **U**: Left singular vectors
   - [m x k] → m terms, k concepts
   - Word Assignment to Topics
3. **∑**: Diagonal matrix of singular values
   - [k x k] → k concepts
   - Topic Importance
4. **V^T**: Right singular vectors
   - [n x k] → n documents, k concepts
   - Topic Distribution Across Documents

$$A = U\Sigma V^T$$



1. **A**: Data Matrix
   - [m x n] → m terms, n documents
2. **U**: Left singular vectors
   - [m x k] → m terms, k concepts
3. **Σ**: Diagonal matrix of singular values
   - [k x k] → k concepts
4. **V^T**: Right singular vectors
   - [n x k] → n documents, k concepts

Diagonal Matrix of Singular Values



- Capture the significance of each latent dimension in the data
- Values arranged in descending order along the diagonal

12.5	0	0
0	9.0	0
0	0	1.4

- Each **singular value** indicates the importance of the concept in the reduced space
- Larger value:
  - Capture more of the variance in the data
  - Represent the most significant underlying pattern
- Hence, Singular Value Decomposition

Singular Value Decomposition: *SVD*

Capture Hidden Patterns

➔ Topics

Dimensionality Reduction

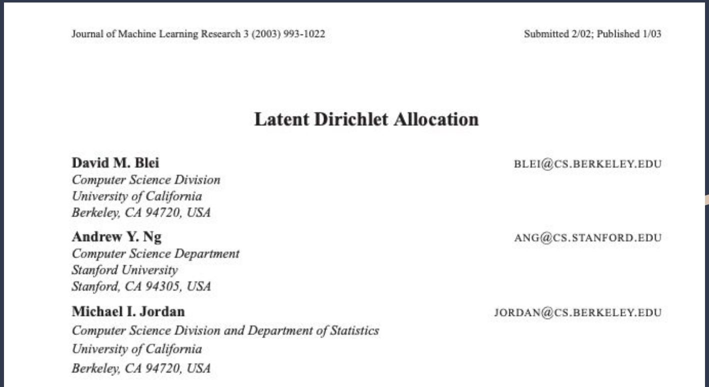
➔ Retaining only the top k singular values

Noise Reduction

➔ SVD in LSA helps filter out noise from the data
➔ Focusing on the most significant singular values and vectors

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation: *LDA*



Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Read the paper

- Introduced in 2003
- **Probabilistic** model
- Latent → Hidden
- Dirichlet → Type of probability distribution
- Allocation → Allocation

- Key Elements:
  - Every document is a mix of topics
    Every document is a distribution of topics
  - Every topic is a mix of words
    Every topic is a distribution of words

Example

5 Documents

The apple was crisp and sweet, bursting with flavor.
The playful puppy chased its tail in circles around the yard.
The lion roared loudly in the jungle, asserting its dominance.
The ripe banana was yellow and fragrant, ready to be enjoyed as a healthy snack.
The curious monkey reached for the juicy mango hanging from the tree.

The apple was crisp and sweet, bursting with flavor.	**TOPIC A**
The playful puppy chased its tail in circles around the yard.	**TOPIC B**
The lion roared loudly in the jungle, asserting its dominance.	**TOPIC B**
The ripe banana was yellow and fragrant, ready to be enjoyed as a healthy snack.	**TOPIC A**
The curious monkey reached for the juicy mango hanging from the tree.	**TOPIC A & B**

The apple was crisp and sweet, bursting with flavor.	**TOPIC A**
The playful puppy chased its tail in circles around the yard.	**TOPIC B**
The lion roared loudly in the jungle, asserting its dominance.	**TOPIC B**
The ripe banana was yellow and fragrant, ready to be enjoyed as a healthy snack.	**TOPIC A**
The curious monkey reached for the juicy mango hanging from the tree.	**TOPIC A & B**

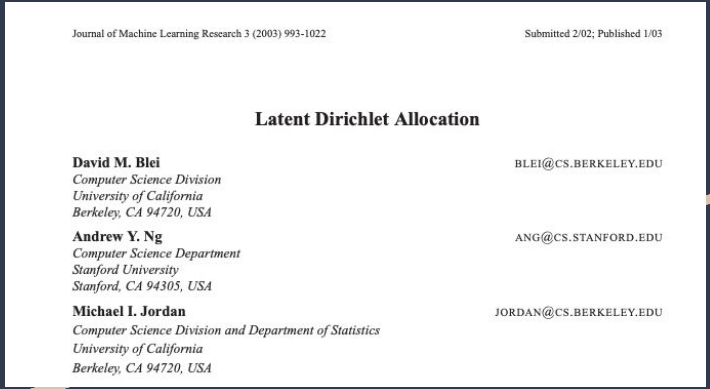
TOPIC A: apple, banana, mango, crisp, sweet, flavor, juicy, …

TOPIC B: puppy, tail, lion, monkey, roared, dominance, curious, …

(each has percentage of distribution)

The apple was crisp and sweet, bursting with flavor.	**FRUIT**
The playful puppy chased its tail in circles around the yard.	**ANIMAL**
The lion roared loudly in the jungle, asserting its dominance.	**ANIMAL**
The ripe banana was yellow and fragrant, ready to be enjoyed as a healthy snack.	**FRUIT**
The curious monkey reached for the juicy mango hanging from the tree.	**FRUIT & ANIMAL**

Latent Dirichlet Allocation: *LDA*


Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Read the paper

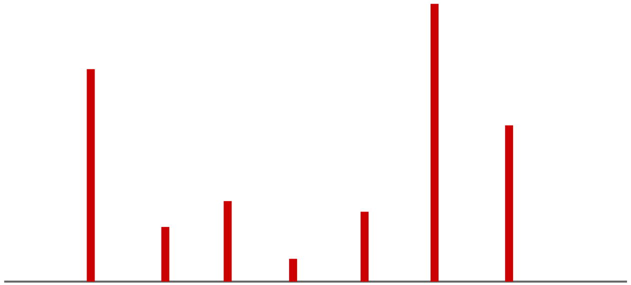
- 2003
- **Probabilistic** model
- Latent → Hidden
- Dirichlet → Type of probability distribution
- Allocation → Allocation

- Key Elements:
  - Every document is a mix of topics
    Every document is a distribution of topics
  - Every topic is a mix of words
    Every topic is a distribution of words

The apple was crisp and sweet, bursting with flavor.	**TOPIC A**
The playful puppy chased its tail in circles around the yard.	**TOPIC B**
The lion roared loudly in the jungle, asserting its dominance.	**TOPIC B**
The ripe banana was yellow and fragrant, ready to be enjoyed as a healthy snack.	**TOPIC A**
The curious monkey reached for the juicy mango hanging from the tree.	**TOPIC A & B**

TOPIC A: apple, banana, mango, crisp, sweet, flavor, juicy, …

TOPIC B: puppy, tail, lion, monkey, roared, dominance, curious, …



Word distribution of TOPIC A

How *LDA* works

FRUIT

ANIMAL

The playful puppy chased its tail in circles around the yard.

- Choose the number of topics beforehand (K)
- Randomly assign each word in each document to one of the K topics
- Go through every word and its assignment
  - How often the topic occurs in the document?
  - How often the word occurs in the topic overall?
- Assign the word to a new topic accordingly
- Perform this for multiple iterations

Walk-Through

Step 1

The playful puppy chased its tail in circles around the yard.

Choose the number of topics beforehand (K)

Let the number of topics be **two**:

1. Fruit
2. Animal

Walk-Through

Step 2

The playful puppy chased its tail in circles around the yard.

- Randomly assign each word in each document to one of the K topics
- Aka **Topic-word Distribution**

Topic 1 (Fruit) might start with words like "chased" and "yard."

Topic 2 (Animal) might start with words like "puppy" and "tail."

chased
circles
yard

Fruit

playful
puppy
tail

Animal

Walk-Through

Step 3 (Iterative)

The playful puppy chased its tail in circles around the yard.

Go through every word and its assignment

- How often the topic occurs in the document?
  - Document-Topic Distribution
- How often the word occurs in the topic overall?
  - Topic-Word Distribution

Fruit	Animal
chased	playful
circles	puppy
yard	tail

- Assign the word to a new topic accordingly

Walk-Through

Step ...

The playful puppy chased its tail in circles around the yard.

After several iterations, we may observe the following refinements:

- Words like "playful," "puppy," and "tail" increasingly align with Animal.
- Words like "yard" and "circles" might stay with Fruit if they lack strong relevance to Animal.

Fruit	Animal
circles	playful
yard	puppy
	chased
	tail

Walk-Through

Final Step

The playful puppy chased its tail in circles around the yard.

Let's talk about the final output..

Document-Topic Distribution:

➔ LDA will estimate that this document is primarily about the Animal topic (e.g., 80% Animal, 20% Fruit).

Topic-Word Distribution:

- **Topic 1** (Fruit): Words like "circles" and "yard" may have a small association.
- **Topic 2** (Animal): Strongly associated with "playful," "puppy," "chased," and "tail."

Multiple Documents

"I love playing football."

"The match was exciting."

"The player scored a goal."

Topics

Sports and **Emotions**

Adjustment of distribution through iterations..

I love playing football.

The match was exciting.

The player scored a goal.

With each iteration, LDA updates:

- Document-Topic Distribution
- Topic-Word Distribution

Let's talk about the final output..

Topic-Word Distribution

Document-Topic Distribution

In Practice

```
class gensim.models.ldamodel.LdaModel(corpus=None, num_topics=100, id2word=None, distributed=False, chunksize=2000, passes=1, update_every=1, alpha='symmetric', eta=None, decay=0.5, offset=1.0, eval_every=10, iterations=50, gamma_threshold=0.001, minimum_probability=0.01, random_state=None, ns_conf=None, minimum_phi_value=0.01, per_word_topics=False, callbacks=None, dtype=<class 'numpy.float32'>)
```

Input

- Document-Term Matrix
- Number of topics (K)
- Number of iterations

Gensim

- Go through every word
- Find the best word & topic distribution
- Assignment

Output

- Top words in each topic
- Adjust parameter as needed

LDA vs *LSA*

	LSA	LDA
Type	Statistical method	Probabilistic method
Technique	Singular Value Decomposition (SVD)	Bayesian inference with Dirichlet distributions
Document Representation	Concepts are derived from word co-occurrence patterns	Topics are probabilistic distributions over words
Scalability	Effective with smaller datasets; performance decreases with size	Scales well with large datasets, widely used in big data applications
Interpretability	Concepts are less interpretable; dimensions represent abstract concepts	Topics are explicit with assigned probabilities
Applications	Better for semantic search, synonym discovery	Better for text generation, topic categorization

Applications

- Text summarization
- Information retrieval
- Recommendation systems
- Sentiment analysis
- And more…

Conclusion

- Intro to Topic Modeling
- **LSA**: Latent Semantic Analysis
- **LDA**: Latent Dirichlet Allocation
- **BERTopic**

Q & A