Applied Text Analytics & Natural Language Processing

with Dr. Mahdi Roozbahani
& Wafa Louhichi

Topic Modeling
Latent Dirichlet Allocation (LDA)

Learning Objectives

In this lesson, you will learn a topic-learning model named Latent Dirichlet Allocation (LDA)

- Cluster of documents by topic
- Cluster of words by topic

What is Topic Modeling?

It is an unsupervised learning technique (no labels needed) to extract topics from documents and find documents that potentially share a common context.

This technique is used to query documents that may not have all the keywords but are still related to a topic which is a hidden concept.

We may retrieve documents that DON’T have the term “system”, but they contain almost everything else (“data”, “retrieval”)

Latent Dirichlet Allocation (LDA)

In LDA, we need to know the number of topics in advance. Let’s say your documents are related to news documents, and you will assign three topics: sports, food, and economy. In LDA, the topics are assigned as Topic 1, Topic 2, and Topic 3. You can later label those topics based on the output of the algorithm.

Now that you know the number of topics, the LDA core suggestion is

- News documents with similar topics share common words.
- These topics can be discovered by finding a group of words occurring together in all documents
- We need to create the document-topic matrix.
- We need to create the topic-word matrix.

Latent Dirichlet Allocation (LDA)

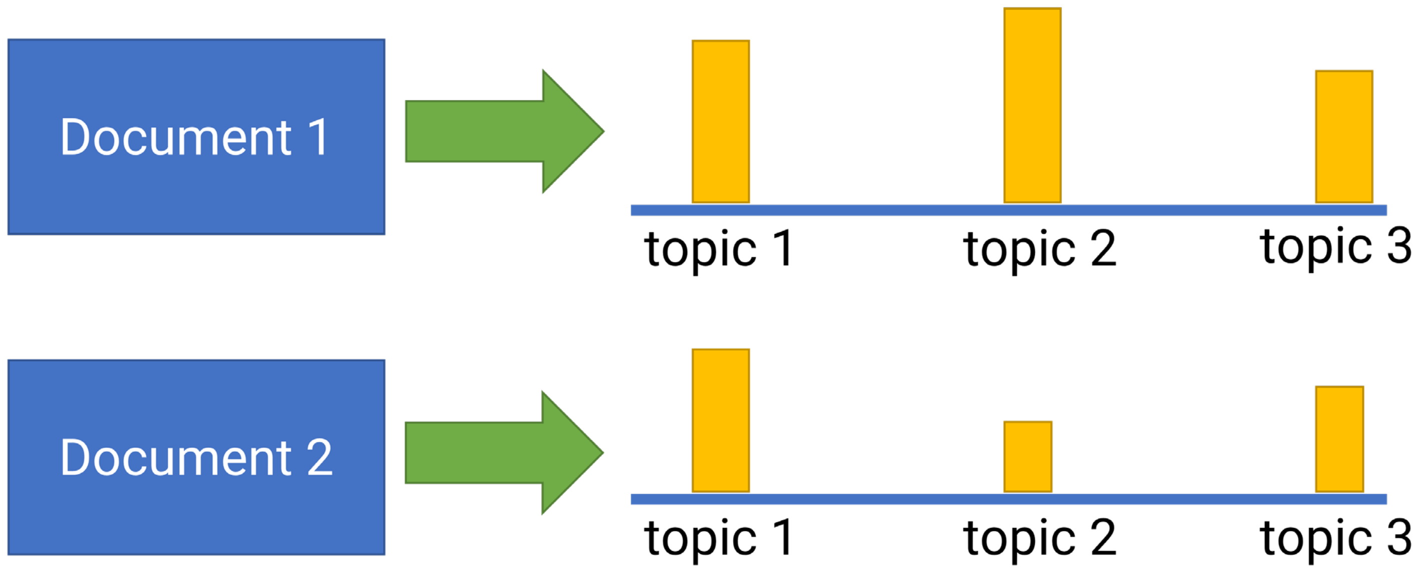
In LDA, we need to know the number of topics in advance. Let’s say your documents are related to news documents, and you will assign three topics: sports, food, and economy. In LDA, the topics are assigned as Topic 1, Topic 2, and Topic 3. You can later label those topics based on the output of the algorithm.

Now that you know the number of topics, the LDA core suggestion is

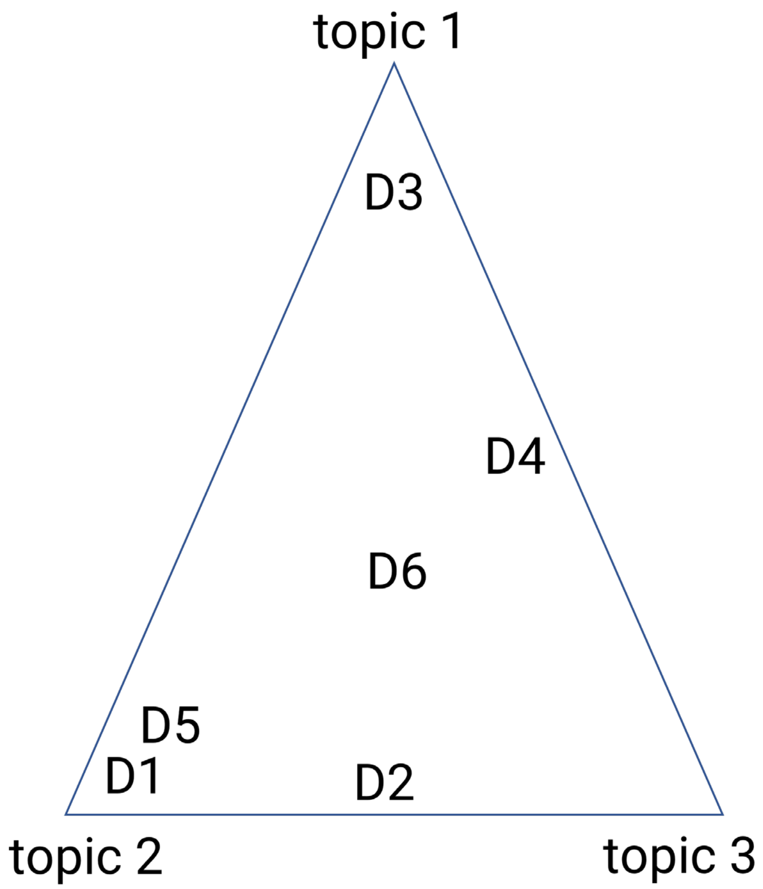
- News documents with similar topics share common words.
- These topics can be discovered by finding a group of words occurring together in all documents
- We need to create the document-topic matrix.
- We need to create the topic-word matrix.

Documents-Topics Matrix (Distribution)

For each document, we need to find the probability of each topic. LDA algorithm starts with randomly assigning each word in our corpus to a topic (topic 1, topic 2, or topic 3). Then, we can calculate in each document the word frequency for each topic, hence its probability.

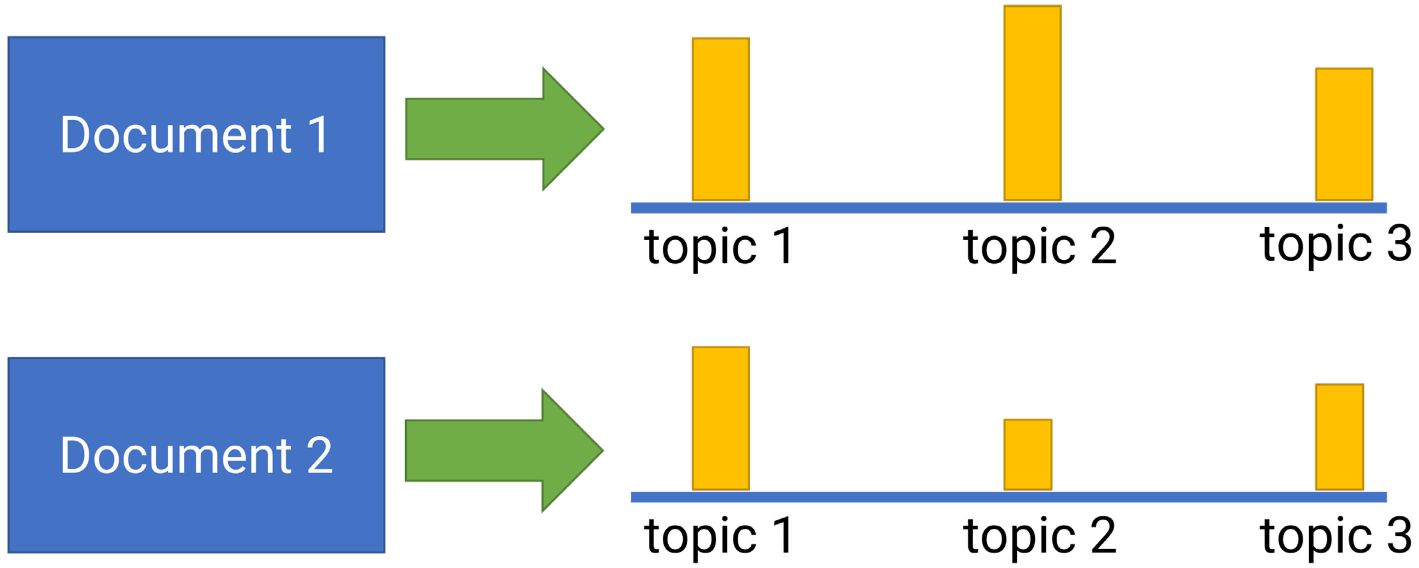


	Topic 1	Topic 2	Topic 3
Document 1	0.1	0.8	0.1

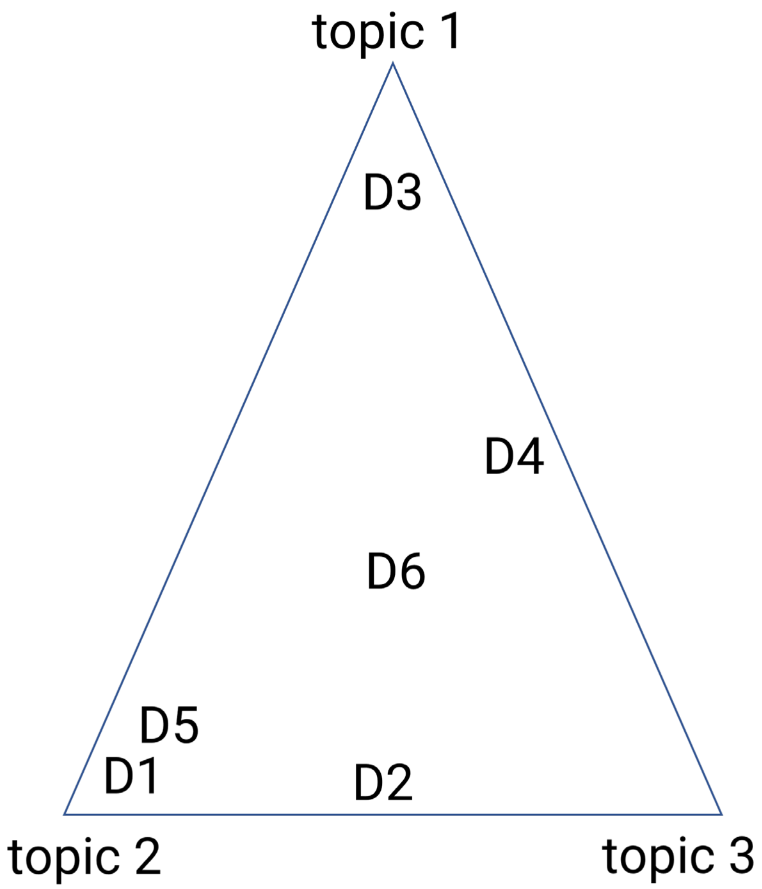


Documents-Topics Matrix (Distribution)

For each document, we need to find the probability of each topic. LDA algorithm starts with randomly assigning each word in our corpus to a topic (topic 1, topic 2, or topic 3). Then, we can calculate in each document the word frequency for each topic, hence its probability.

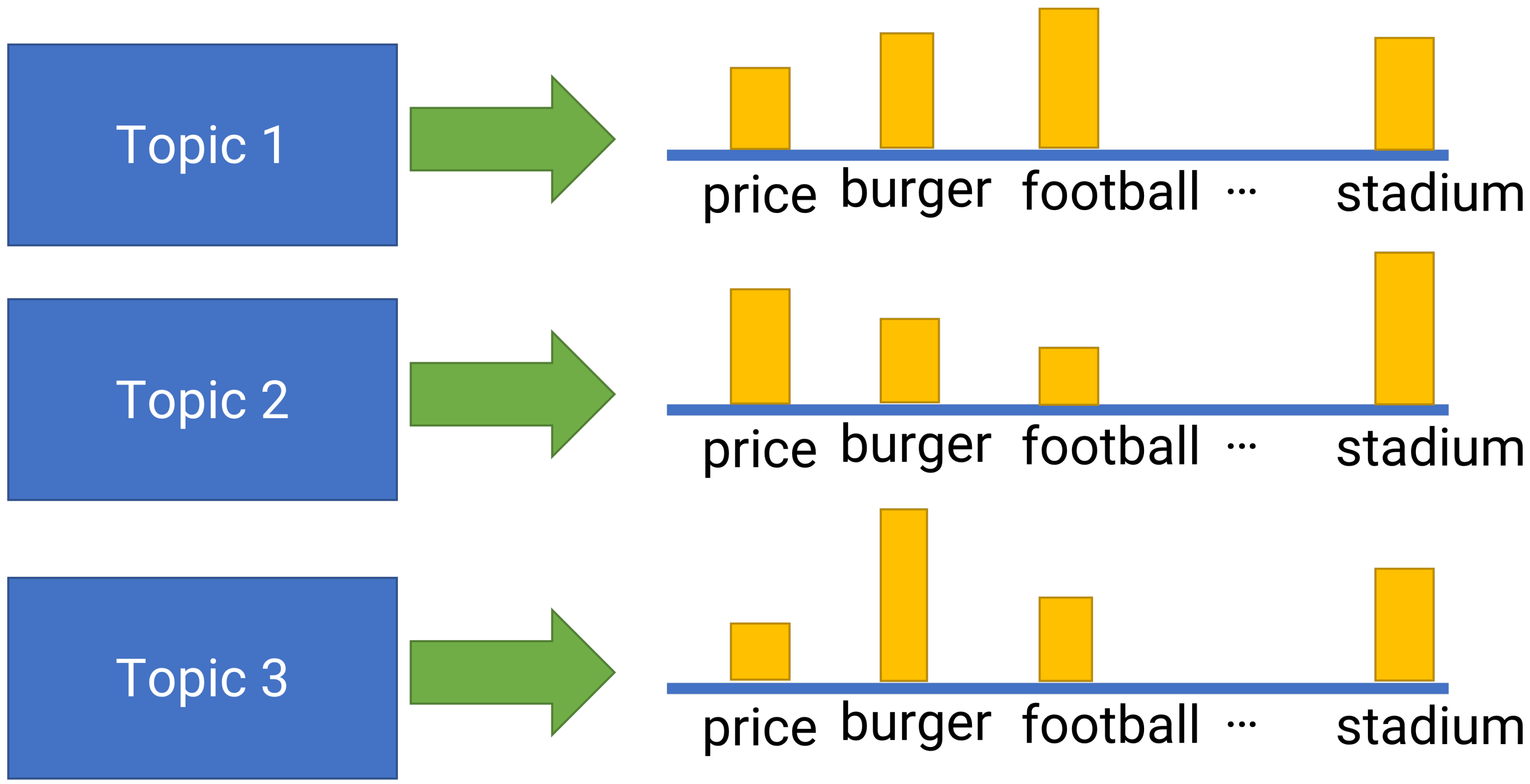


	Topic 1	Topic 2	Topic 3
Document 1	0.1	0.8	0.1

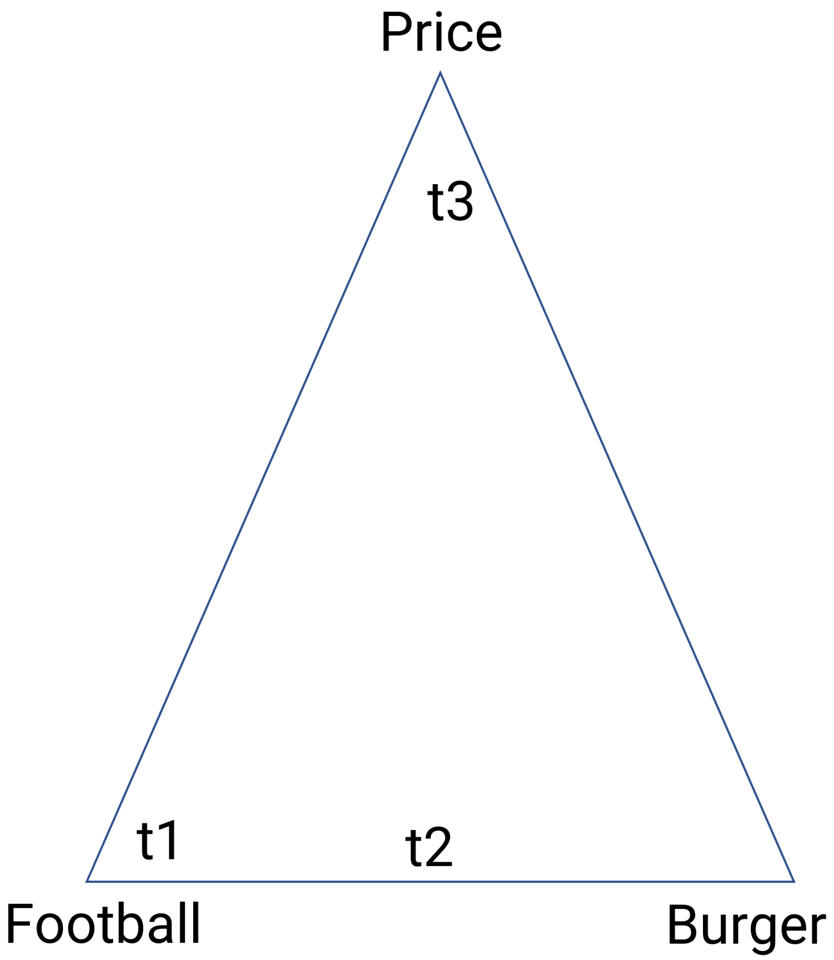


Topics-Words Matrix (Distribution)

For each word, we can calculate the frequency of each word appearing in each topic and hence its probability.



	price	burger	football	…	stadium
Topic 1	0.01	0.02	0.4		0.4



Probability of a document

$$p(w,z,\theta,\phi,|\alpha,\beta) = \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{i=1}^{K} p(\phi_i|\beta) \left(\prod_{t=1}^{N} p\left(z_{j,t}|\theta_j\right) p(w_{j,t}|\phi,z_{j,t}) \right)$$

Probability of document

Dirichlet Distributions Documents-topics (Triangle with topics on corners). M is the number of documents.

The multinomial is calculated based on the Dirichlet Distributions Documents-topics

Multinomial Distributions to calculate the probability of each topic given the Dirichlet Distributions (documents-topics) for document j where z is the topic of word t in document j. N is the number of words in document j.

Probability of a document

$$\underline{p(w, z, \theta, \phi, |\alpha, \beta)} = \prod_{j=1}^{M} \underline{p(\theta_j|\alpha)} \prod_{i=1}^{K} p(\phi_i|\beta) \left(\prod_{t=1}^{N} \underline{p\left(z_{j,t}|\theta_j\right)}\, p(w_{j,t}|\phi, z_{j,t}) \right)$$

Probability of document

Dirichlet Distributions Documents-topics (Triangle with topics on corners). M is the number of documents.

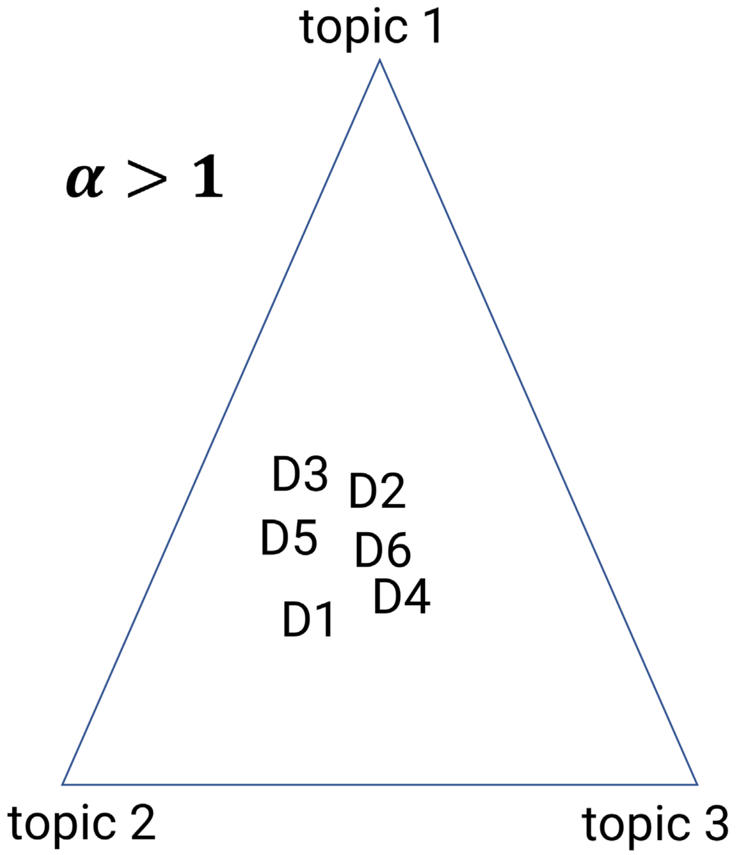
The multinomial is calculated based on the Dirichlet Distributions Documents-topics

Multinomial Distributions to calculate the probability of each topic given the Dirichlet Distributions (documents-topics) for document j where z is the topic of word t in document j. N is the number of words in document j.

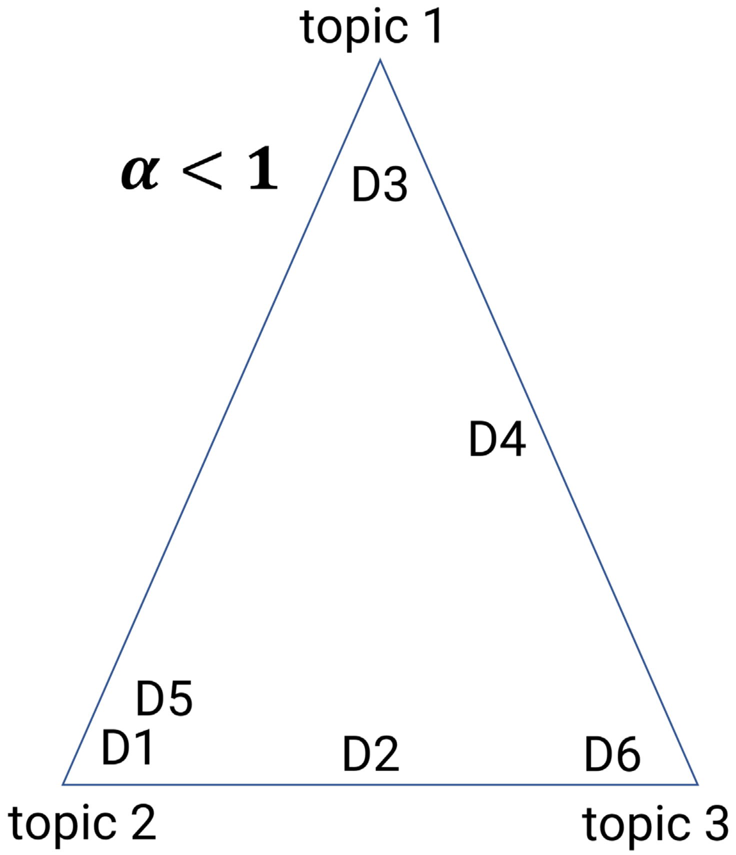
Effects of Dirichlet Parameters α or β

$$\underline{p(w,z,\theta,\phi,|\alpha,\beta)} = \prod_{j=1}^{M} p(\theta_j|\underline{\alpha}) \prod_{i=1}^{K} p(\phi_i|\beta) \left(\prod_{t=1}^{N} p\left(z_{j,t}|\theta_j\right) p(w_{j,t}|\phi,z_{j,t}) \right)$$

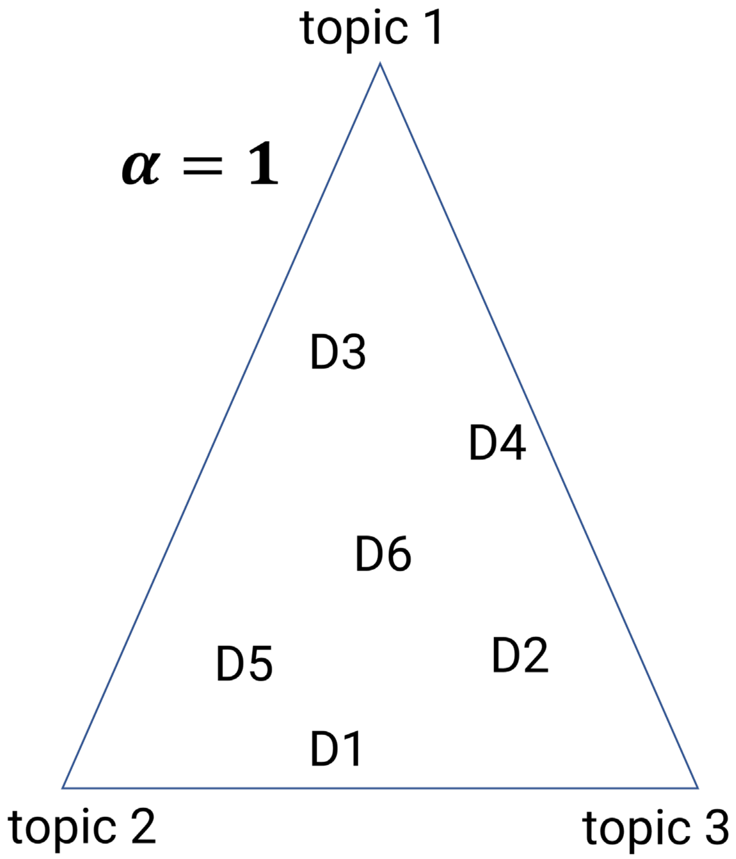
$\boldsymbol{\alpha > 1}$



$\boldsymbol{\alpha < 1}$



$\boldsymbol{\alpha = 1}$



Probability of a Document

$$p(w, z, \theta, \phi, |\alpha, \beta) = \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{i=1}^{K} p(\phi_i|\beta) \left(\prod_{t=1}^{N} p\left(z_{j,t}\middle|\theta_j\right) p(w_{j,t}|\phi, z_{j,t}) \right)$$

Probability of document

Dirichlet Distributions Topics-words (Triangle with words on corners). K is the number of topics.

Multinomial Distributions to calculate the probability of each word given Dirichlet Distributions (topics-words) and based on chosen topics by $p\left(z_{j,t}\middle|\theta_j\right)$ for the document j and word t; where z is the topic of word t in document j. N is the number of words in document j.

Duty of Each Probability

$$\prod_{j=1}^{M} p(\theta_j|\alpha)$$

	t 1	t 2	t 3
d 1	0.1	0.8	0.1
d2	…	…	…

$$\prod_{i=1}^{K} p(\phi_i|\beta)$$

	price	burger	football	…	stadium
t 1	0.01	0.02	0.4	…	0.4
t 2	0.01	0.5	0.3	…	0.1
t 3	0.9	0.02	0.01	…	0.001

$$\prod_{t=1}^{N} p(z_{j,t}|\theta_j)$$

Let's say Document 1 has 5 words; now we need to use the Documents-topics Dirichlet Distribution to randomly assign a topic to each word. For example, the chance of each word being assigned as t2 in document 1 is 80% and is higher than other topics.

$$p(w_{j,t}|\phi, z_{j,t})$$

Now by knowing the topics of all 5 words in document using the $p(z_{j,t}|\theta_j)$, we can use the Topics-words Dirichlet Distribution to randomly re-assign a word to that topic. For example, if word 1 (price) was randomly assigned to topic 1 in $p(z_{j,t}|\theta_j)$, and if we were to assign a new word to this topic, the chance of ”stadium” or “football” would be the highest among others.

Duty of Each Probability

$$\prod_{j=1}^{M} p(\boldsymbol{\theta}_j|\boldsymbol{\alpha})$$

	t 1	t 2	t 3
d 1	0.1	0.8	0.1
d2	…	…	…

$$\prod_{i=1}^{K} p(\boldsymbol{\phi}_i|\boldsymbol{\beta})$$

	price	burger	football	…	stadium
t 1	0.01	0.02	0.4	…	0.4
t 2	0.01	0.5	0.3	…	0.1
t 3	0.9	0.02	0.01	…	0.001

$$\prod_{t=1}^{N} p(\boldsymbol{z}_{j,t}|\boldsymbol{\theta}_j)$$

Let’s say Document 1 has 5 words; now we need to use the Documents-topics Dirichlet Distribution to randomly assign a topic to each word. For example, the chance of each word being assigned as t2 in document 1 is 80% and is higher than other topics.

$$p(\boldsymbol{w}_{j,t}|\boldsymbol{\phi}, \boldsymbol{z}_{j,t})$$

Now by knowing the topics of all 5 words in document using the $p(\boldsymbol{z}_{j,t}|\boldsymbol{\theta}_j)$, we can use the Topics-words Dirichlet Distribution to randomly re-assign a word to that topic. For example, if word 1 (price) was randomly assigned to topic 1 in $p(\boldsymbol{z}_{j,t}|\boldsymbol{\theta}_j)$, and if we were to assign a new word to this topic, the chance of ”stadium” or “football” would be the highest among others.

Duty of Each Probability

Topics probability	Words Probability	Randomly Generating topics for words	Randomly assigning words to selected topics
$\prod_{j=1}^{M} p(\theta_j \vert \alpha)$	$\prod_{i=1}^{K} p(\phi_i \vert \beta)$	$\prod_{t=1}^{N} p(z_{j,t} \vert \theta_j)$	$p(w_{j,t} \vert \phi, z_{j,t})$

- We can generate M documents based on the predefined values for Dirichlet parameters α and β
- We can change the Dirichlet parameters α and β and generate another M documents
- We will check what sets of Dirichlet parameters α and β would re-generate documents closest to the original ones.
- To maximize the probability of generated document to be similar to the original one, we use Gibbs Sampling.
- This maximization tries to cluster similar words and cluster similar documents for a specific topic together

Duty of Each Probability

Topics probability	Words Probability	Randomly Generating topics for words	Randomly assigning words to selected topics
$\prod_{j=1}^{M} p(\theta_j \vert \alpha)$	$\prod_{i=1}^{K} p(\phi_i \vert \beta)$	$\prod_{t=1}^{N} p(z_{j,t} \vert \theta_j)$	$p(w_{j,t} \vert \phi, z_{j,t})$

- We can generate M documents based on the predefined values for Dirichlet parameters α and β
- We can change the Dirichlet parameters α and β and generate another M documents
- We will check what sets of Dirichlet parameters α and β would re-generate documents closest to the original ones.
- To maximize the probability of generated document to be similar to the original one, we use Gibbs Sampling.
- This maximization tries to cluster similar words and cluster similar documents for a specific topic together

Summary

- We learned an unsupervised technique (topic modeling)
- LDA
- Obtain topics of words and documents