Gibbs Sampling for LDA and Applications to RAG

Kyle Torres
Advisor: Prof. Hardin

March 7, 2025

Motivating Example

Document 1

ball
score
goal
brownie
ball

Document 2

policy
vote
pie
policy
state

Document 3

pizza
pie
pizza
brownie
ball

Motivating Example

Document 1	Document 2	Document 3
ball score goal brownie ball	policy vote pie policy state	pizza pie pizza brownie ball
Sports		

Motivating Example

Document 1

ball
score
goal
brownie
ball

Sports

Document 2

policy
vote
pie
policy
state

Politics

Document 3

pizza
pie
pizza
brownie
ball

Motivating Example

Document 1	**Document 2**	**Document 3**
ball score goal brownie ball	policy vote pie policy state	pizza pie pizza brownie ball
Sports	**Politics**	**Food**

Motivating Example (con.)

Document 1

ball
score
goal
brownie
ball

Document 2

policy
vote
pie
policy
state

Document 3

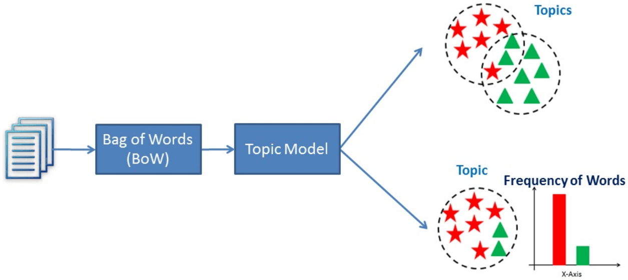
pizza
pie
pizza
brownie
ball

Definition

A method for unsupervised classification of documents that identifies clusters of similar words (topics).

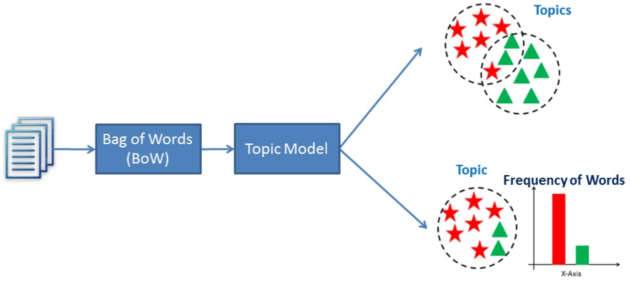
Definition

A method for unsupervised classification of documents that identifies clusters of similar words (topics).



Definition

A method for unsupervised classification of documents that identifies clusters of similar words (topics).



Applications

Sentiment analysis, recommender systems, **information retrieval**, etc.

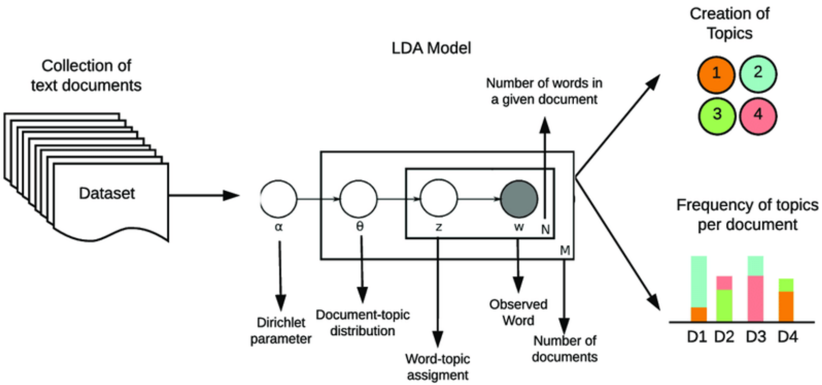
Definition

Latent Dirichlet Allocation (LDA) is a probabilistic model that extracts topics from a corpus of text.

Latent Dirichlet Allocation (LDA)

Definition

Latent Dirichlet Allocation (LDA) is a probabilistic model that extracts topics from a corpus of text.



Some Bayesian Terminology

Prior

The *prior* $P(\theta)$ is the probability distribution that represents one's beliefs about some parameter θ before some evidence is taken into account.

Some Bayesian Terminology

Prior

The *prior* $P(\theta)$ is the probability distribution that represents one's beliefs about some parameter θ before some evidence is taken into account.

Likelihood

The *likelihood* $P(Y \mid \theta)$ measures how well a model explains observed data.

Some Bayesian Terminology

Prior

The *prior* $P(\theta)$ is the probability distribution that represents one's beliefs about some parameter θ before some evidence is taken into account.

Likelihood

The *likelihood* $P(Y \mid \theta)$ measures how well a model explains observed data.

Posterior

The *posterior* $P(\theta \mid Y)$ is the probability distribution that combines information from the prior and the data using Bayes' Theorem.

Some Bayesian Terminology

Prior

The *prior* $P(\theta)$ is the probability distribution that represents one's beliefs about some parameter θ before some evidence is taken into account.

Likelihood

The *likelihood* $P(Y \mid \theta)$ measures how well a model explains observed data.

Posterior

The *posterior* $P(\theta \mid Y)$ is the probability distribution that combines information from the prior and the data using Bayes' Theorem.

$$\underbrace{P(\theta|Y)}_{\text{Posterior}} = \frac{P(Y|\theta)\,P(\theta)}{P(Y)} \propto \underbrace{P(Y|\theta)}_{\text{Likelihood}}\,\underbrace{P(\theta)}_{\text{Prior}}$$

LDA Model Assumptions/Generative Process

1. Draw a topic distribution for each document:
   - Each document d has a corresponding topic distribution θ_d
   - Sample $\theta_d \sim \text{Dirichlet}(\alpha)$

LDA Model Assumptions/Generative Process

1. Draw a topic distribution for each document:
   - Each document d has a corresponding topic distribution θ_d
   - Sample $\theta_d \sim \text{Dirichlet}(\alpha)$

2. Draw a word distribution for each topic:
   - Each topic k has a word distribution φ_k
   - Sample $\varphi_k \sim \text{Dirichlet}(\beta)$

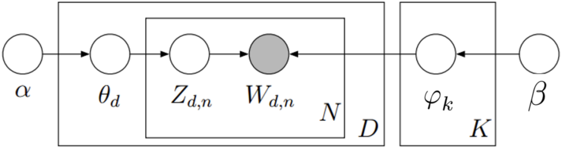
LDA Model Assumptions/Generative Process

1. Draw a topic distribution for each document:
   - Each document d has a corresponding topic distribution θ_d
   - Sample $\theta_d \sim \text{Dirichlet}(\alpha)$

2. Draw a word distribution for each topic:
   - Each topic k has a word distribution φ_k
   - Sample $\varphi_k \sim \text{Dirichlet}(\beta)$

3. Generate words for each document:

   For each word position n in document d:
   - Sample a topic $z_{d,n} \sim \text{Multinomial}(\theta_d)$
   - Sample a word $w_{d,n} \sim \text{Multinomial}(\varphi_k)$



Definition

Gibbs Sampling is a Markov chain Monte Carlo (MCMC) algorithm that samples from a multivariate probability distribution.

Definition

Gibbs Sampling is a Markov chain Monte Carlo (MCMC) algorithm that samples from a multivariate probability distribution.

- Usually used when direct sampling from the joint distribution is intractable and sampling from the conditional distribution is more practical

Definition

Gibbs Sampling is a Markov chain Monte Carlo (MCMC) algorithm that samples from a multivariate probability distribution.

- Usually used when direct sampling from the joint distribution is intractable and sampling from the conditional distribution is more practical
- Instead of calculating the joint probability of topic assignments $p(z_1, z_2, \cdots, z_n)$, we will be calculating conditionals $p(z_i | z_1, z_2 \cdots, z_n)$, rewritten as $p(z_i | z_{-i})$

Definition

Gibbs Sampling is a Markov chain Monte Carlo (MCMC) algorithm that samples from a multivariate probability distribution.

- Usually used when direct sampling from the joint distribution is intractable and sampling from the conditional distribution is more practical
- Instead of calculating the joint probability of topic assignments $p(z_1, z_2, \cdots, z_n)$, we will be calculating conditionals $p(z_i | z_1, z_2 \cdots, z_n)$, rewritten as $p(z_i | z_{-i})$
- By updating variables while keeping others fixed, we create a Markov Chain, and after many iterations, the chain converges to the joint distribution

Definition

Gibbs Sampling is a Markov chain Monte Carlo (MCMC) algorithm that samples from a multivariate probability distribution.

- Usually used when direct sampling from the joint distribution is intractable and sampling from the conditional distribution is more practical
- Instead of calculating the joint probability of topic assignments $p(z_1, z_2, \cdots, z_n)$, we will be calculating conditionals $p(z_i|z_1, z_2 \cdots, z_n)$, rewritten as $p(z_i|z_{-i})$
- By updating variables while keeping others fixed, we create a Markov Chain, and after many iterations, the chain converges to the joint distribution
- We assume the Markov Chain is ergodic, meaning that is converges to some stationary distribution regardless of the initial state

Gibbs Sampling Formula

Conditional Probability

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

Gibbs Sampling Formula

Conditional Probability

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

$P(z_i = k')$: probability that the topic assigned at the i^{th} token is k' where i corresponds to the topic and word, $\{z_i, w_i = (d', j')\}$

Gibbs Sampling Formula

Conditional Probability

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

$P(z_i = k')$: probability that the topic assigned at the i^{th} token is k' where i corresponds to the topic and word, $\{z_i, w_i = (d', j')\}$

$C(d', k')_{-i}$: $\#$ of words in document d' assigned to topic k', excluding w_i

Gibbs Sampling Formula

Conditional Probability

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

$P(z_i = k')$: probability that the topic assigned at the i^{th} token is k' where i corresponds to the topic and word, $\{z_i, w_i = (d', j')\}$

$C(d', k')_{-i}$: # of words in document d' assigned to topic k', excluding w_i

$C(k', v')_{-i}$: # of times word v' has been assigned to topic k', excluding w_i

Gibbs Sampling Formula

Conditional Probability

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

$P(z_i = k')$: probability that the topic assigned at the i^{th} token is k' where i corresponds to the topic and word, $\{z_i, w_i = (d', j')\}$

$C(d', k')_{-i}$: # of words in document d' assigned to topic k', excluding w_i

$C(k', v')_{-i}$: # of times word v' has been assigned to topic k', excluding w_i

α: Dirichlet hyperparameter that defines prior information for document-topic distributions θ_d

Gibbs Sampling Formula

Conditional Probability

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

$P(z_i = k')$: probability that the topic assigned at the i^{th} token is k' where i corresponds to the topic and word, $\{z_i, w_i = (d', j')\}$

$C(d', k')_{-i}$: # of words in document d' assigned to topic k', excluding w_i

$C(k', v')_{-i}$: # of times word v' has been assigned to topic k', excluding w_i

α: Dirichlet hyperparameter that defines prior information for document-topic distributions θ_d

β: Dirichlet hyperparameter that defines prior information for topic-word distributions φ_k

Gibbs Sampling Process

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

For each token i:

1. Compute the probability for each topic using the Gibbs Sampling formula

Gibbs Sampling Process

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

For each token i:

1. Compute the probability for each topic using the Gibbs Sampling formula
2. Normalize these probabilities so they sum to 1

$$P(z_i = k' \mid Z_{-i}, W) \propto \left[\frac{\alpha + C(d', k')_{-i}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-i})}\right] \cdot \left[\frac{\beta + C(k', v')_{-i}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-i})}\right]$$

For each token i:

1. Compute the probability for each topic using the Gibbs Sampling formula
2. Normalize these probabilities so they sum to 1
3. Sample a new topic from this multinomial distribution

Example $(i = 47)$

Suppose we have 3 topics:

1. Compute the probability for each topic using the Gibbs Sampling formula

$$P(z_{47} = k' \mid Z_{-47}, W) \propto \left[\frac{\alpha + C(d', k')_{-47}}{\sum_{k=1}^{K} (\alpha + C(d', k)_{-47})} \right] \cdot \left[\frac{\beta + C(k', v')_{-47}}{\sum_{v=1}^{V} (\beta + C(k', v)_{-47})} \right.$$

Gibbs Sampling Process (con.)

Example $(i = 47)$

Suppose we have 3 topics:

1. Compute the probability for each topic using the Gibbs Sampling formula

$$P(z_{47} = k' \mid Z_{-47}, W) \propto \left[\frac{\alpha + C(d', k')_{-47}}{\sum_{k=1}^{K}(\alpha + C(d', k)_{-47})}\right] \cdot \left[\frac{\beta + C(k', v')_{-47}}{\sum_{v=1}^{V}(\beta + C(k', v)_{-47})}\right]$$

2. Normalize these probabilities so they sum to 1

Topic k	Probability $P(z_{47} = k \mid Z_{-47}, W)$
Topic 1	0.2
Topic 2	0.5
Topic 3	0.3

Example $(i = 47)$

Suppose we have 3 topics:

1. Compute the probability for each topic using the Gibbs Sampling formula

$$P(z_{47} = k' \mid Z_{-47}, W) \propto \left[\frac{\alpha + C(d', k')_{-47}}{\sum_{k=1}^{K} (\alpha + C(d', k)_{-47})} \right] \cdot \left[\frac{\beta + C(k', v')_{-47}}{\sum_{v=1}^{V} (\beta + C(k', v)_{-47})} \right]$$

2. Normalize these probabilities so they sum to 1

Topic k	Probability $P(z_{47} = k \mid Z_{-47}, W)$
Topic 1	0.2
Topic 2	0.5
Topic 3	0.3

3. Sample a new topic using probabilities from this multinomial distribution (Topic 2 is most likely but isn't always chosen)

Retrieval Augmented Generation (RAG)

Definition

Retrieval Augmented Generation (RAG) is a technique for incorporating information retrieval capabilities for generative artificial intelligence models.

Definition

Retrieval Augmented Generation (RAG) is a technique for incorporating information retrieval capabilities for generative artificial intelligence models.

- sequence-to-sequence model: takes query as input and generates response as output

Retrieval Augmented Generation (RAG)

Definition

Retrieval Augmented Generation (RAG) is a technique for incorporating information retrieval capabilities for generative artificial intelligence models.

- sequence-to-sequence model: takes query as input and generates response as output
- non-parametric memory (dynamic external database)

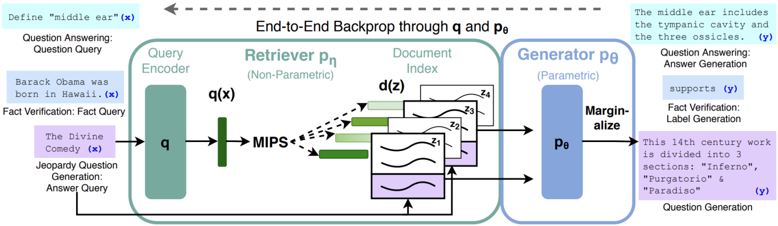
Retrieval Augmented Generation (RAG)

Definition

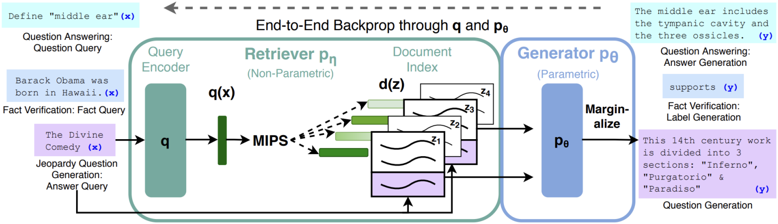
Retrieval Augmented Generation (RAG) is a technique for incorporating information retrieval capabilities for generative artificial intelligence models.

- sequence-to-sequence model: takes query as input and generates response as output
- non-parametric memory (dynamic external database)
- retrieval improves the reliability of responses by decreasing the chances of “hallucinating”

Applying LDA to RAG

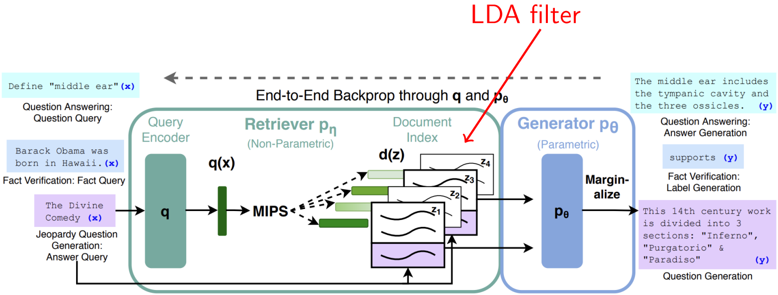


Applying LDA to RAG



- enhanced retrieval: match query with documents that contain the same topic

Applying LDA to RAG



- enhanced retrieval: match query with documents that contain the same topic
- serves as a filtering method in the pre-processing stage that reduces the search space

Future Directions

- Comparing accuracy of outputs from this experiment with results from the original paper
- Quantifying the extent to which this filtering method improves model efficiency

Thank you for listening!

Special thanks to Prof Hardin.