Topic Modeling

What is Topic Modeling?

Anything that tries to answer…

“What is the underlying topic(s) that this document is about?”

Let’s motivate with an example…

3D → 2D Reduction with text data (bag of words model)

“I love my pet rabbit.”
“That dish yesterday was amazing.”
“She cooked the best rabbit dish ever.”
“I gave leftovers of that dish to my pet, mr. rabbit”
“Rabbits make messy pets.”
“My rabbit growls when I pet her.”
“He has five rabbits.”
“I had this weird dish with fried rabbit.”
“That’s my pet rabbit’s favorite dish.”
…

3D → 2D Reduction with text data (bag of words model)

“I love my pet rabbit.”
“That dish yesterday was amazing.”
“She cooked the best rabbit dish ever.”
“I gave leftovers of that dish to my pet, mr. rabbit”
“Rabbits make messy pets.”
“My rabbit growls when I pet her.”
“He has five rabbits.”
“I had this weird dish with fried rabbit.”
“That’s my pet rabbit’s favorite dish.”
…

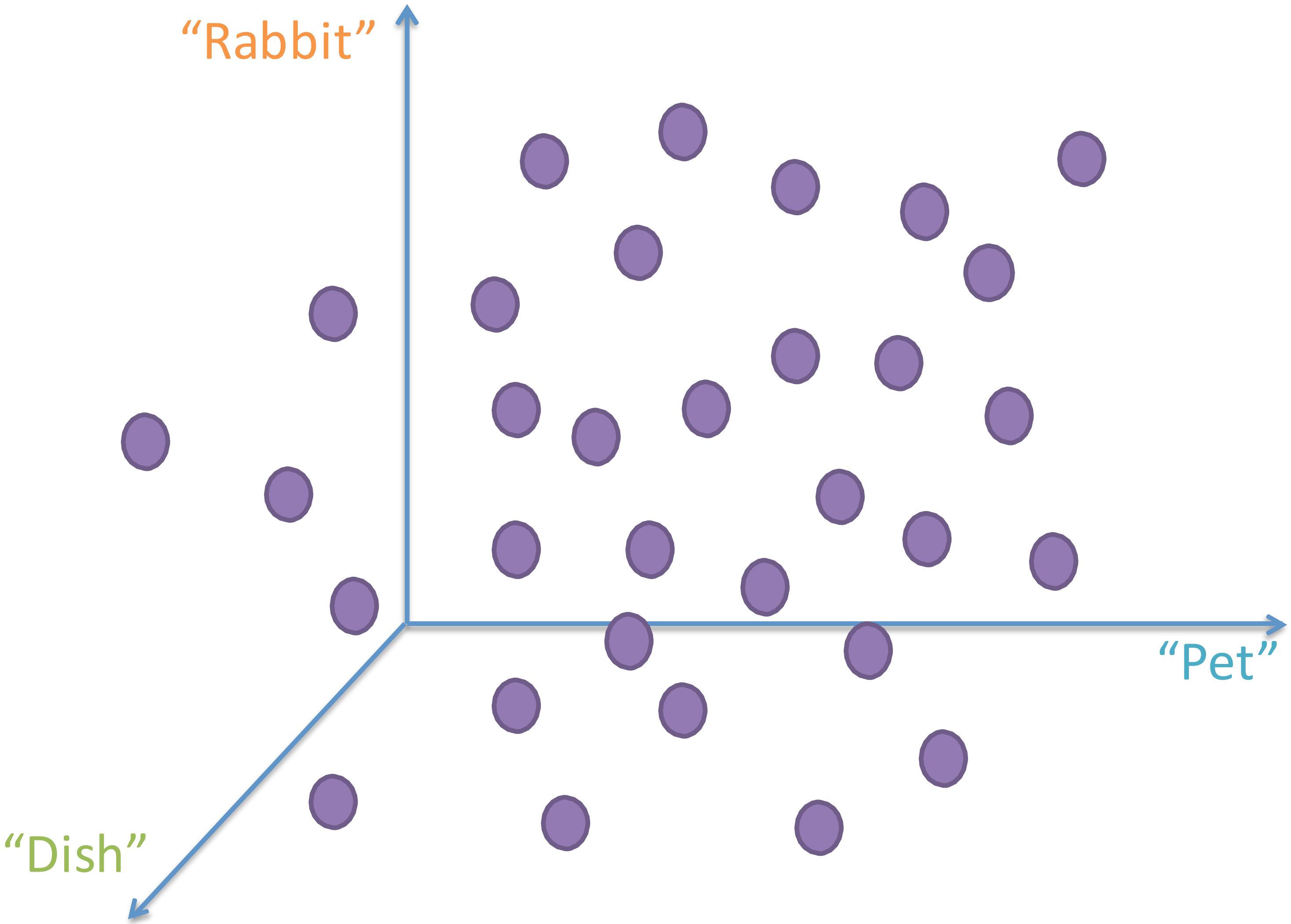
Remove stop words, only keep nouns, end up with 3 features: “rabbit”, “pet”, “dish”

3D → 2D Reduction with text data (bag of words model)

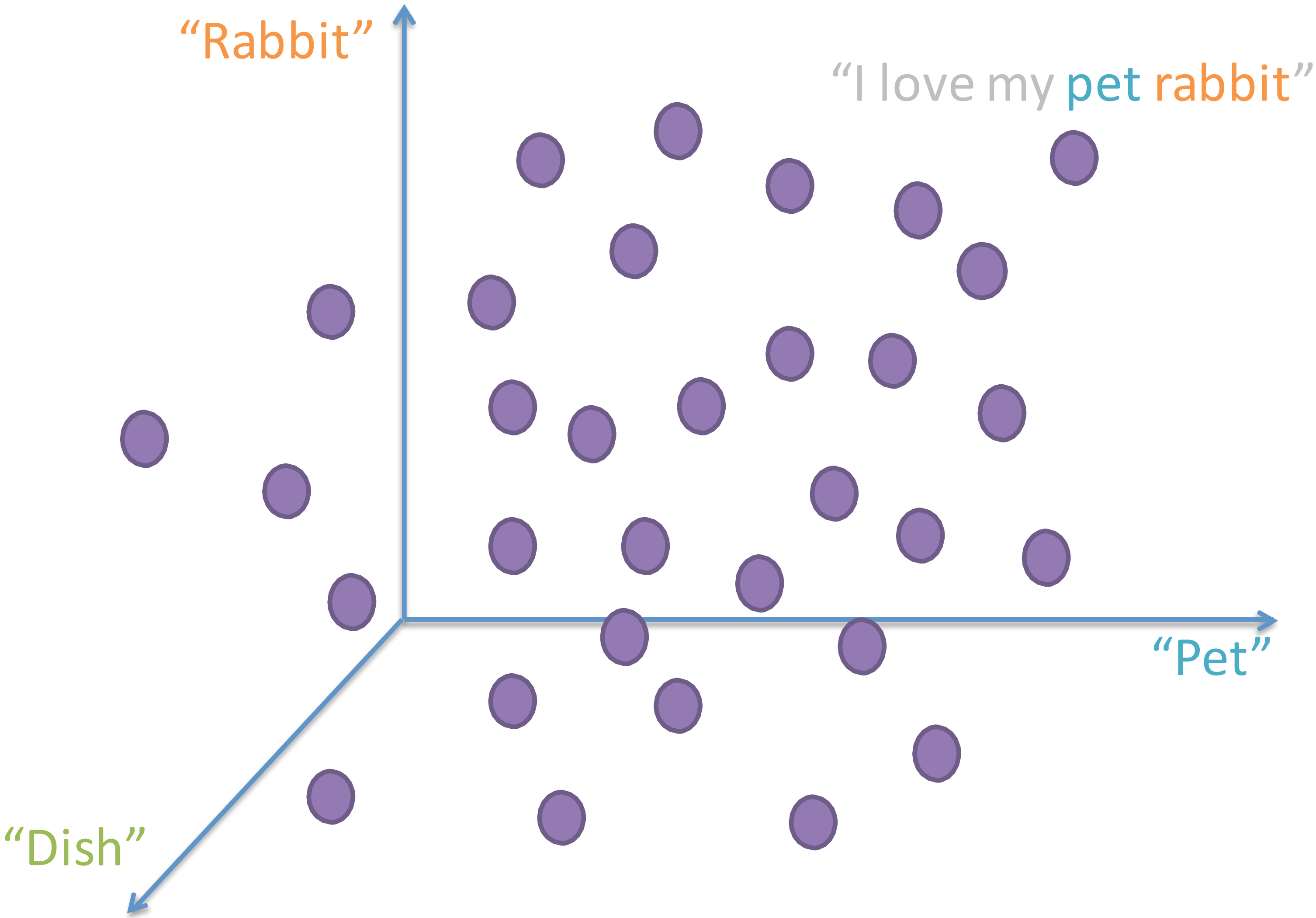
“I love my pet rabbit.”
“That dish yesterday was amazing.”
“She cooked the best rabbit dish ever.”
“I gave leHovers of that dish to my pet, mr. rabbit”
“Rabbits make messy pets.”
“My rabbit growls when I pet her.”
“He has five rabbits.”
“I had this weird dish with fried rabbit.”
“That’s my pet rabbit’s favorite dish.”
…

Remove stop words, only keep nouns, end up with 3 features: “rabbit”, “pet”, “dish”

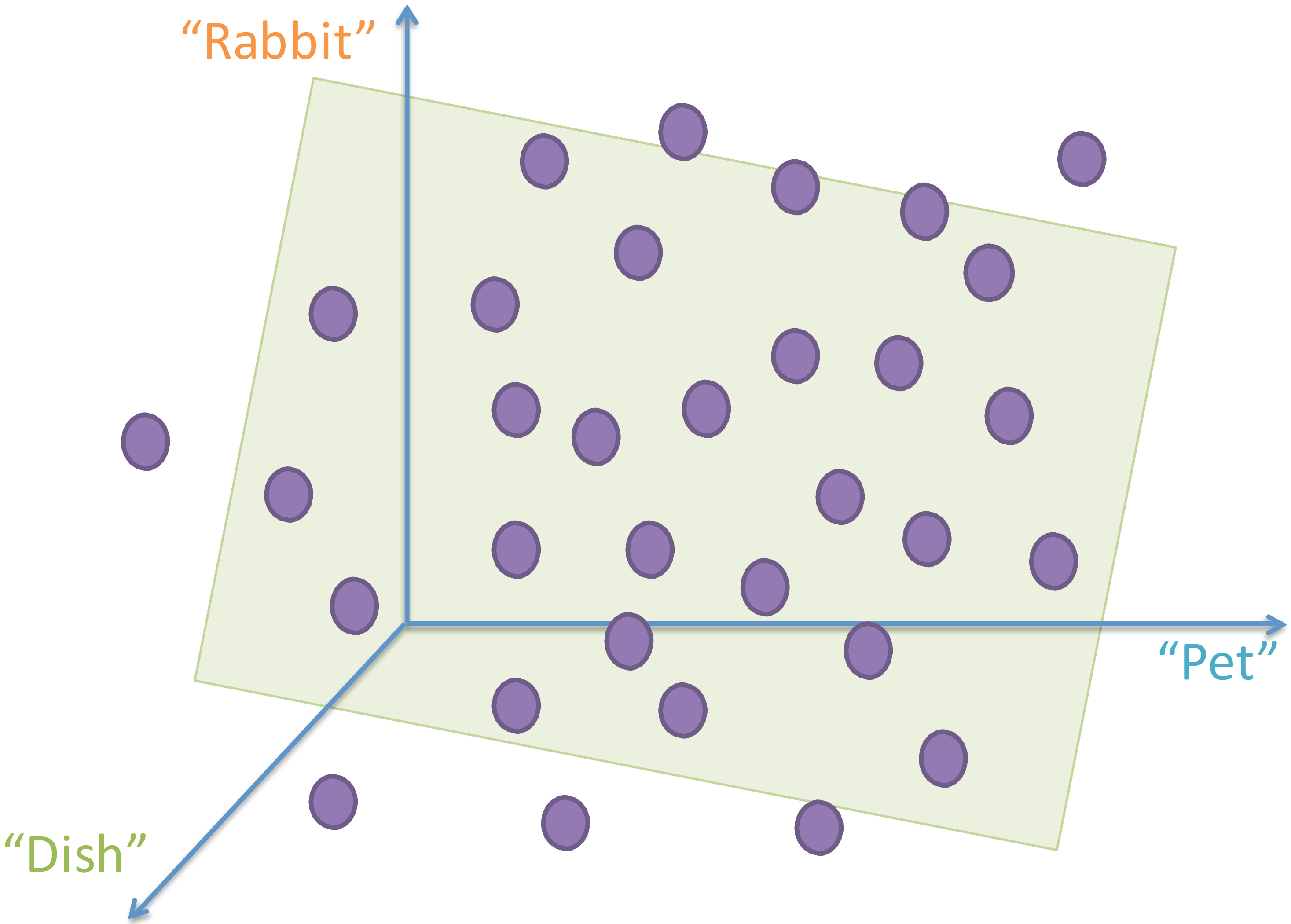
3D → 2D Reduction with text data (bag of words model)



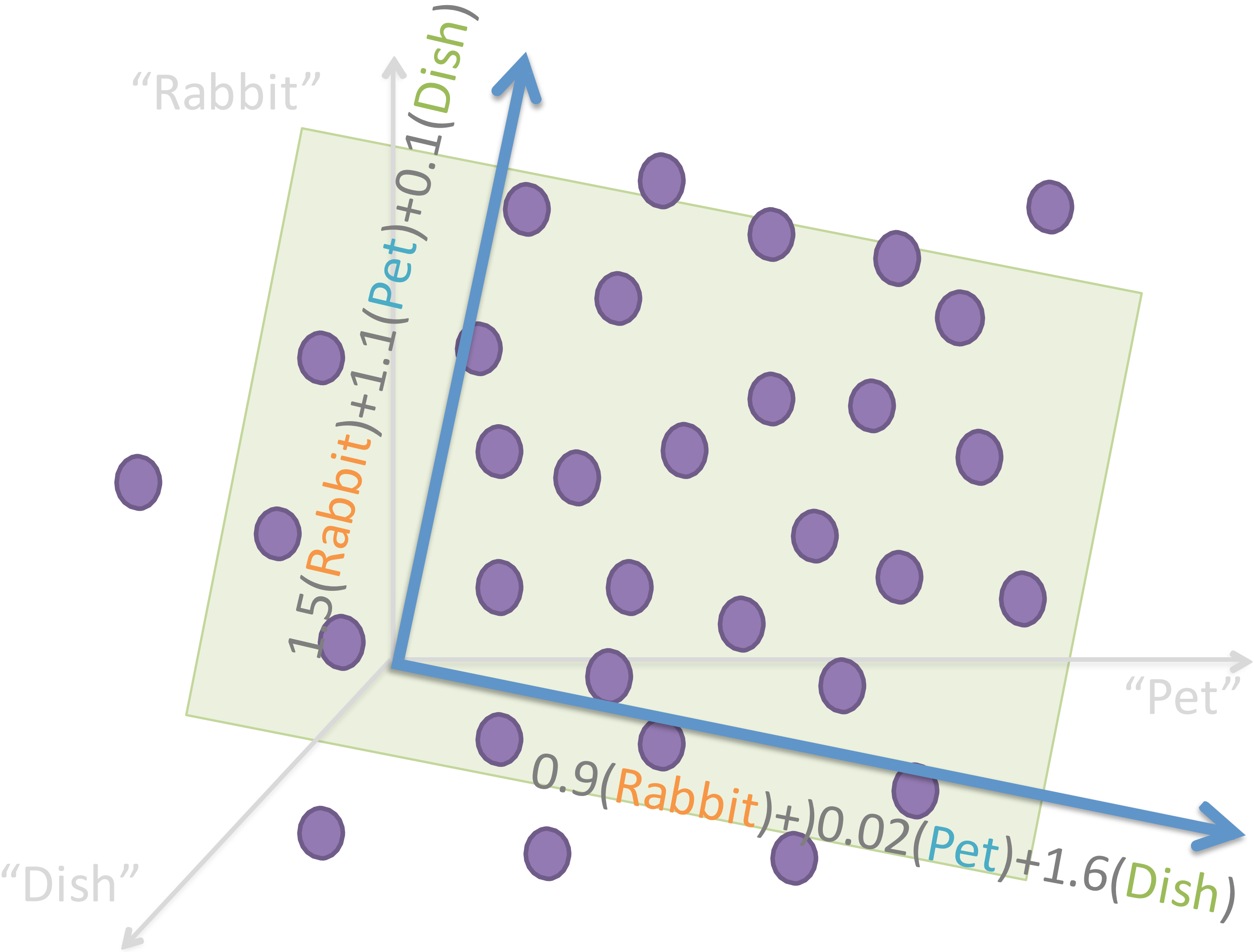
3D → 2D Reduction with text data (bag of words model)



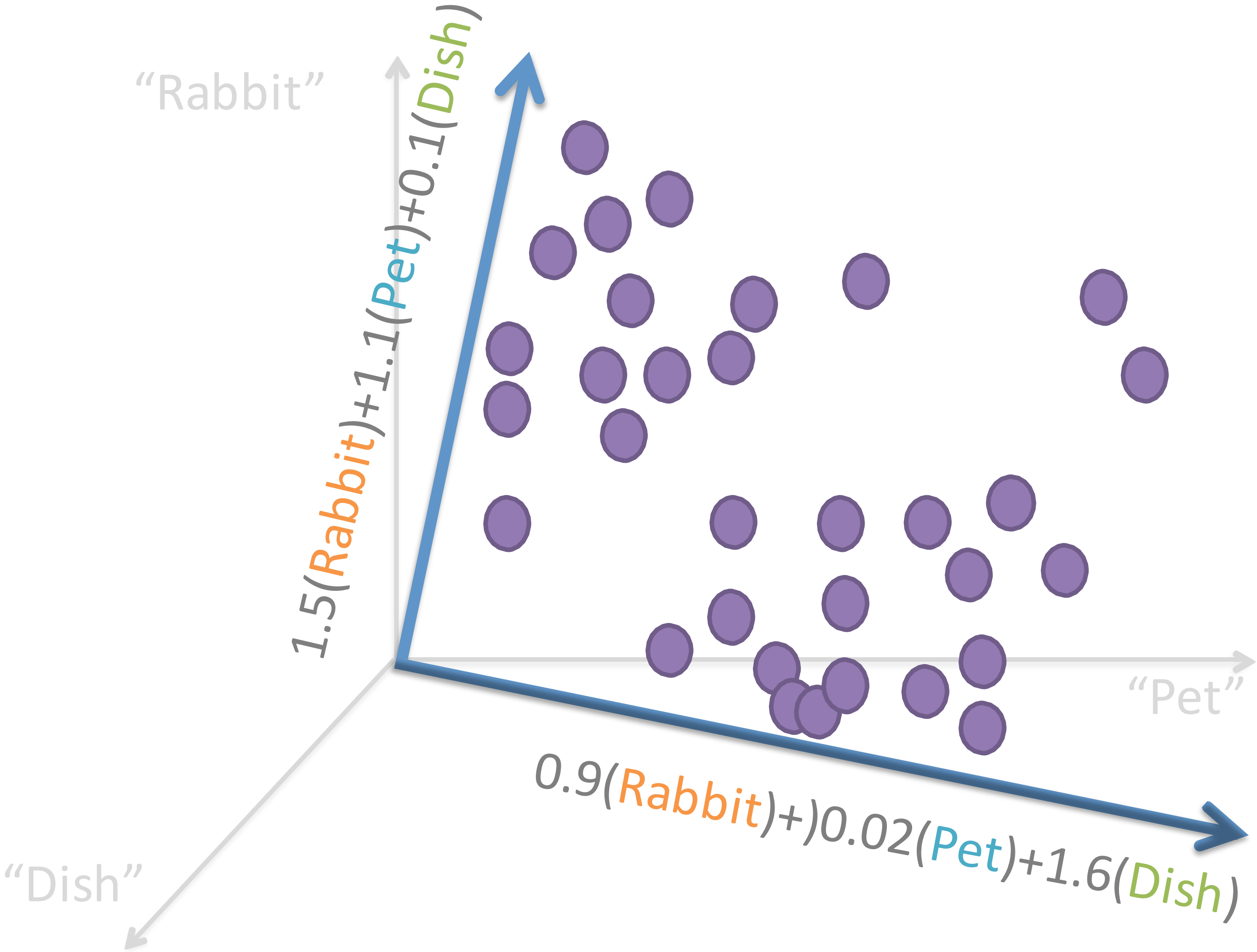
3D → 2D Feature Extraction



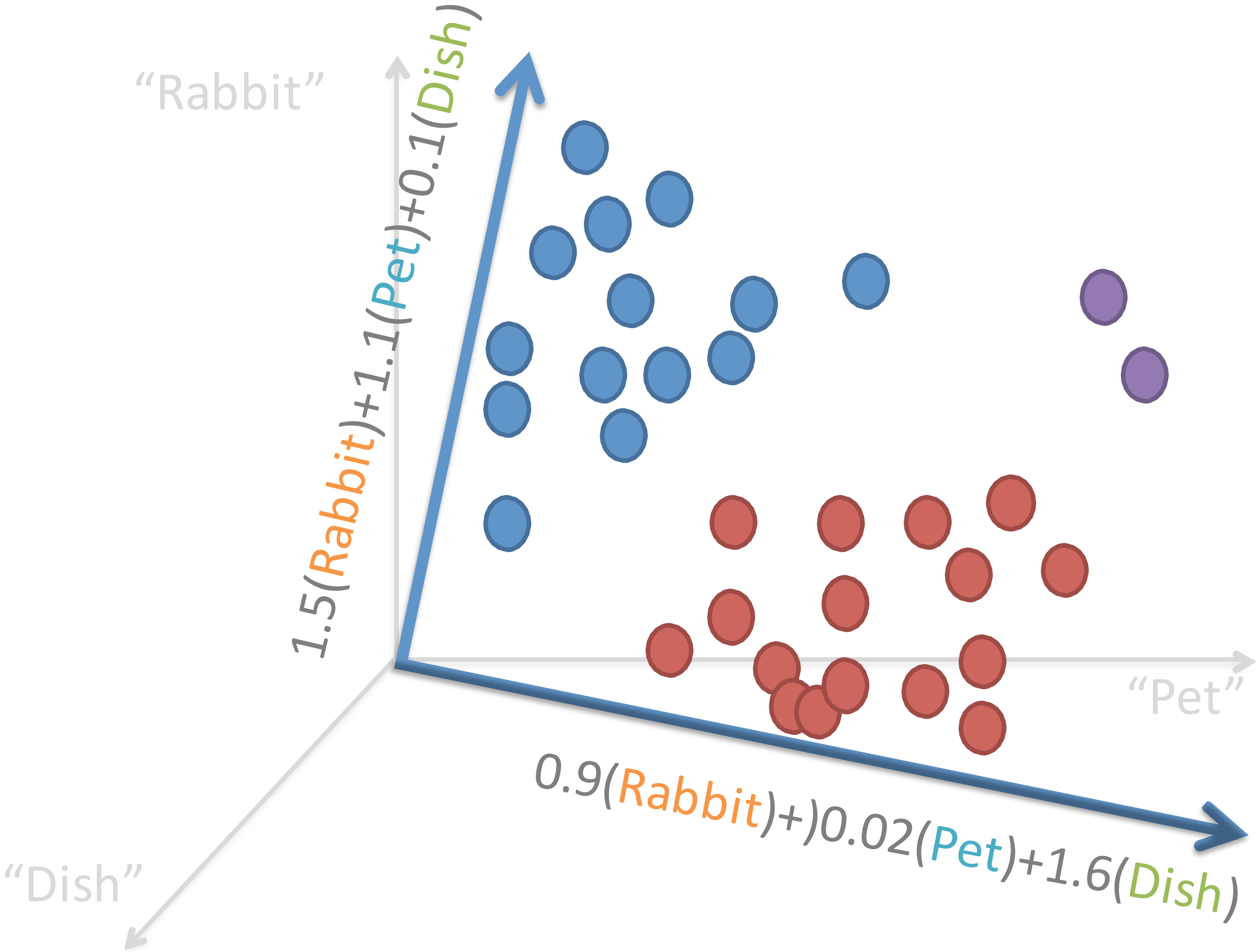
3D → 2D Feature Extraction


"Rabbit"
"Pet"
"Dish"
1.5(Rabbit)+1.1(Pet)+0.1(Dish)
0.9(Rabbit)+)0.02(Pet)+1.6(Dish)

3D → 2D Feature Extraction


“Rabbit”
1.5(Rabbit)+1.1(Pet)+0.1(Dish)
“Pet”
0.9(Rabbit)+)0.02(Pet)+1.6(Dish)
“Dish”

Clustering is easier on this space



What are the clusters?

"I love my pet rabbit."
"Rabbits make messy pets."
"My rabbit growls when I pet her."
"He has five rabbits."

"That dish yesterday was amazing."
"She cooked the best rabbit dish ever."
"I had this weird dish with fried rabbit."

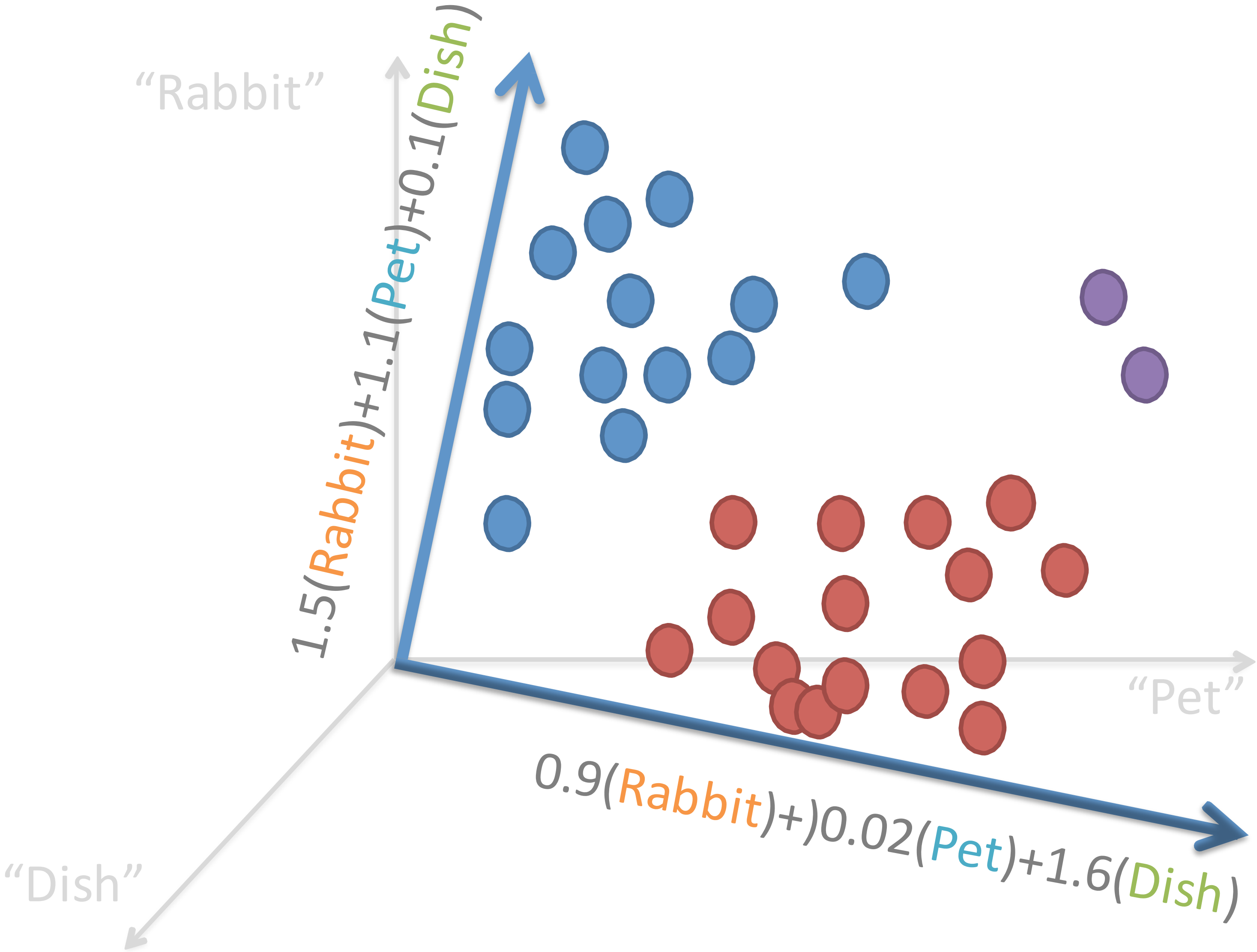
"I gave leftovers of that dish to my pet, Mr. Rabbit"
"That's my pet rabbit's favorite dish."

Axis 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish)
Axis 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish)

"I love my pet rabbit."
"Rabbits make messy pets."
"My rabbit growls when I pet her."
"He has five rabbits."

"That dish yesterday was amazing."
"She cooked the best rabbit dish ever."
"I had this weird dish with fried rabbit."

"I gave leftovers of that dish to my pet, Mr. Rabbit"
"That's my pet rabbit's favorite dish."


“Rabbit”
1.5(Rabbit)+1.1(Pet)+0.1(Dish)
“Pet”
0.9(Rabbit)+)0.02(Pet)+1.6(Dish)
“Dish”

Axis 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish)
Axis 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish)

"I love my **pet rabbit**."
"**Rabbits** make messy **pets**."
"My **rabbit** growls when I **pet** her."
"He has five **rabbits**."

"That **dish** yesterday was amazing."
"She cooked the best **rabbit dish** ever."
"I had this weird **dish** with fried **rabbit**."

"I gave leftovers of that **dish** to my **pet**, mr. **rabbit**"
"That's my **pet rabbit**'s favorite **dish**."

Axis 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish)
Axis 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish)

Axis1: High
Axis2: Low

"I love my **pet rabbit**."
"**Rabbits** make messy **pets**."
"My **rabbit** growls when I **pet** her."
"He has five **rabbits**."

"That **dish** yesterday was amazing."
"She cooked the best **rabbit dish** ever."
"I had this weird **dish** with fried **rabbit**."

"I gave leftovers of that **dish** to my **pet**, mr. **rabbit**"
"That's my **pet rabbit**'s favorite **dish**."

Axis 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish)
Axis 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish)

Axis1: High
Axis2: Low

“I love my **pet rabbit**.”
“**Rabbits** make messy **pets**.”
“My **rabbit** growls when I **pet** her.”
“He has five **rabbits**.”

Axis1: Low
Axis2: High

“That **dish** yesterday was amazing.”
“She cooked the best **rabbit dish** ever.”
“I had this weird **dish** with fried **rabbit**.”

“I gave leftovers of that **dish** to my **pet**, mr. **rabbit**”
“That’s my **pet rabbit**’s favorite **dish**.”

Axis 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish)

Axis 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish)

Axis1: High

Axis2: Low

"I love my **pet rabbit**."

"**Rabbits** make messy **pets**."

"My **rabbit** growls when I **pet** her."

"He has five **rabbits**."

Axis1: Low

Axis2: High

"That **dish** yesterday was amazing."

"She cooked the best **rabbit dish** ever."

"I had this weird **dish** with fried **rabbit**."

Axis1: High

Axis2: High

"I gave leftovers of that **dish** to my **pet**, mr. **rabbit**"

"That's my **pet rabbit**'s favorite **dish**."

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

Topic1: High
Topic2: Low

“I love my **pet rabbit**.”
“**Rabbits** make messy **pets**.”
“My **rabbit** growls when I **pet** her.”
“He has five **rabbits**.”

Topic1: Low
Topic2: High

“That **dish** yesterday was amazing.”
“She cooked the best **rabbit dish** ever.”
“I had this weird **dish** with fried **rabbit**.”

Topic1: High
Topic2: High

“I gave leftovers of that **dish** to my **pet**, mr. **rabbit**”
“That’s my **pet rabbit**’s favorite **dish**.”

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

T1	T2	
87%	13%	“I love my **pet rabbit**.”
88%	12%	“**Rabbits** make messy **pets**.”
80%	20%	“My **rabbit** growls when I **pet** her.”
66%	34%	“He has five **rabbits**.”
2%	98%	“That **dish** yesterday was amazing.”
16%	84%	“She cooked the best **rabbit dish** ever.”
15%	85%	“I had this weird **dish** with fried **rabbit**.”
47%	53%	“I gave leftovers of that **dish** to my **pet**, mr. **rabbit**”
42%	58%	“That’s my **pet rabbit**’s favorite **dish**.”

Topics are not (hard) clusters. A document does not belong to a single topic. Each topic is present in the document up to a certain degree. For each doc, we have a distribution over topics.

T1	T2	
87%	13%	“I love my **pet rabbit**.”
88%	12%	“**Rabbits** make messy **pets**.”
80%	20%	“My **rabbit** growls when I **pet** her.”
66%	34%	“He has five **rabbits**.”
2%	98%	“That **dish** yesterday was amazing.”
16%	84%	“She cooked the best **rabbit dish** ever.”
15%	85%	“I had this weird **dish** with fried **rabbit**.”
47%	53%	“I gave leftovers of that **dish** to my **pet**, mr. **rabbit**”
42%	58%	“That’s my **pet rabbit**’s favorite **dish**.”

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

What is a topic?

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

What is a topic?

When writing about a specific topic (like pet rabbits), we use some words more often than others.

Words like “pet”, “rabbit”, “lettuce”, “cage”, “fluffy”, etc. are more likely to appear, words like “dish”, “transmission”, “opaque”, “affair” are less likely to appear.

A topic can be thought of as a
Probability distribution over all possible words

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

What is a topic?

Probability distribution over all possible words

Word	Prob in [Pet Rabbits]	Prob in [Food]
pet	2.3x10^{--7}	1.2x10^{--10}
rabbit	7.9x10^{--7}	3.4x10^{--8}
dish	6.8x10^{--11}	4.5x10^{--7}
car	3.1x10^{--12}	1.8x10^{--12}
hello	8.3x10^{--9}	1.4x10^{--9}
the	7.4x10^{--4}	7.3x10^{--4}
love	5.4x10^{--8}	3.9x10^{--8}
affair	3.0x10^{--13}	2.1x10^{--13}
delicious	9.1x10^{--9}	9.8x10^{--8}

…

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

What is a topic?

Probability distribution over all possible words

Word	Prob in [Pet Rabbits]	Prob in [Food]
pet	$2.3x10^{--7}$	$1.2x10^{--10}$
rabbit	$7.9x10^{--7}$	$3.4x10^{--8}$
dish	$6.8x10^{--11}$	$4.5x10^{--7}$
car	$3.1x10^{--12}$	$1.8x10^{--12}$
hello	$8.3x10^{--9}$	$1.4x10^{--9}$
the	$7.4x10^{--4}$	$7.3x10^{--4}$
love	$5.4x10^{--8}$	$3.9x10^{--8}$
affair	$3.0x10^{--13}$	$2.1x10^{--13}$
delicious	$9.1x10^{--9}$	$9.8x10^{--8}$

…

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

What is a topic?

Probability distribution over all possible words

Word	Prob in [Pet Rabbits]	Prob in [Food]
pet	$2.3x10^{--7}$	$1.2x10^{--10}$
rabbit	$7.9x10^{--7}$	$3.4x10^{--8}$
dish	$6.8x10^{--11}$	$4.5x10^{--7}$
car	$3.1x10^{--12}$	$1.8x10^{--12}$
hello	$8.3x10^{--9}$	$1.4x10^{--9}$
the	$7.4x10^{--4}$	$7.3x10^{--4}$
love	$5.4x10^{--8}$	$3.9x10^{--8}$
affair	$3.0x10^{--13}$	$2.1x10^{--13}$
delicious	$9.1x10^{--9}$	$9.8x10^{--8}$

…

TOPIC 1: 1.5(Rabbit) + 1.1 (Pet) + 0.1(Dish) ← Pet rabbits, pets

TOPIC 2: 0.9(Rabbit) + 0.02(Pet) + 1.6(Dish) ← Food, rabbit dishes

What is a topic?

Probability distribution over all possible words

Word	Prob in [Pet Rabbits]	Prob in [Food]
pet	2.3x10^{--7}	1.2x10^{--10}
rabbit	7.9x10^{--7}	3.4x10^{--8}
dish	6.8x10^{--11}	4.5x10^{--7}
car	3.1x10^{--12}	1.8x10^{--12}
hello	8.3x10^{--9}	1.4x10^{--9}
the	7.4x10^{--4}	7.3x10^{--4}
love	5.4x10^{--8}	3.9x10^{--8}
affair	3.0x10^{--13}	2.1x10^{--13}
delicious	9.1x10^{--9}	9.8x10^{--8}

…

Topic Modeling

Let's use an algorithm specifically developed to find topics.

Topic Modeling: LDA

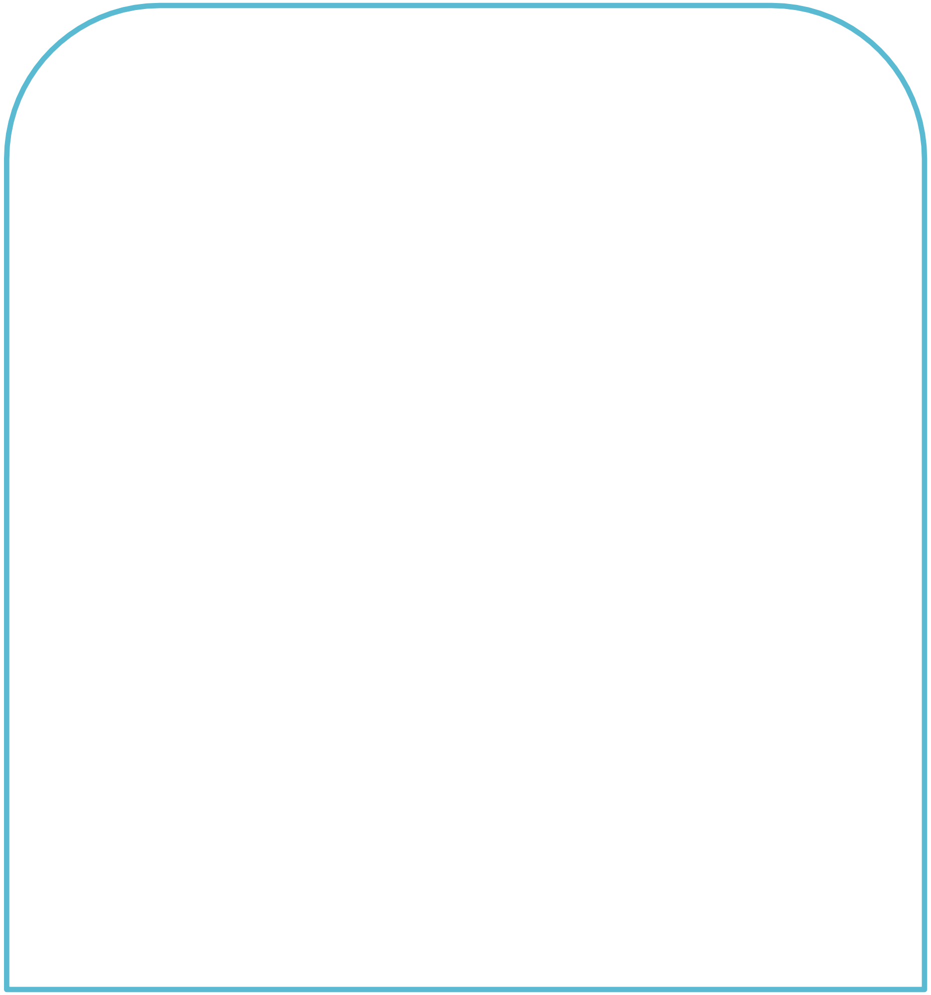
Let’s use an algorithm specifically developed to find topics.

Model the process of writing

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

Topic Modeling: LDA

Let's use an algorithm specifically developed to find topics.

Model the process of writing

Empty page: I'll write a document.

First, I'll decide what topics to write on. Choose the topic distribution.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

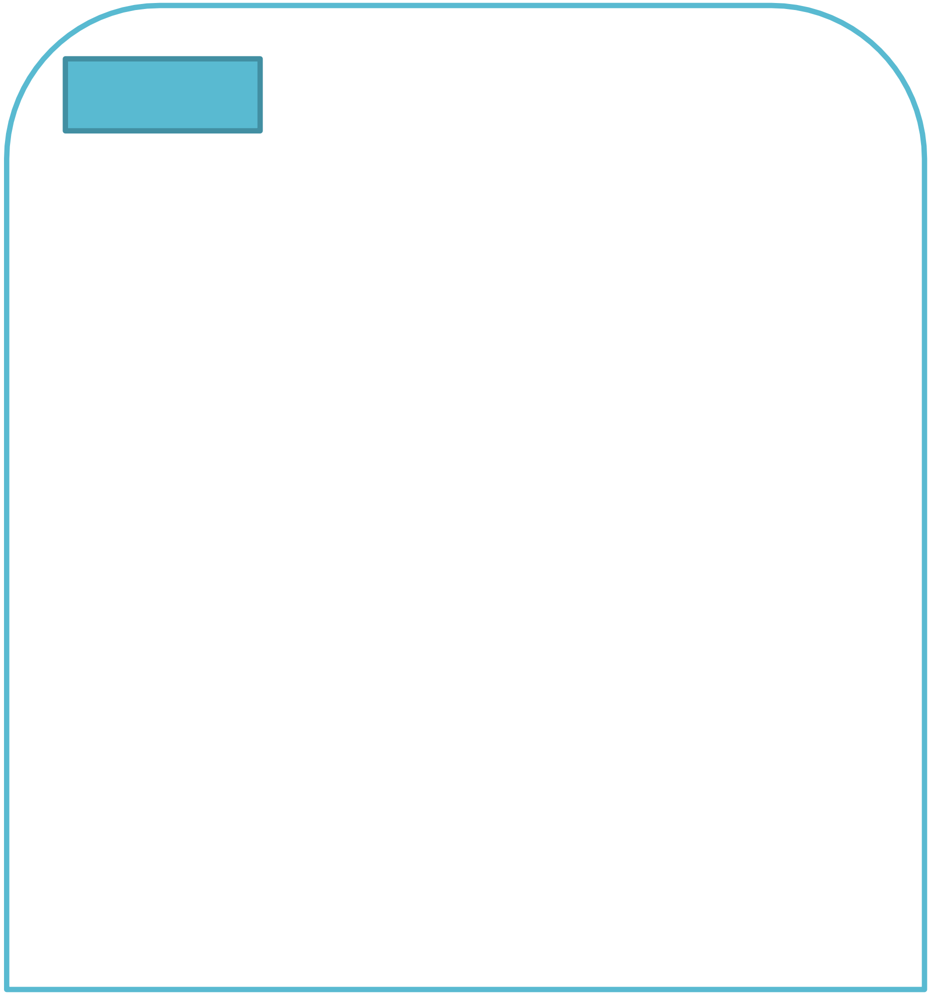
Sex:2%, Drugs:33%, Rock’n Roll:65%

Ok. I’ll write the document word by word (bag of words). First word!

Topic Modeling: LDA

Let's use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I'll write a document.

First, I'll decide what topics to write on. Choose the topic distribution.

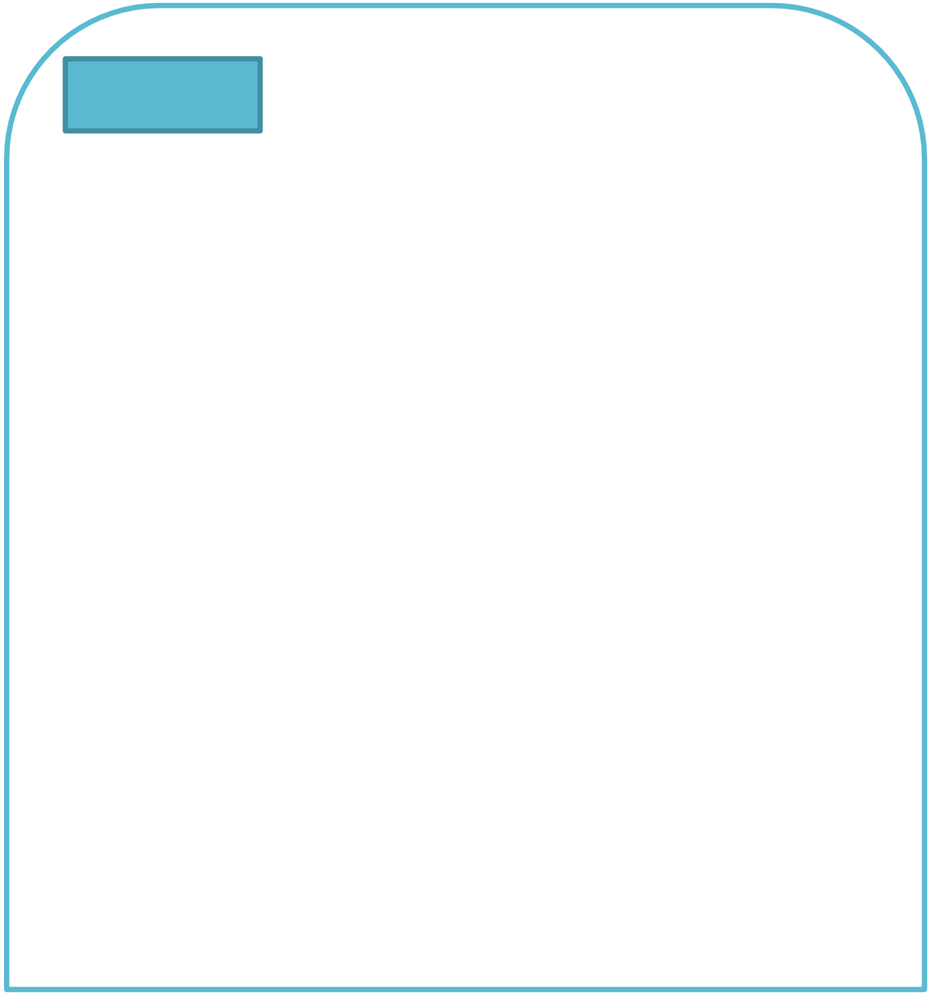
Sex:2%, Drugs:33%, Rock'n Roll:65%

Ok. I'll write the document word by word (bag of words). First word!

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

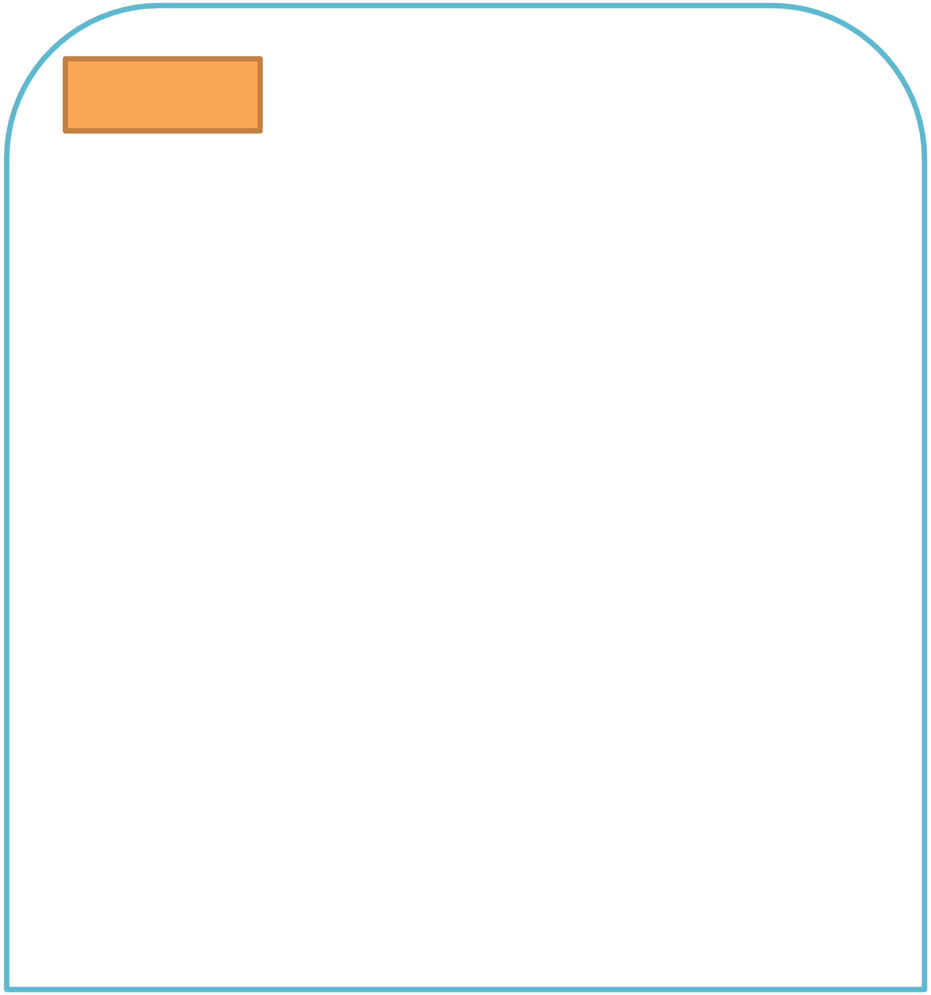
First, I’ll decide what topics to write on. Choose the topic distribution.
Sex:2%, Drugs:33%, Rock’n Roll:65%

Ok. I’ll write the document word by word (bag of words). First word! Choose which topic this word will be about. Roll the dice, pick randomly from the topic distribution for the doc.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

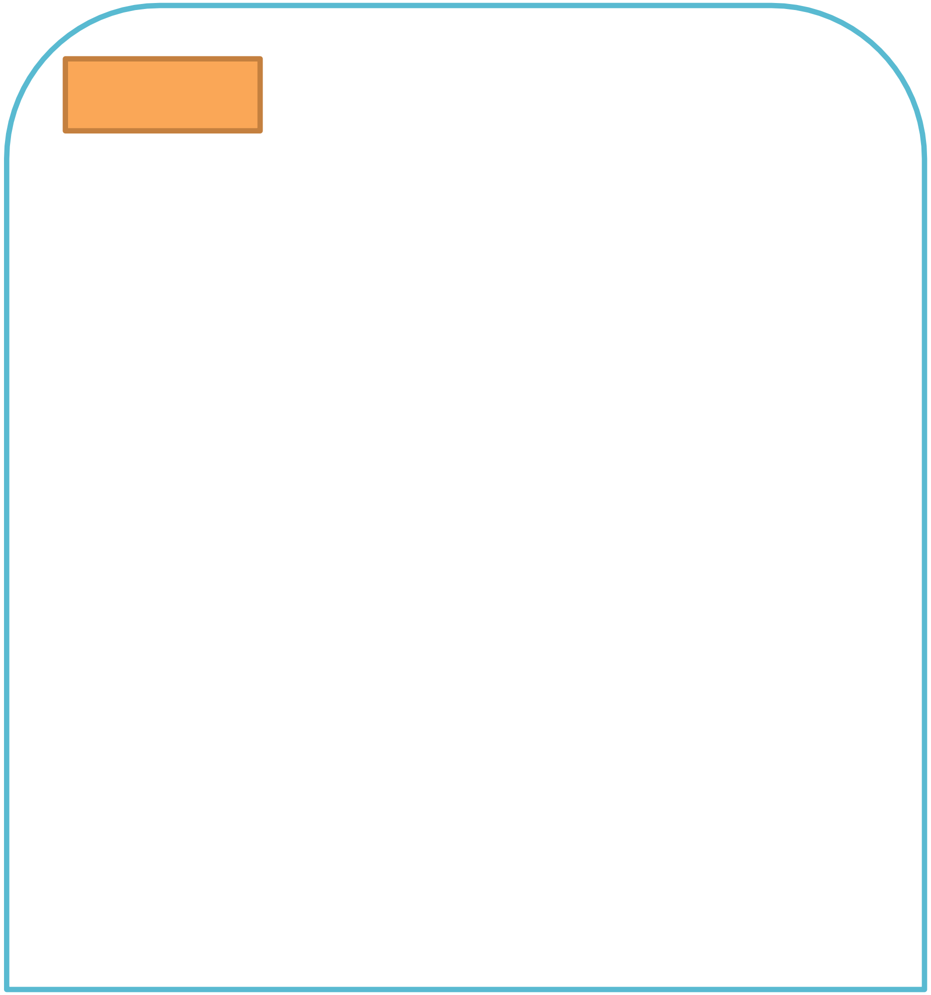
Sex:2%, Drugs:33%, Rock’n Roll:65%

Ok. I’ll write the document word by word (bag of words). First word! Choose which topic this word will be about. Roll the dice, pick randomly from the topic distribution for the doc.

Topic Modeling: LDA

Let's use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I'll write a document.

First, I'll decide what topics to write on. Choose the topic distribution.
Sex:2%, Drugs:33%, Rock'n Roll:65%

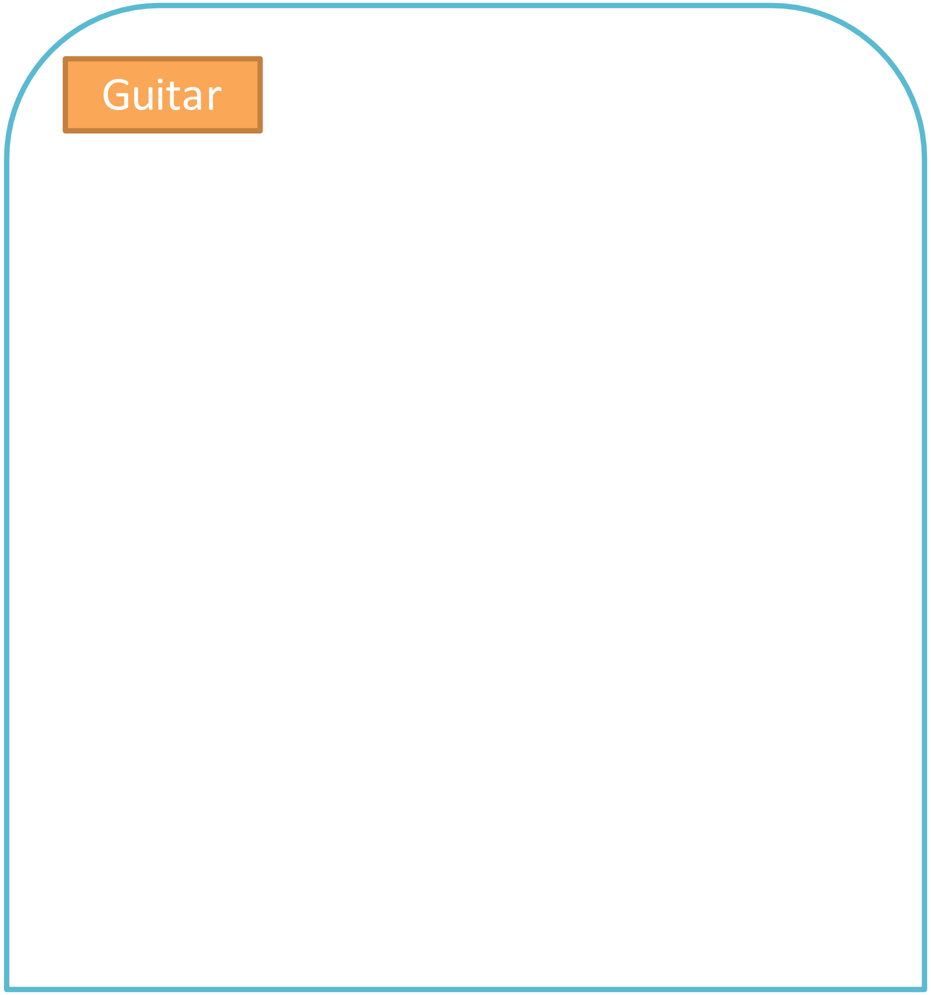
Ok. I'll write the document word by word (bag of words). First word! Choose which topic this word will be about. Roll the dice, pick randomly from the topic distribution for the doc.

A Rock'n Roll word. Randomly pick a word according to the prob. distribution of the Rock'n Roll topic.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

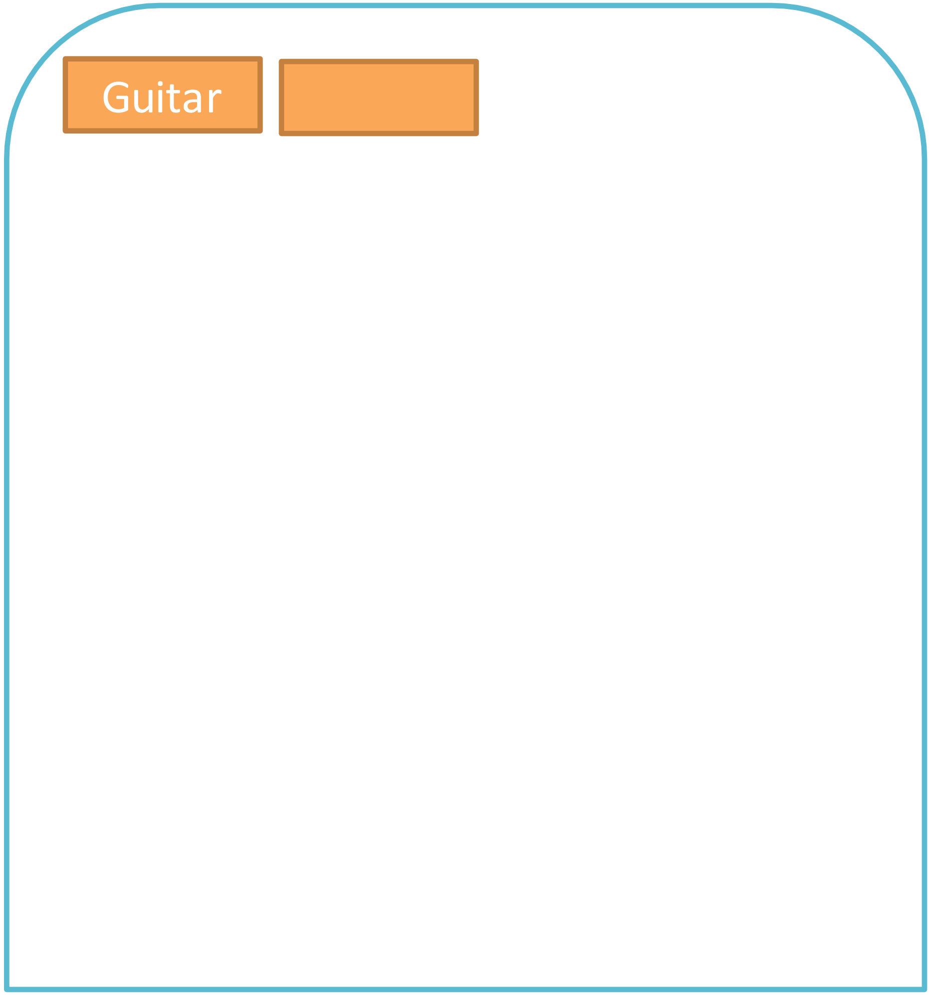
Ok. I’ll write the document word by word (bag of words). First word! Choose which topic this word will be about. Roll the dice, pick randomly from the topic distribution for the doc.

A Rock’n Roll word. Randomly pick a word according to the prob. distribution of the Rock’n Roll topic.

Topic Modeling: LDA

Let's use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I'll write a document.

First, I'll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock'n Roll:65%

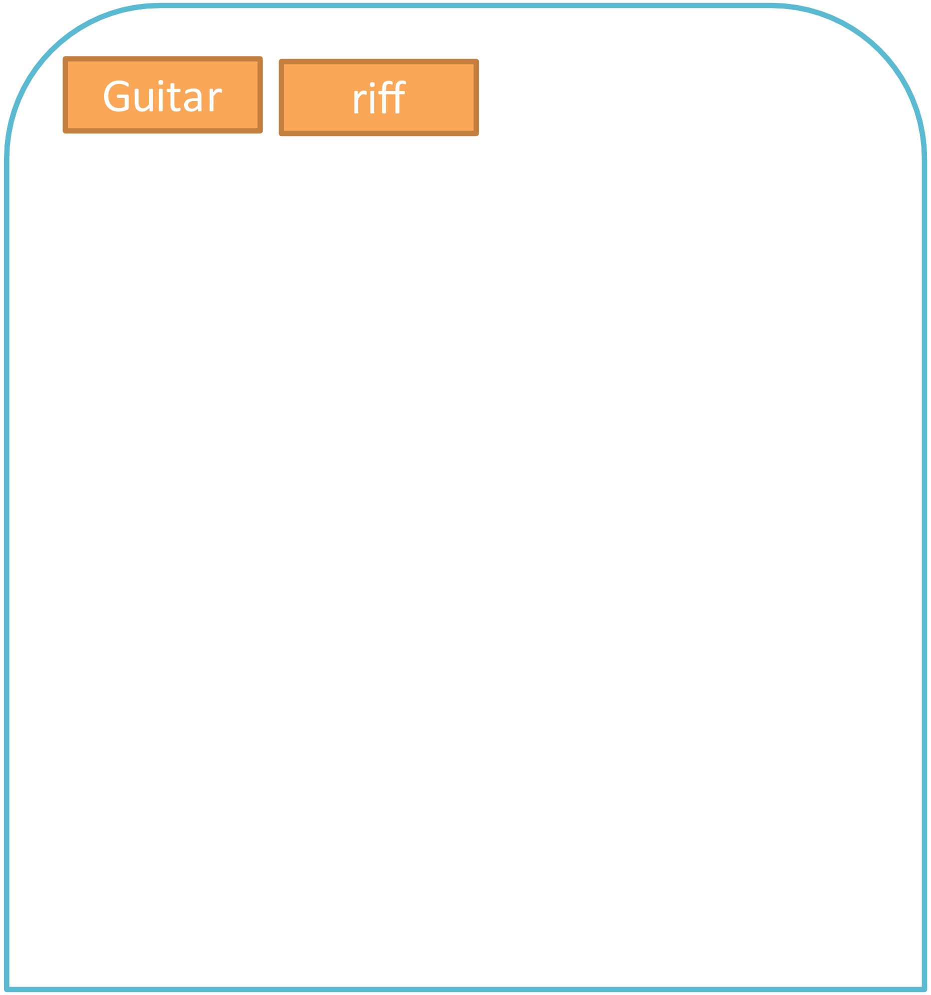
Choose next word's topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

Choose next word’s topic.

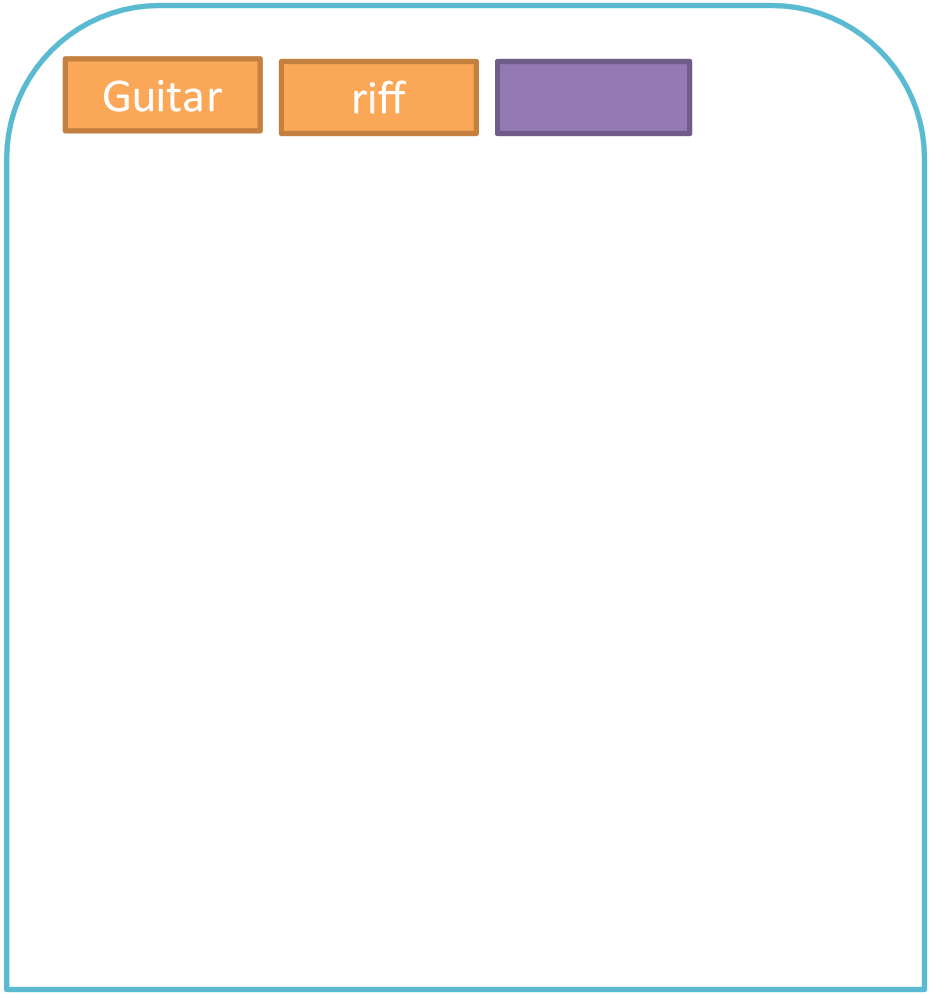
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar

riff

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

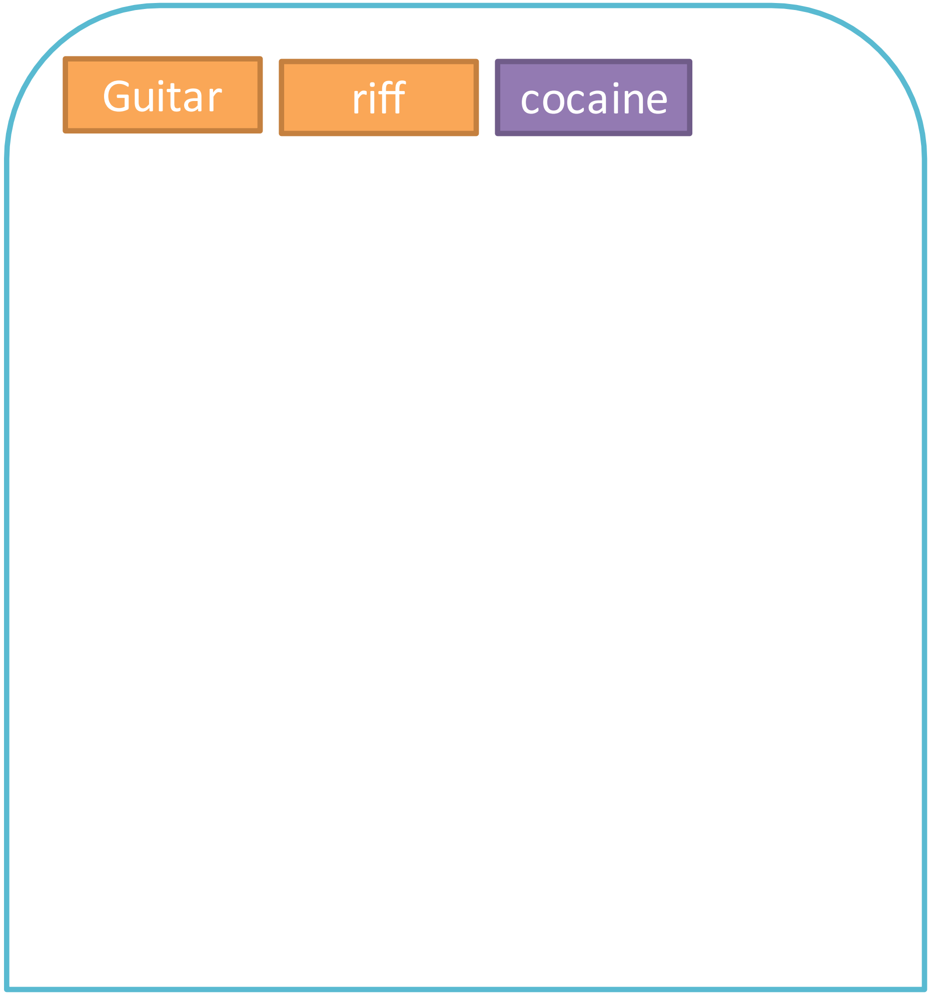
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

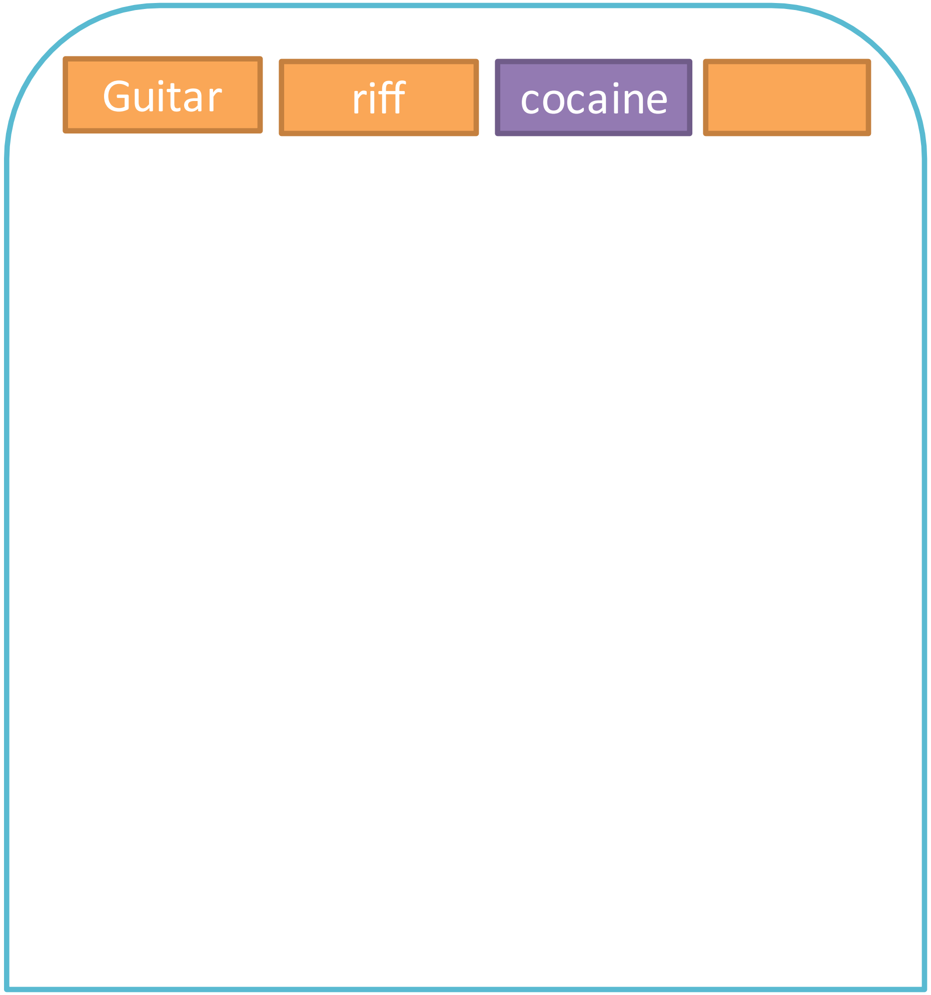
Choose next word’s topic.
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

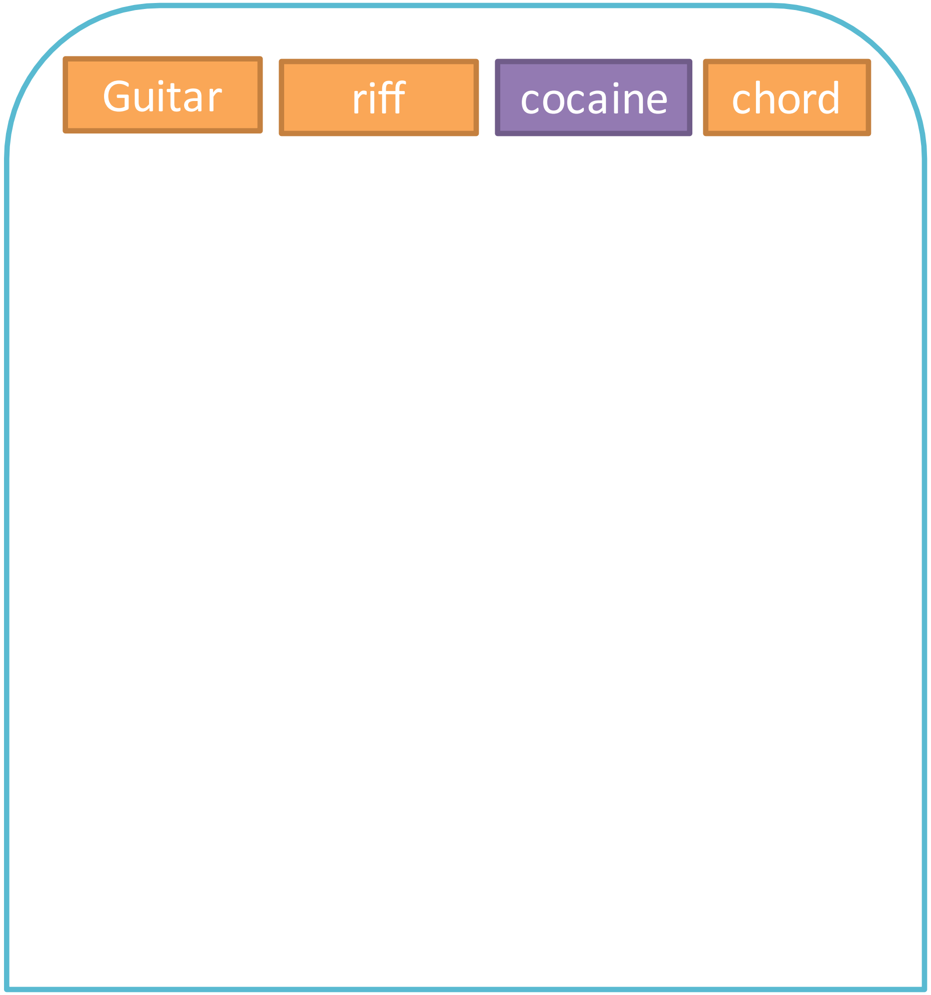
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

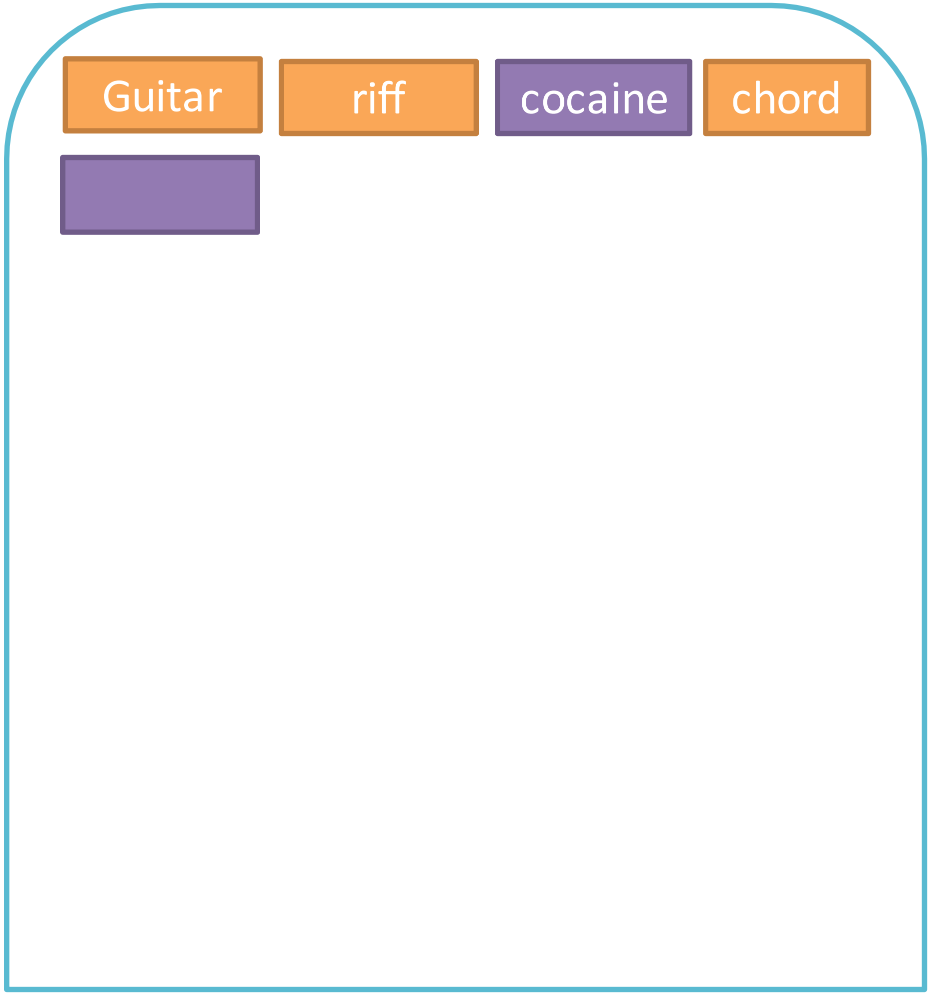
Choose next word’s topic.
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

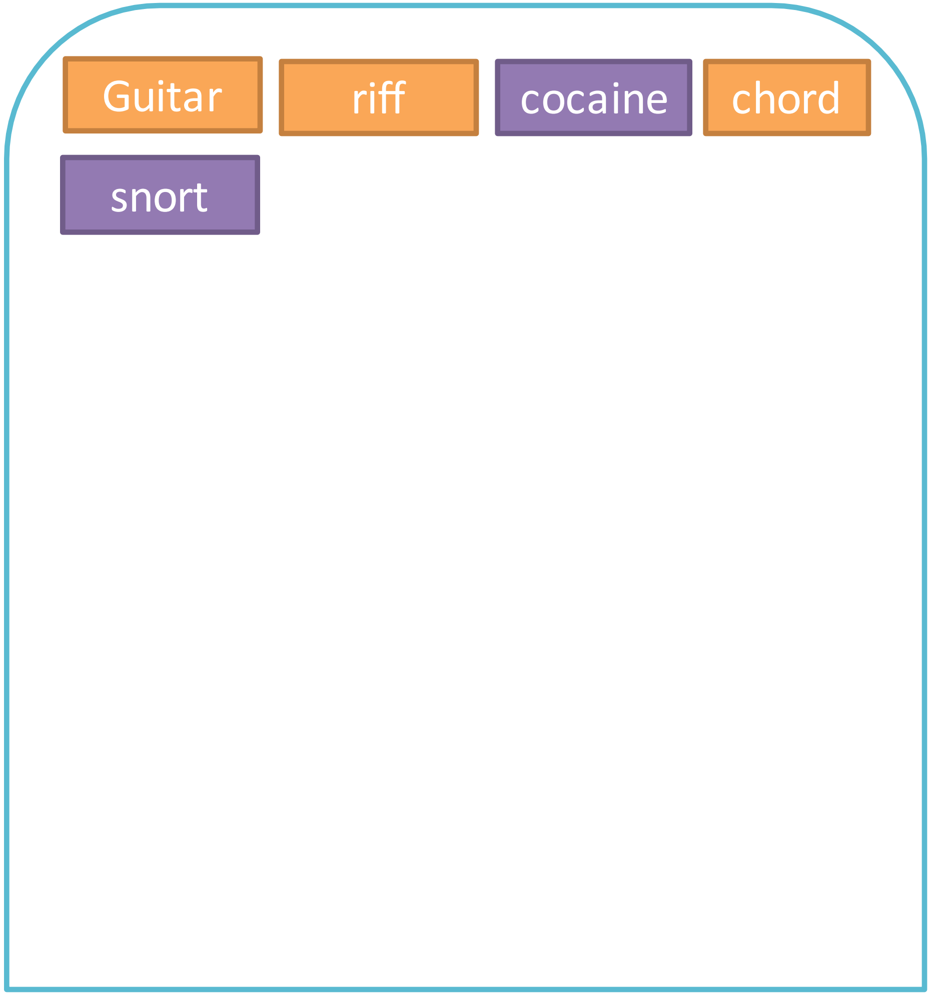
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let's use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I'll write a document.

First, I'll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock'n Roll:65%

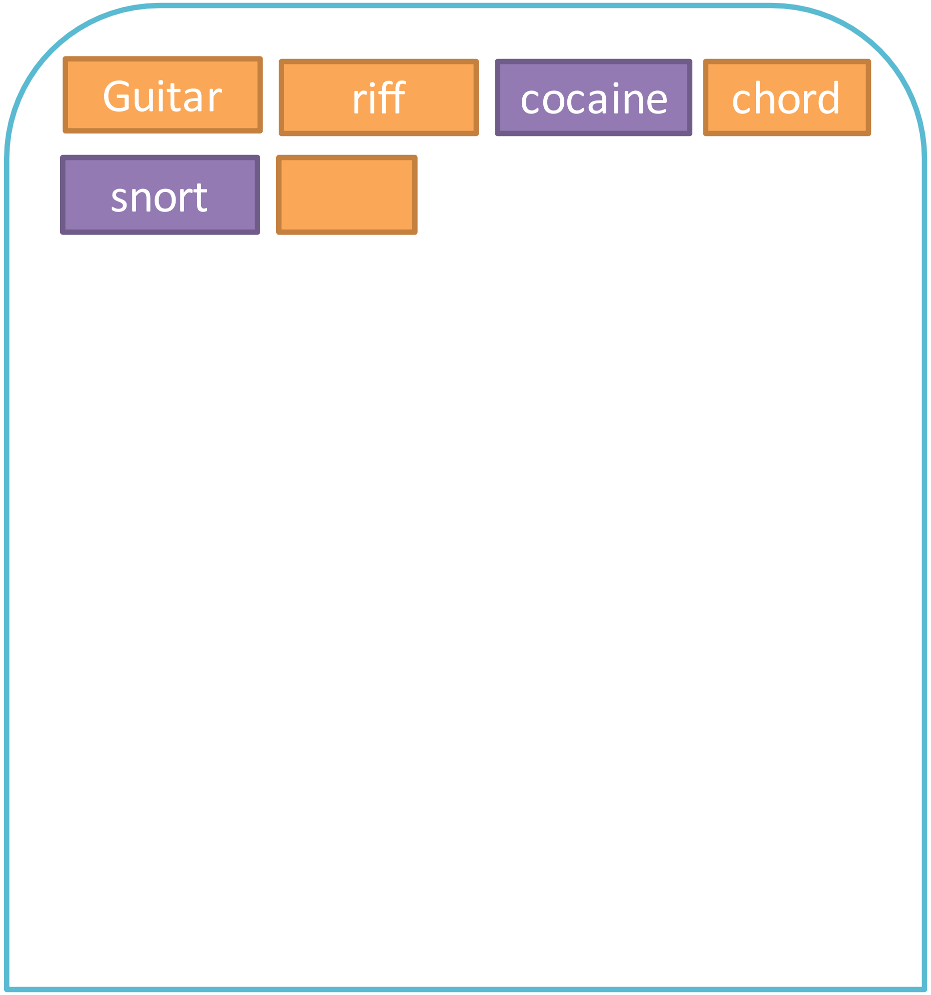
Choose next word's topic.
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

snort

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

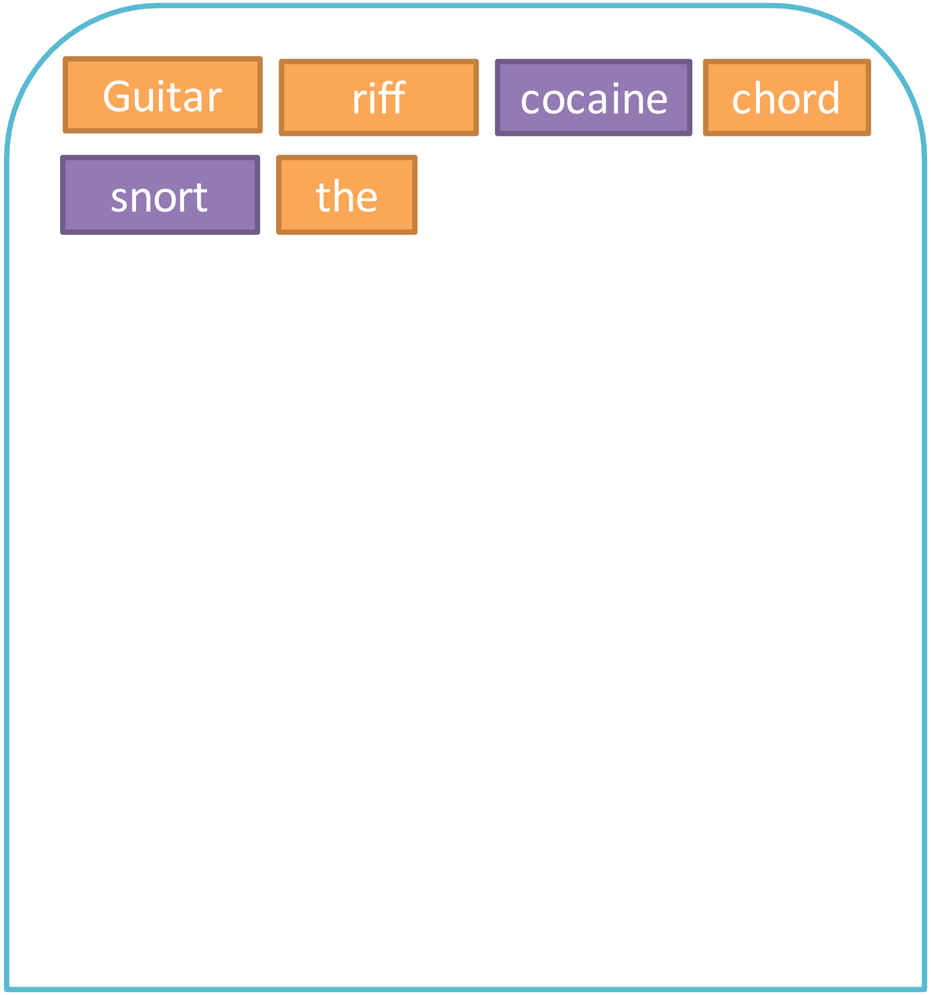
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

snort the

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

Choose next word’s topic.

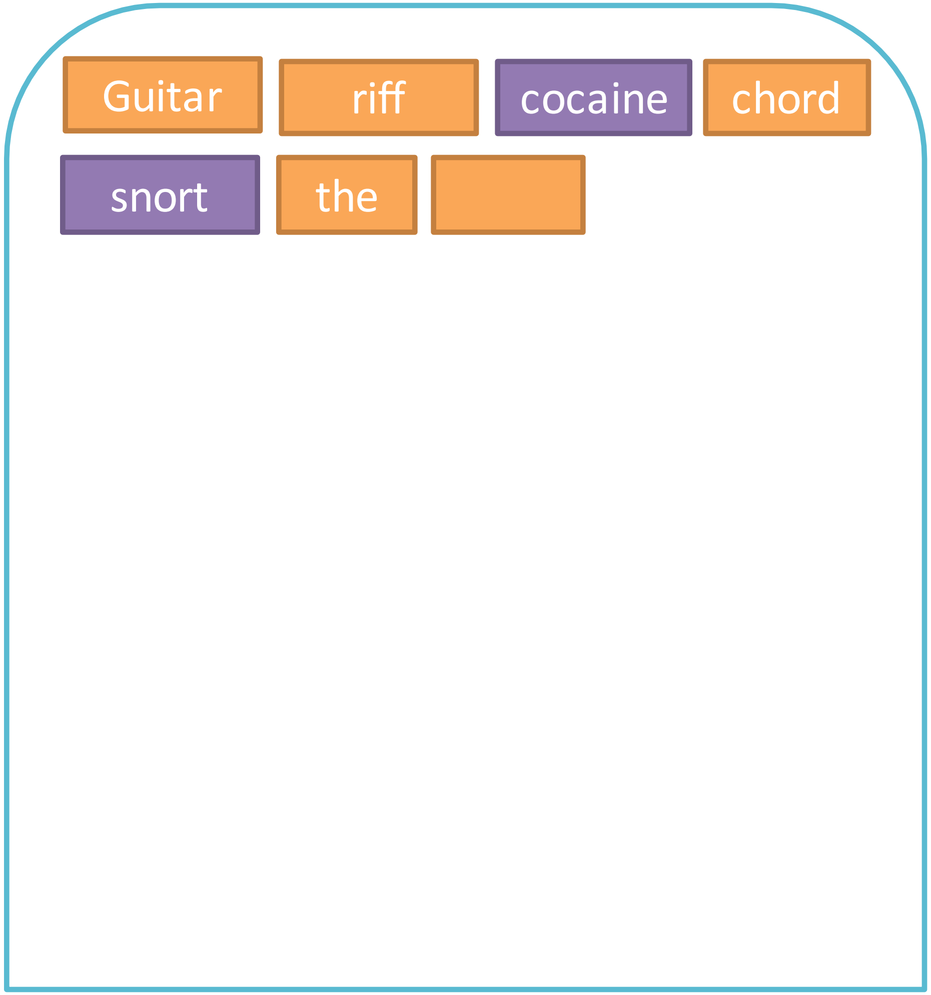
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

snort the

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

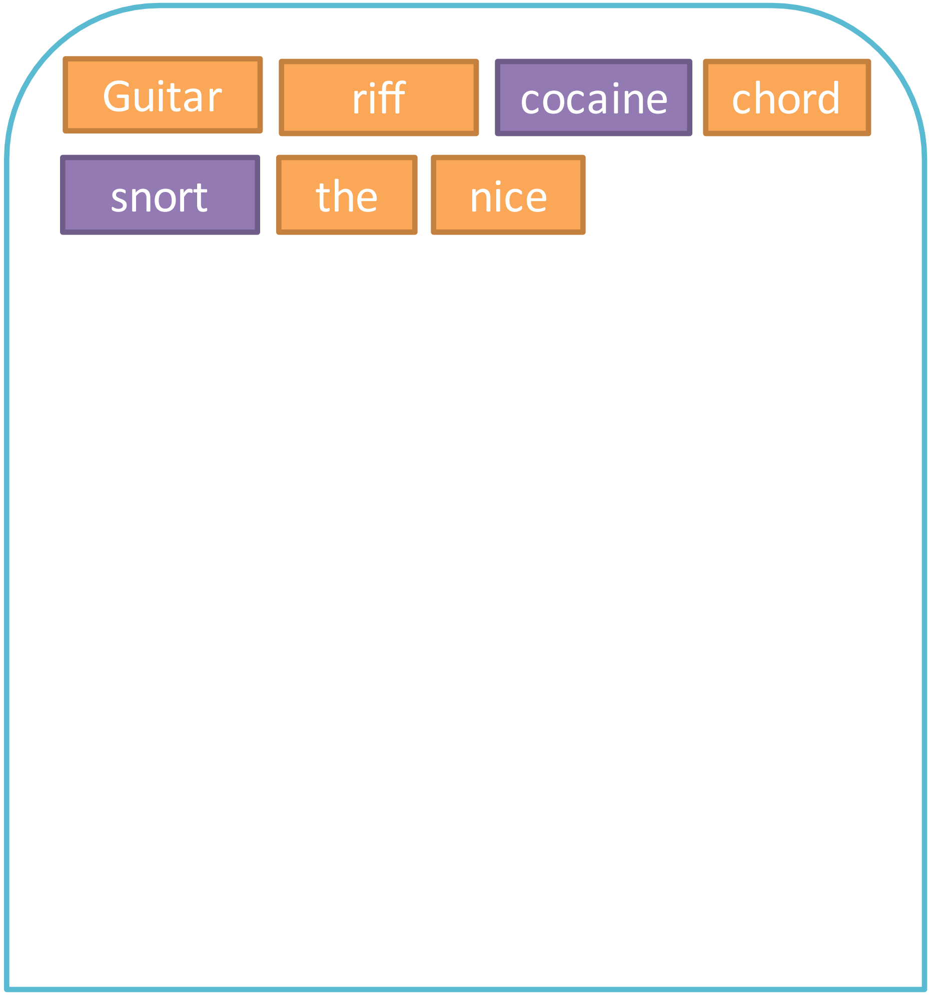
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

snort the nice

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

Choose next word’s topic.

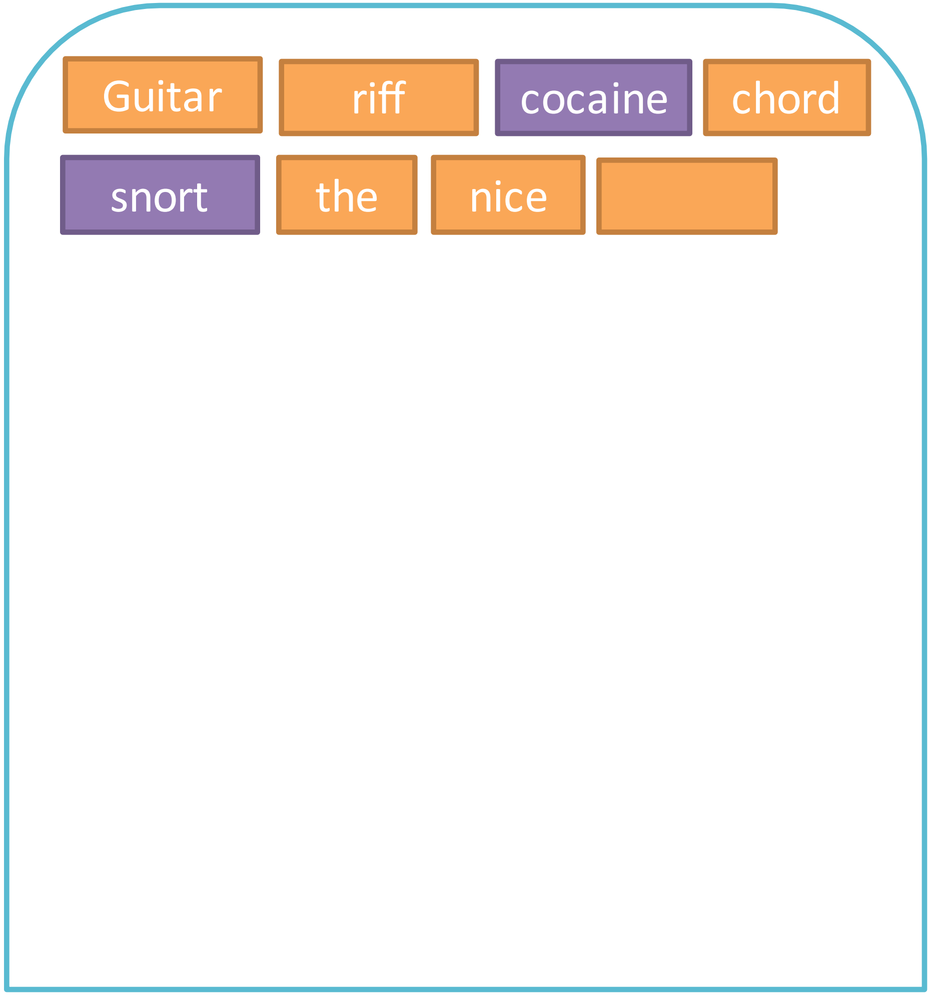
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let's use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

snort the nice

Empty page: I'll write a document.

First, I'll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock'n Roll:65%

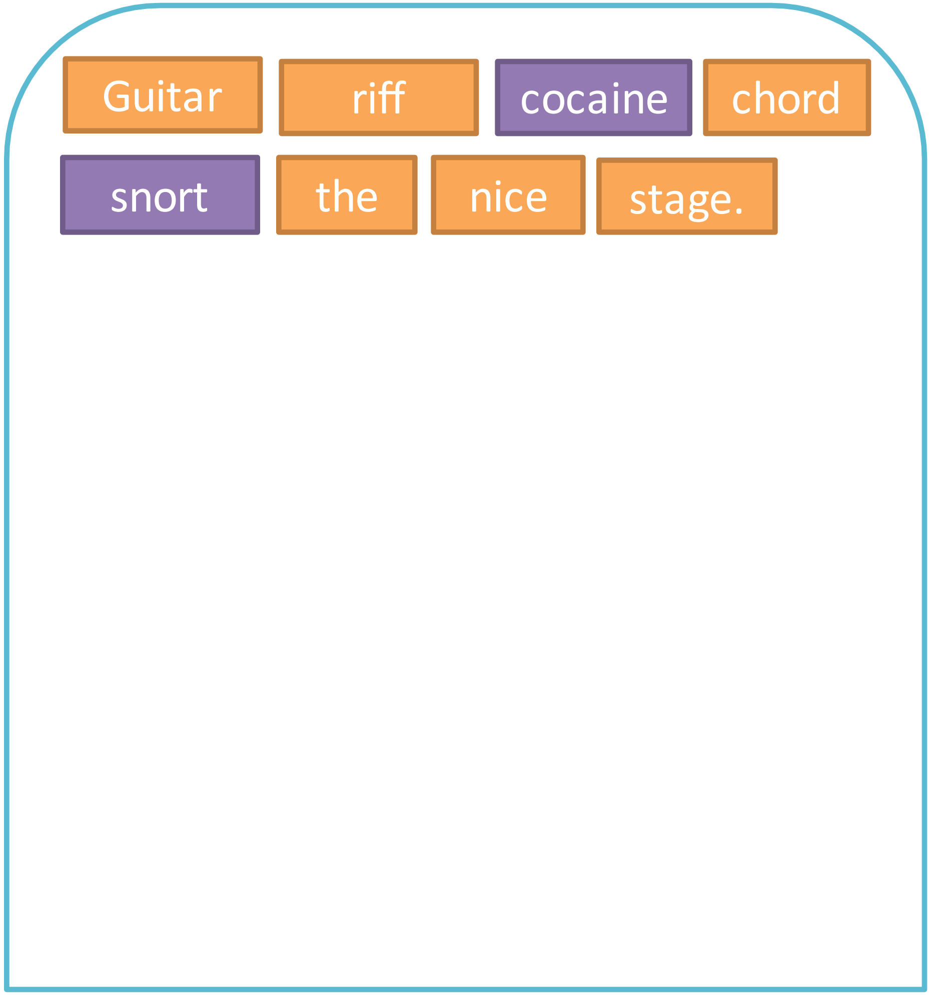
Choose next word's topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.
Sex:2%, Drugs:33%, Rock’n Roll:65%

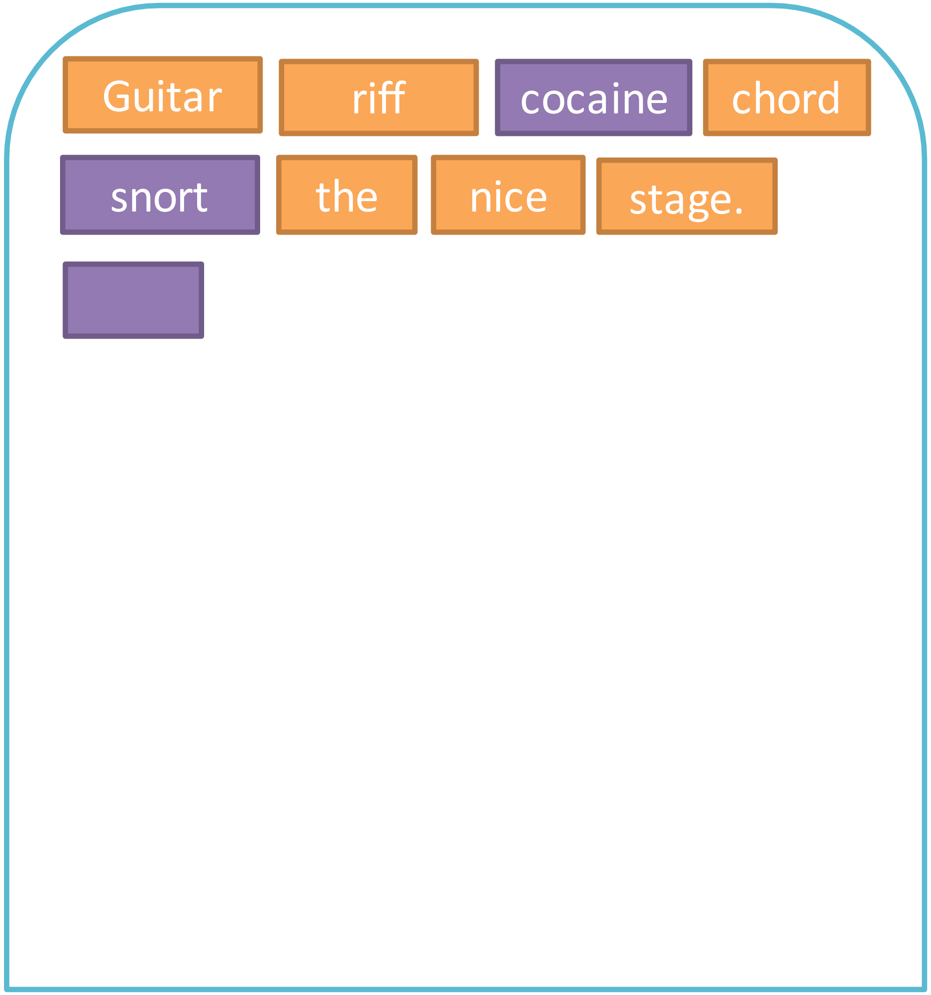
Choose next word’s topic.
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord snort the nice stage.

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

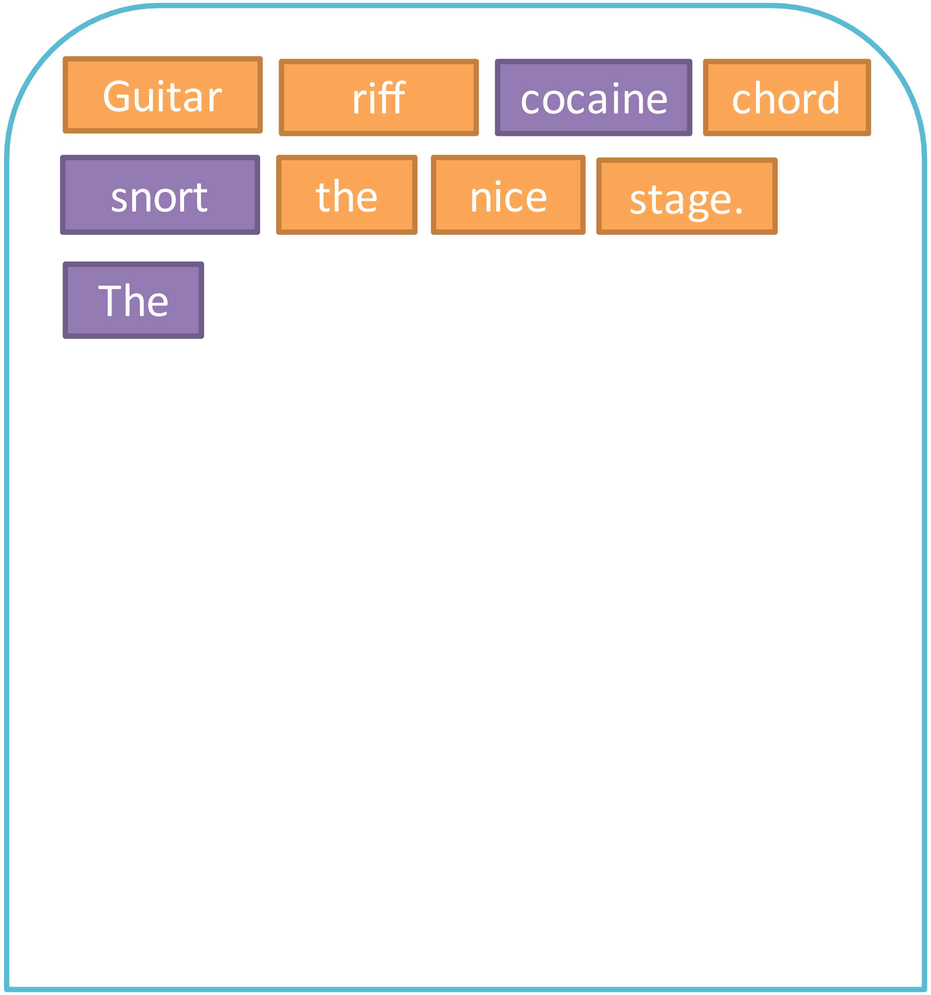
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord snort the nice stage. The

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

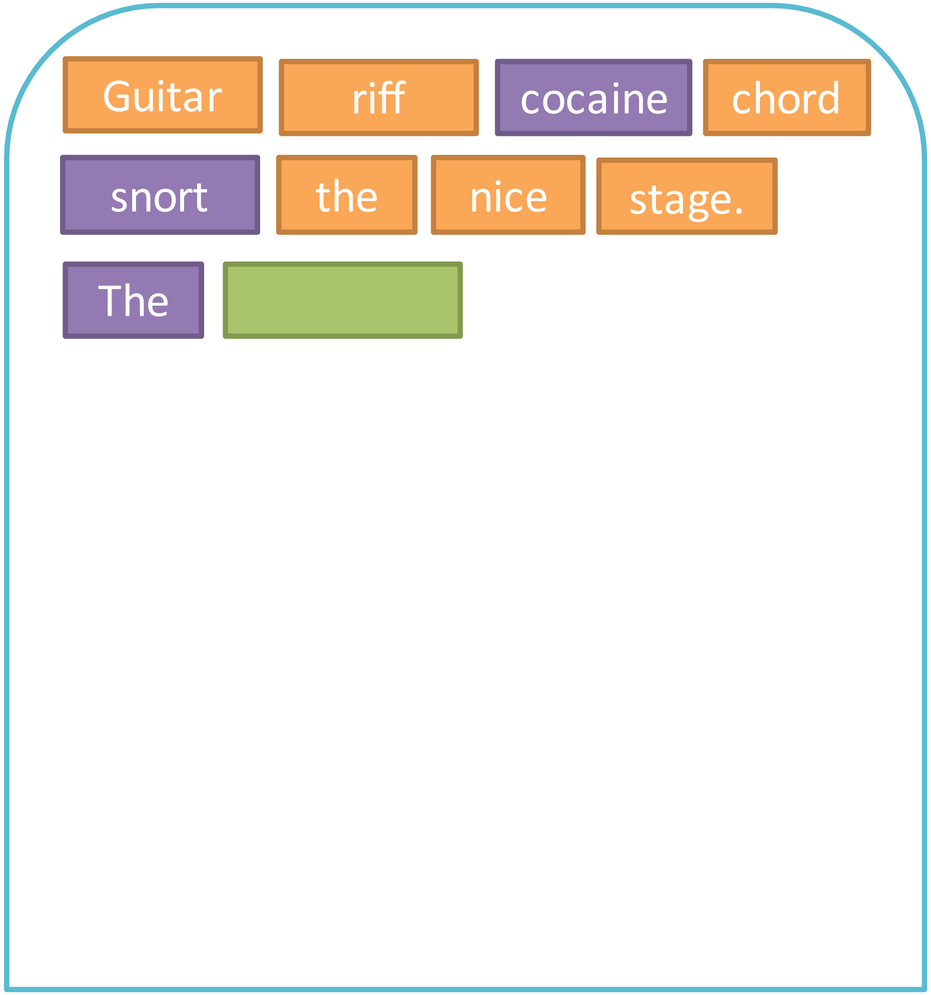
Choose next word’s topic.
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

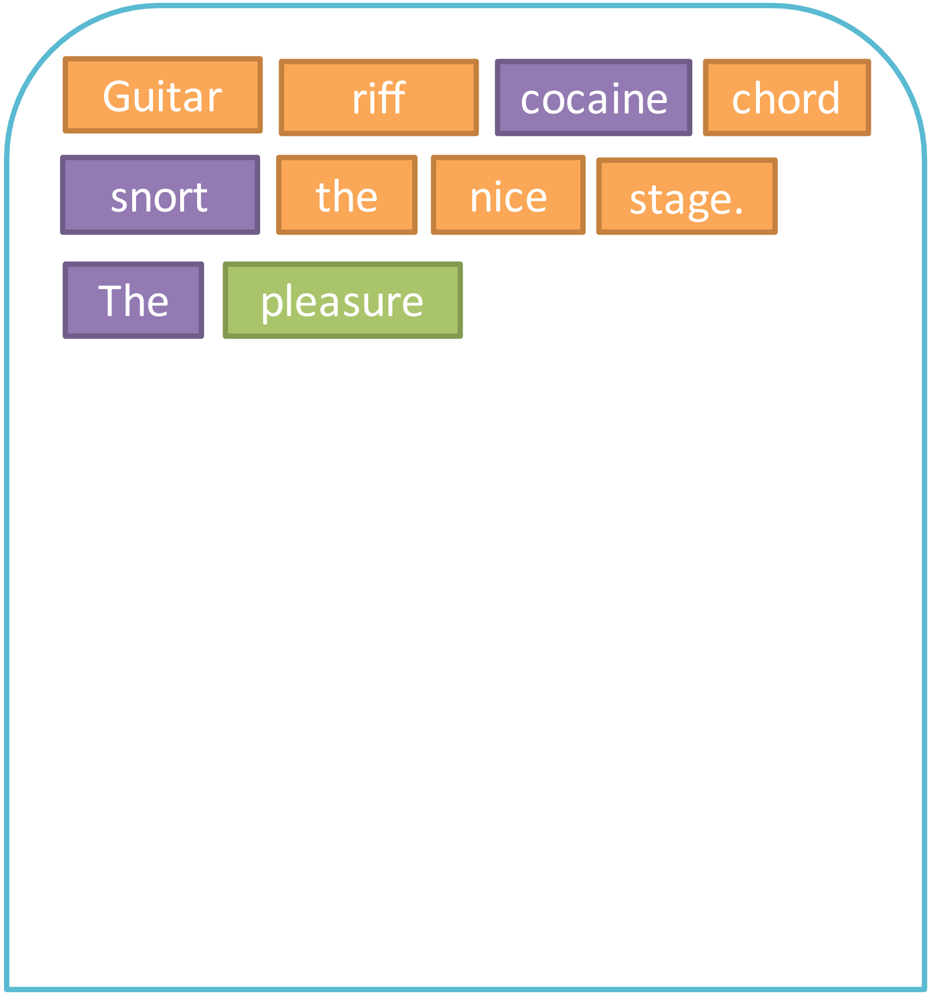
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord snort the nice stage. The pleasure

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

Choose next word’s topic.

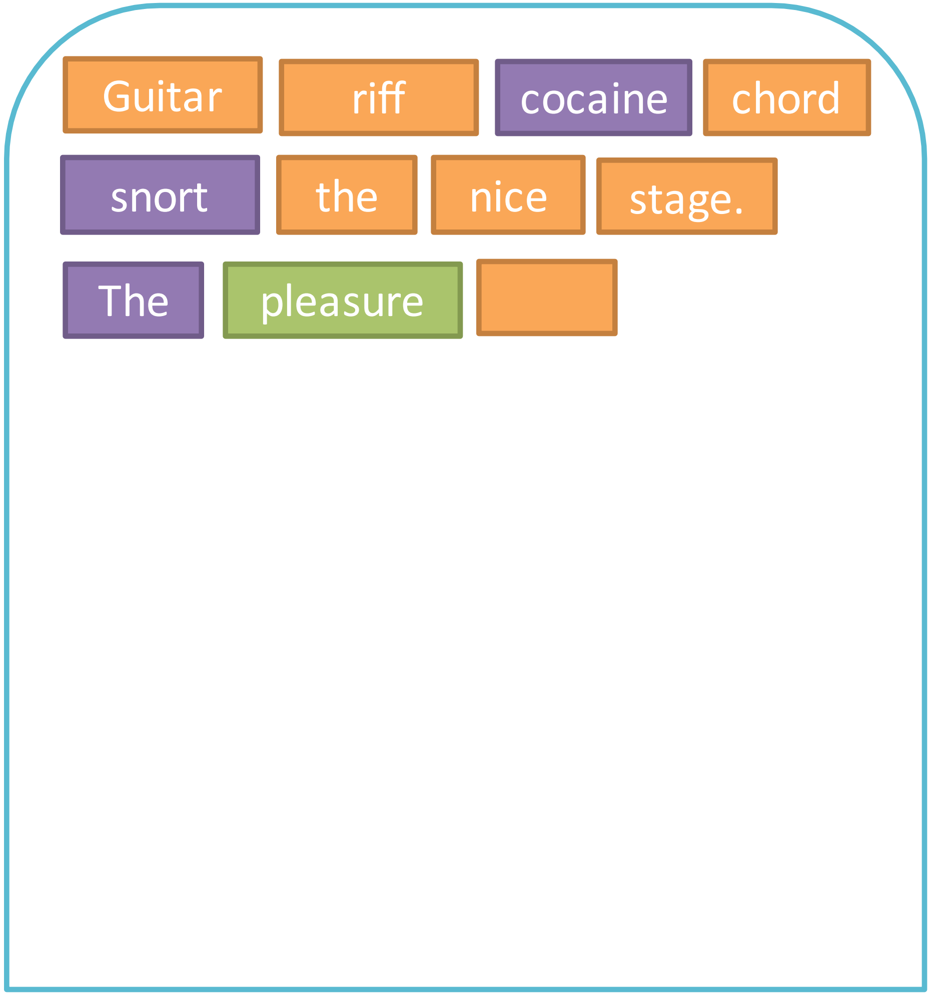
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let's use an algorithm specifically developed to find topics.

Model the process of writing



Empty page: I'll write a document.

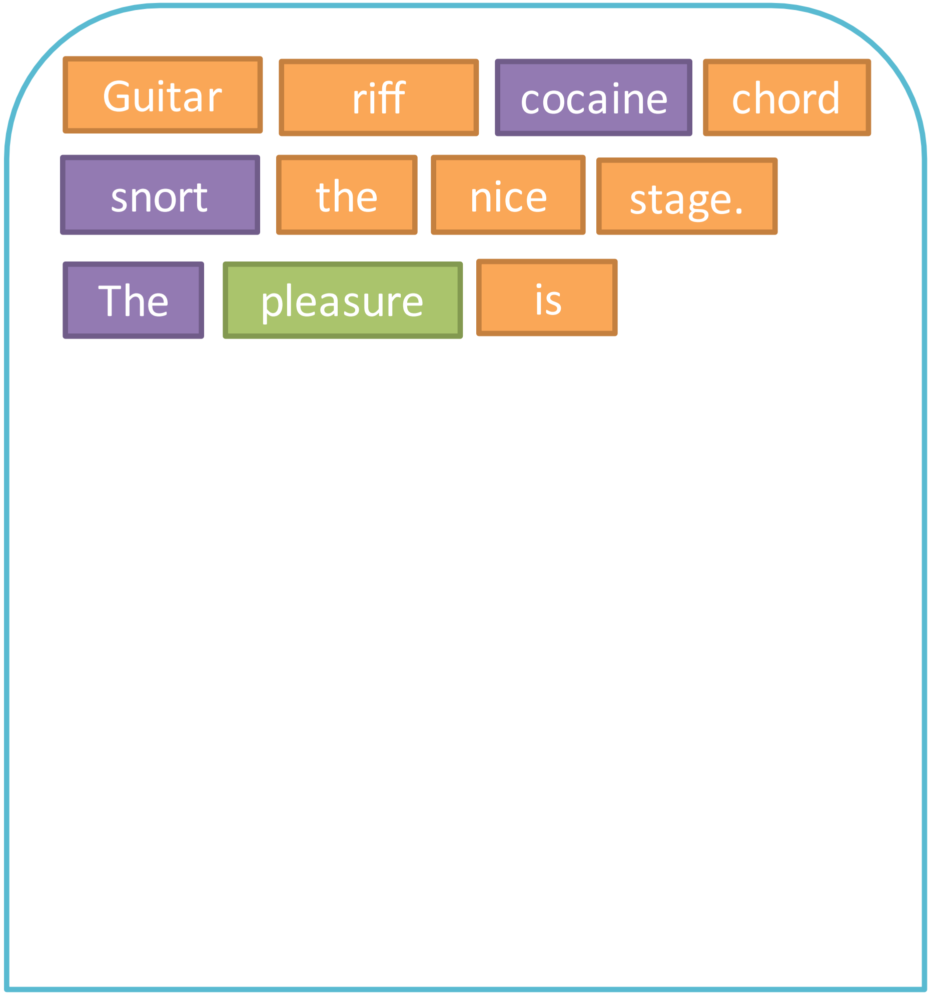
First, I'll decide what topics to write on. Choose the topic distribution.
Sex:2%, Drugs:33%, Rock'n Roll:65%

Choose next word's topic.
Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

snort the nice stage.

The pleasure is

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

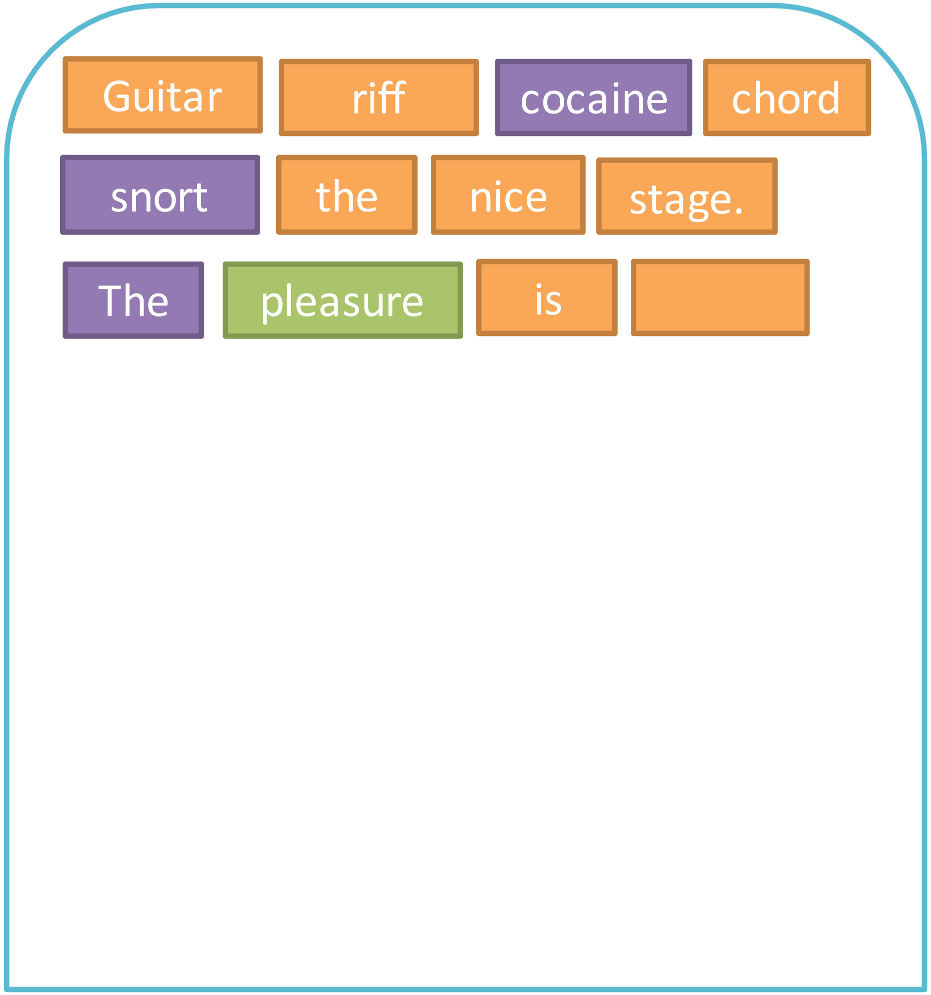
Choose next word’s topic.
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord snort the nice stage. The pleasure is

Empty page: I’ll write a document.

First, I’ll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock’n Roll:65%

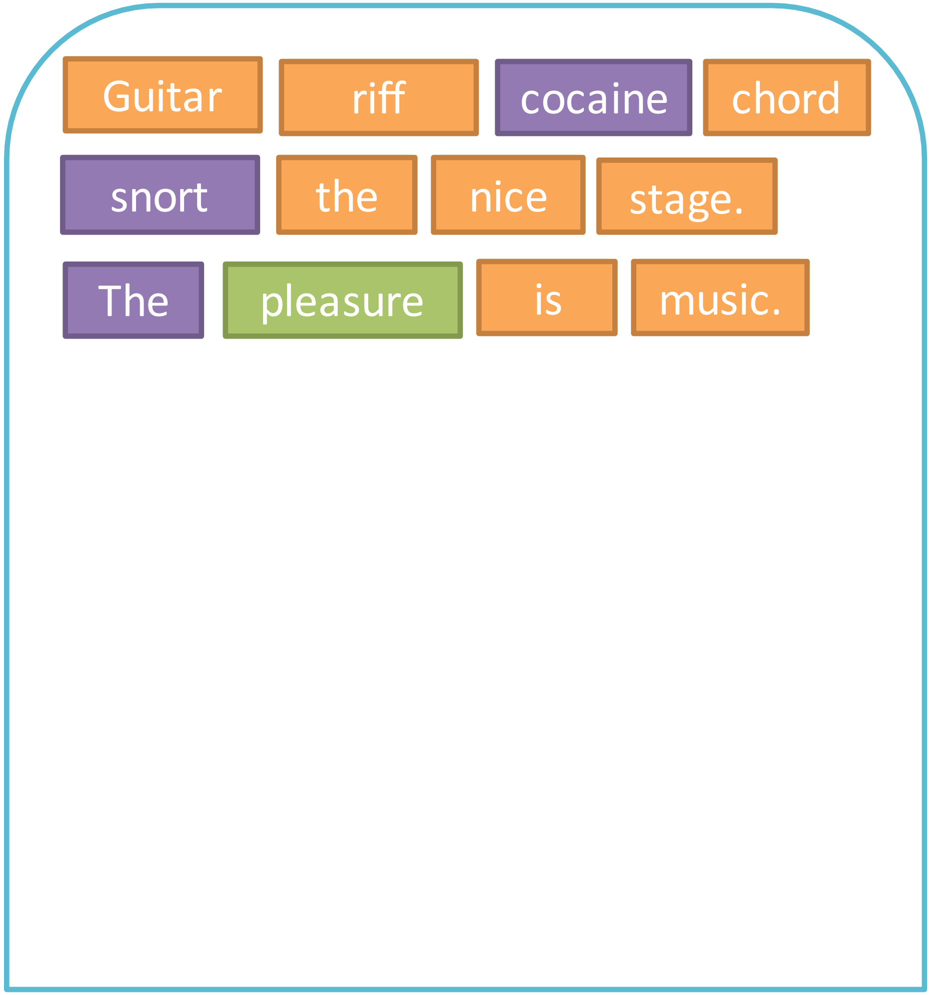
Choose next word’s topic.

Roll the dice.

Topic Modeling: LDA

Let’s use an algorithm specifically developed to find topics.

Model the process of writing



Guitar riff cocaine chord

snort the nice stage.

The pleasure is music.

Empty page: I’ll write a document.

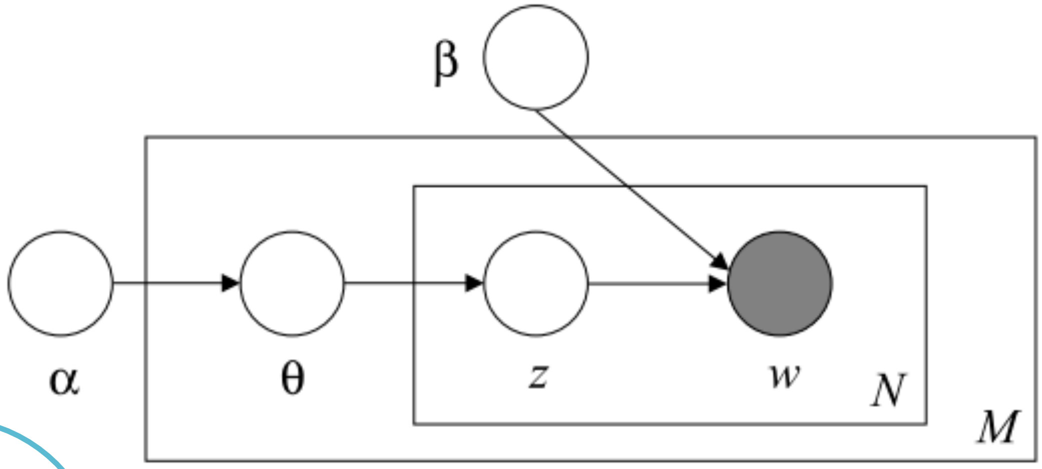
First, I’ll decide what topics to write on. Choose the topic distribution.

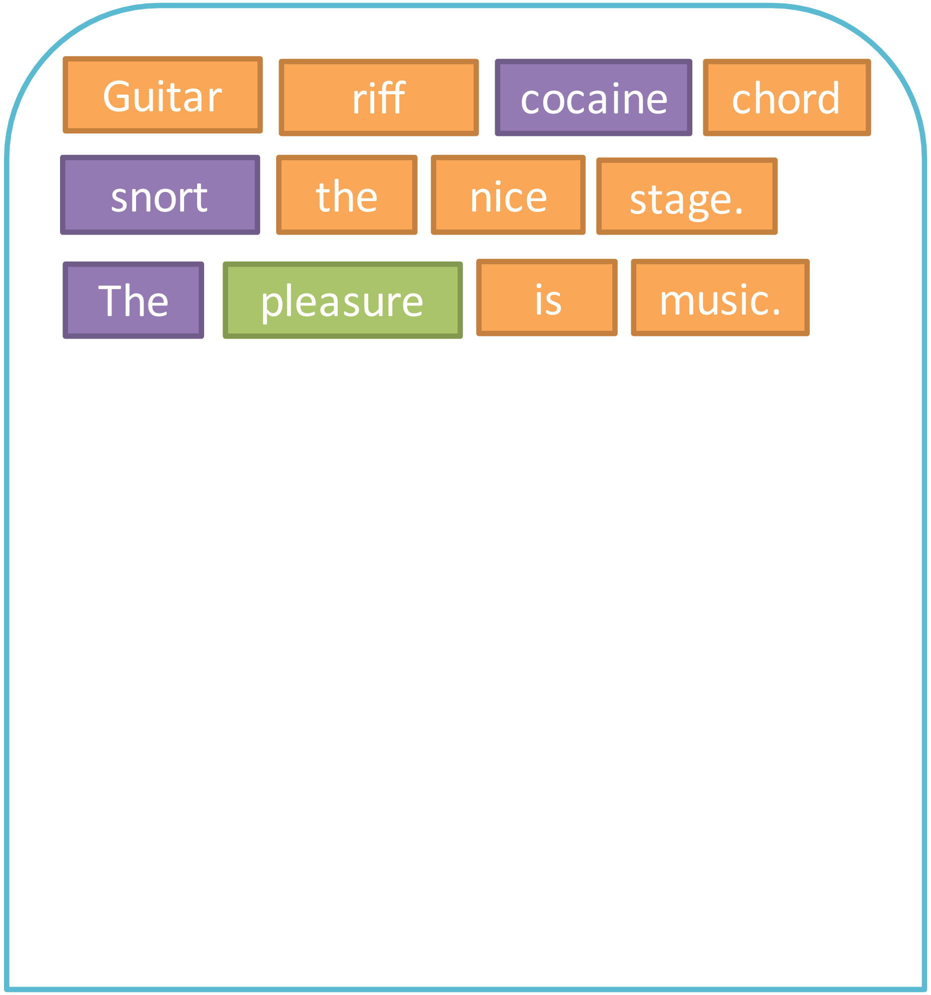
Sex:2%, Drugs:33%, Rock’n Roll:65%

Choose next word’s topic.
Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA





Guitar riff cocaine chord snort the nice stage. The pleasure is music.

First, I'll decide what topics to write on. Choose the topic distribution.

Sex:2%, Drugs:33%, Rock'n Roll:65%

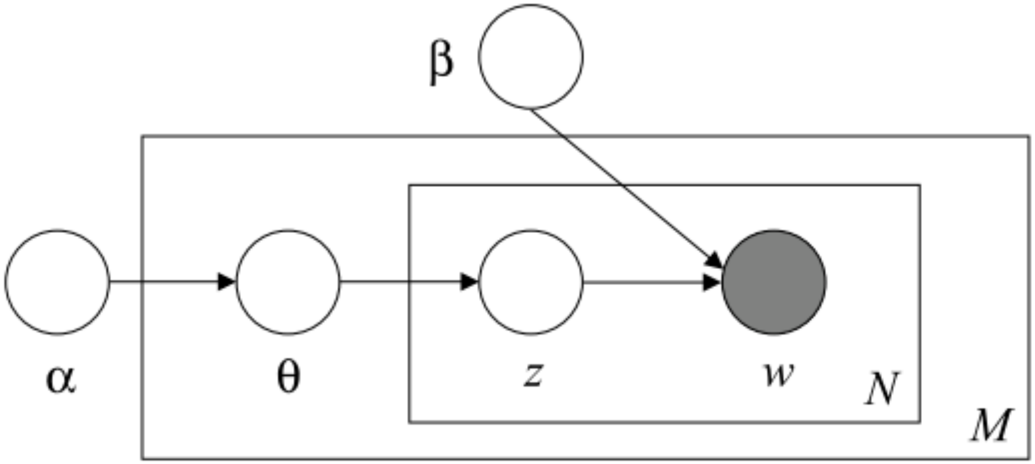
Choose next word's topic.

Roll the dice.

Choose the word according to this topic. Roll the dice.

Topic Modeling: LDA

Choosing word for a topic:
topic’s prob dist. over words



Corpus:
possible topic distributions for documents

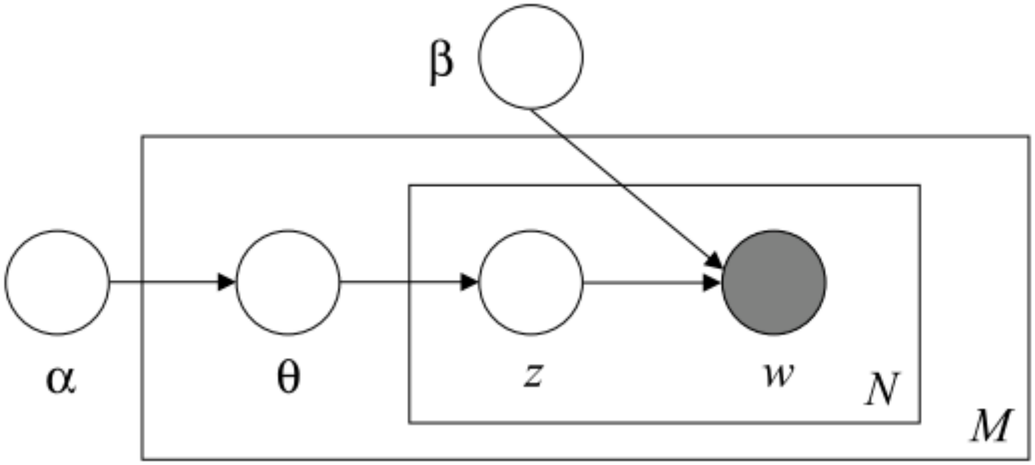
Each doc:
Topic dist. for doc

Each word:
which topic

Each word:
which actual word

Topic Modeling: LDA

Dirichlet
topic’s prob dist. over words



Dirichlet
Topic
dist.
for doc

Multinomial
which
topic

Multinomial
which
actual
word

Topic Modeling: LDA

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

How to solve this?

Math is tricky, but it’s still Bayes at heart…

Options (see resources):

- Gibbs Sampling
- Variational Bayesian Inference

Moving to Topic Space

What we’ve essentially done is transformed our higher dimensional space into a much reduced “topic space”.

We can use this topic space as we would any vector space of observations, for ML purposes.

Topic Modeling

What and why

Rotating the coordinate space

We regard documents as made of different portions of topics
Instead of different proportions of words.
Word space → Topic space

Similarity of docs

Searching for similar documents may be more meaningful in topic space

Dimensionality reduction

Clustering/classifying in topic space can be easier/meaningful

Intuition, Understanding

Look at prob. Dist. For topics, and how they are distributed over docs.