Computational Text Analysis for Legal Practice

(Class 4)

Charles Crabtree & Kevin L. Cope

January 21, 2022

github.com/cdcrabtree/uva-2022

Class Outline

- Class 4: Topic model human rights documents.
- Goal: Represent texts as a mixture of topics.

Class Outline

- Class 4: Topic model human rights documents.
- Goal: Represent texts as a mixture of topics.
- Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)

Class Outline

- Class 4: Topic model human rights documents.
- Goal: Represent texts as a mixture of topics.
- Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)

Plan

Class Outline

- Class 4: Topic model human rights documents.
- Goal: Represent texts as a mixture of topics.
- Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)

Plan

1. Single versus Mixed Membership models.

Class Outline

- Class 4: Topic model human rights documents.
- Goal: Represent texts as a mixture of topics.
- Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)

Plan

1. Single versus Mixed Membership models.
2. Topic modeling intuition, output, decision points.

Class Outline

- Class 4: Topic model human rights documents.
- Goal: Represent texts as a mixture of topics.
- Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)

Plan

1. Single versus Mixed Membership models.
2. Topic modeling intuition, output, decision points.
3. Interpretation and applications.

Class Outline

- Class 4: Topic model human rights documents.
- Goal: Represent texts as a mixture of topics.
- Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)

Plan

1. Single versus Mixed Membership models.
2. Topic modeling intuition, output, decision points.
3. Interpretation and applications.

Key Words

- Mixed membership model.
- Topic models.

Key Words

- Mixed membership model.
- Topic models.
- Topic and topic proportions.

Key Words

- Mixed membership model.
- Topic models.
- Topic and topic proportions.
- Latent Dirichlet Allocation (LDA).

Key Words

- Mixed membership model.
- Topic models.
- Topic and topic proportions.
- Latent Dirichlet Allocation (LDA).
- Structural Topic Modeling (STM).

Key Words

- Mixed membership model.
- Topic models.
- Topic and topic proportions.
- Latent Dirichlet Allocation (LDA).
- Structural Topic Modeling (STM).

Key R packages

- `stm`

Single vs. Mixed Membership Models

Clustering

Document ⇝ one cluster

Doc 1

Doc 2

Doc 3

Doc *N*

Cluster 1

Cluster *N*

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ one cluster

Doc 1
Doc 2
Doc 3
Doc *N*

Cluster 1
Cluster 2
Cluster *N*

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ one cluster

Doc 1 Cluster 1
Doc 2 Cluster 2
Doc 3 Cluster *N*
Doc *N*

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ one cluster

Doc 1
Doc 2
Doc 3
Doc N

Cluster 1
Cluster 2
Cluster N

Single vs. Mixed Membership Models

Clustering

Document ⇝ one cluster

Doc 1
Doc 2
Doc 3
Doc N

Cluster 1
Cluster 2
Cluster N

Single vs. Mixed Membership Models

Topic Models (Mixed Membership)

Document ⇝ many clusters

Doc 1

Doc 2

Doc 3

Doc *N*

Cluster 1

Cluster 2

Cluster *N*

Single vs. Mixed Membership Models

Topic Models (Mixed Membership)

Document ⇝ many clusters

Doc 1

Doc 2

Doc 3

Doc N

Cluster 1

Cluster 2

Cluster N

What is Topic Modeling?

Topic modeling is an algorithm used to code the content of a corpus into substantively meaningful categories, or "topics," using the statistical correlations between words.

It is unsupervised because we don't tell it the topics beforehand. The algorithm "discovers" abstract topics that can be thought of as a constellation of words that tend to show up together.

What is Topic Modeling?

Topic modeling is an algorithm used to code the content of a corpus into substantively meaningful categories, or "topics," using the statistical correlations between words.

It is unsupervised because we don't tell it the topics beforehand. The algorithm "discovers" abstract topics that can be thought of as a constellation of words that tend to show up together.

It is mixed membership because it considers each document to be a mixture of different topics.

What is Topic Modeling?

Topic modeling is an algorithm used to code the content of a corpus into substantively meaningful categories, or "topics," using the statistical correlations between words.

It is unsupervised because we don't tell it the topics beforehand. The algorithm "discovers" abstract topics that can be thought of as a constellation of words that tend to show up together.

It is mixed membership because it considers each document to be a mixture of different topics.

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`
- `I ate a banana and spinach smoothie for breakfast.`

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`
- `I ate a banana and spinach smoothie for breakfast.`
- `Hamsters and kittens are cute.`

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`
- `I ate a banana and spinach smoothie for breakfast.`
- `Hamsters and kittens are cute.`
- `My sister adopted a kitten yesterday.`

How does Topic Modeling Work?

Goal: Topic model the following documents:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Hamsters and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`
- `I ate a banana and spinach smoothie for breakfast.`
- `Hamsters and kittens are cute.`
- `My sister adopted a kitten yesterday.`
- `Look at this cute hamster munching on a piece of broccoli.`

We suspect that this corpus contains 2 topics. We want to reverse engineer those topics from the co-occurrence of words in each document.

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`
- `I ate a banana and spinach smoothie for breakfast.`
- `Hamsters and kittens are cute.`
- `My sister adopted a kitten yesterday.`
- `Look at this cute hamster munching on a piece of broccoli.`

We suspect that this corpus contains 2 topics. We want to reverse engineer those topics from the co-occurrence of words in each document.

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`

How does Topic Modeling Work?

Goal: Topic model the following documents:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`
- `I ate a banana and spinach smoothie for breakfast.`
- `Hamsters and kittens are cute.`

How does Topic Modeling Work?

Goal: Topic model the following documents:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Hamsters and kittens are cute.
- My sister adopted a kitten yesterday.

How does Topic Modeling Work?

Goal: Topic model the following documents:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Hamsters and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

How does Topic Modeling Work?

Goal: Topic model the following documents:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Hamsters and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

Topic A (interpreted to be about Food)

Topic B (interpreted to be about Pets)

How does Topic Modeling Work?

Goal: Topic model the following documents:

- `I like to eat broccoli and bananas.`
- `I ate a banana and spinach smoothie for breakfast.`
- `Hamsters and kittens are cute.`
- `My sister adopted a kitten yesterday.`
- `Look at this cute hamster munching on a piece of broccoli.`

Topic A (interpreted to be about Food)

Topic B (interpreted to be about Pets)

Latent Dirichlet Allocation

LDA: Popular topic modeling method.

Inputs.

1. A document term matrix (or any multidimensional dataset).
2. K: the expected number of topics.

Latent Dirichlet Allocation

LDA: Popular topic modeling method.

Inputs.

1. A document term matrix (or any multidimensional dataset).
2. K: the expected number of topics.

Outputs.

1. π_k: Topic distribution over words.

Latent Dirichlet Allocation

LDA: Popular topic modeling method.

Inputs.

1. A document term matrix (or any multidimensional dataset).
2. K: the expected number of topics.

Outputs.

1. π_k: Topic distribution over words.
2. θ_i: Document distribution over topics.

Latent Dirichlet Allocation

LDA: Popular topic modeling method.

Inputs.

1. A document term matrix (or any multidimensional dataset).
2. K: the expected number of topics.

Outputs.

1. π_k: Topic distribution over words.
2. θ_i: Document distribution over topics.

LDA Decisions

Small decisions have potentially huge consequences.

1. How should we preprocess the data?
    - Topic models are sensitive to feature selection
    - Common to remove sparse words, but there is much debate.

LDA Decisions

Small decisions have potentially huge consequences.

1. How should we preprocess the data?
    - Topic models are sensitive to feature selection
    - Common to remove sparse words, but there is much debate.
2. How to chose K?
    - User must assign the number of topics (K).
    - Different values of K will lead to different partitions.

LDA Decisions

Small decisions have potentially huge consequences.

1. How should we preprocess the data?
    - Topic models are sensitive to feature selection
    - Common to remove sparse words, but there is much debate.
2. How to chose *K*?
    - User must assign the number of topics (K).
    - Different values of K will lead to different partitions.
3. Random starting values!
    - Results will depend on the initial assignments.
    - Important to run the algorithm multiple times from different random starting values.

LDA Decisions

Small decisions have potentially huge consequences.

1. How should we preprocess the data?
   - Topic models are sensitive to feature selection
   - Common to remove sparse words, but there is much debate.
2. How to chose K?
   - User must assign the number of topics (K).
   - Different values of K will lead to different partitions.
3. Random starting values!
   - Results will depend on the initial assignments.
   - Important to run the algorithm multiple times from different random starting values.

How do we decide these issues?

LDA Decisions

Small decisions have potentially huge consequences.

1. How should we preprocess the data?
   - Topic models are sensitive to feature selection
   - Common to remove sparse words, but there is much debate.
2. How to chose *K*?
   - User must assign the number of topics (K).
   - Different values of K will lead to different partitions.
3. Random starting values!
   - Results will depend on the initial assignments.
   - Important to run the algorithm multiple times from different random starting values.

How do we decide these issues?

What Makes a Good Topic Model?

A good topic model is one for which topics are substantially / semantically interpretable.

How do we interpret the topics?

1. Look at top / distinctive words for each topic.
2. Read most representative documents for each topic.

What Makes a Good Topic Model?

A good topic model is one for which topics are substantially / semantically interpretable.

How do we interpret the topics?

1. Look at top / distinctive words for each topic.
2. Read most representative documents for each topic.

Let's Program

Launch RStudio and we can get started programming.

kcope@law.virginia.edu | crabtree@dartmouth.edu