Introduction to Machine Learning for Social Science

Class 14: Topic Modeling

Rochelle Terman

Postdoctoral Fellow
Center for International Security Cooperation
Stanford University

March 1, 2018

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?
- **Argument:** U.S. news coverage portrays Muslim societies differently than non-Muslim societies.

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?
- **Argument:** U.S. news coverage portrays Muslim societies differently than non-Muslim societies.
- **Data:** 35 years of reporting in the *New York Times* and *Washington Post.*

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?
- **Argument:** U.S. news coverage portrays Muslim societies differently than non-Muslim societies.
- **Data:** 35 years of reporting in the *New York Times* and *Washington Post*.
- **Method:** Topic Modeling

Today: Topic model American news coverage of women abroad
Goal: represent each article as a mixture of topics:

- Describe each topic.
- Measure proportion of each article addressing each topic.

Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)
Game Plan:

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

Key Terms:

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

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Cluster 1

Doc 2

Cluster 2

Doc 3

$\vdots$

$\vdots$

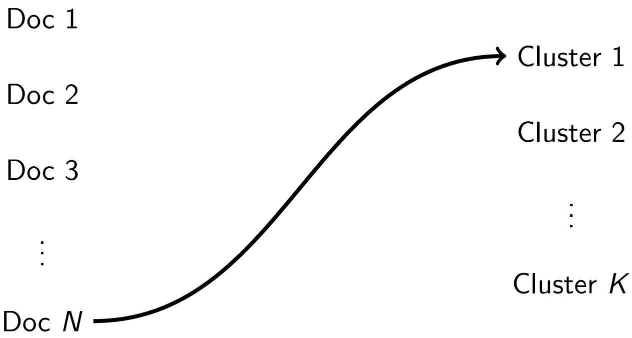
Cluster K

Doc N

Single vs. Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster



Single vs. Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

Doc 1

Cluster 1

Doc 2

Cluster 2

Doc 3

$\vdots$

$\vdots$

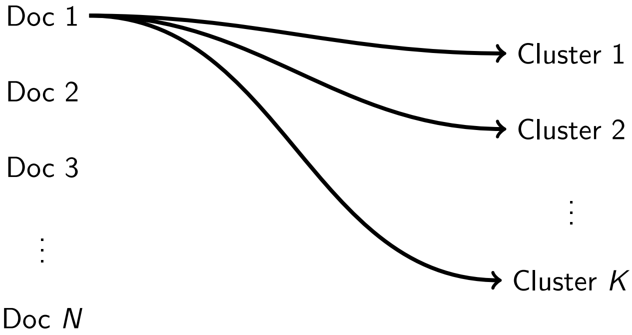
Cluster K

Doc N

Single vs. Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters



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.

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.

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.`

Topic A (interpreted to be about Food)
Topic B (interpreted to be about Pets)

Latent Dirichlet Allocation

LDA: Popular topic modeling method.

Latent Dirichlet Allocation

LDA: Popular topic modeling method.

Inputs

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

Latent Dirichlet Allocation

LDA: Popular topic modeling method.

Inputs

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

Outputs

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

LDA: Outputs

- 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.

1) Topic distribution over words (π_k).

Topic	broccoli	banana	breakfast	kitten	cute	hamster	like	yesterday	*Total*
A	.30	.25	.20	.01	.01	.01	.12	.10	1
B	.01	01	01	.35	.24	.25	.08	.05	1

LDA: Outputs

- 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.

2) Document distribution over topics (θ_i).

Document	Topic A Weight	Topic B Weight	Total
1	.99	.01	1
2	.99	.01	1
3	.01	99	1
4	.01	99	1
4	.60	.40	1

LDA: Decisions

Small Decisions with Big Consequences:

LDA: Decisions

Small Decisions with Big 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 with Big 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 with Big 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!
    - Like clustering, esults will depend on the initial assignments.
    - Important to run the algorithm multiple times from different random starting values.

LDA: Decisions

Small Decisions with Big 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!
   - Like clustering, esults will depend on the initial assignments.
   - Important to run the algorithm multiple times from different random starting values.

How do we decide?

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.

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?
- **Argument:** U.S. news coverage stigmatizes Muslim societies as distinctly sexist.

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?
- **Argument:** U.S. news coverage stigmatizes Muslim societies as distinctly sexist.
- **Data:** 35 years of reporting in the *New York Times* and *Washington Post.*

An Example

Islamophobia and Media Portrayals of Muslim Women (*International Studies Quarterly*)

- **Question**: How do U.S. news media report about women's rights abroad?
- **Argument:** U.S. news coverage stigmatizes Muslim societies as distinctly sexist.
- **Data:** 35 years of reporting in the *New York Times* and *Washington Post*.
- **Method:** Structural Topic Modeling

Structural Topic Model

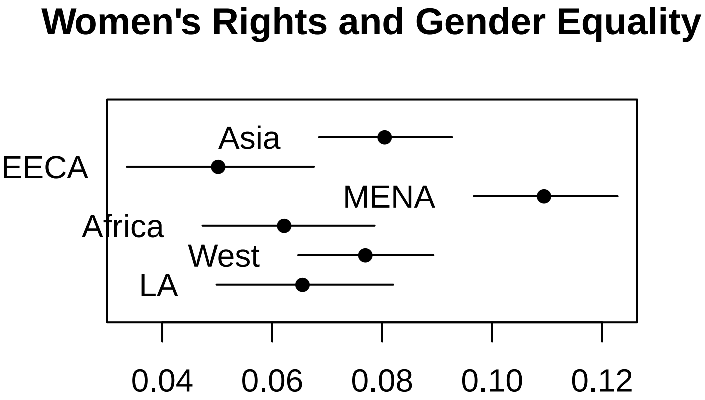
The structural topic model is an extension of LDA.

How does the prevalence of topics vary across groups of documents (by region, author, etc)?

Label	Probability Keywords	FREX Keywords
Business	said, work, compani, year, percent, job, busi, worker, million, market	compani, bank, industri, factori, employ, market, employe, busi, corpor, manag
Sports	team, women, game, play, world, said, olymp, sport, player, first	game, olymp, sport, player, soccer, athlet, coach, team, medal, championship
Fashion	black, dress, one, cloth, wear, design, street, fashion, citi, white	restaur, jacket, shirt, color, skirt, blue, worn, cloth, fashion, pant
Arts	film, book, show, art, work, stori, life, one, play, write	film, artist, novel, art, museum, theater, movi, charact, fiction, reader
Women's Rights & Gender Equality	women, men, femal, law, right, chang, male, equal, mani, issu	equal, male, gender, femal, discrimin, men, women, law, status, chang
Politics	polit, minist, govern, elect, parti, presid, said, vote, leader, prime	elect, vote, minist, prime, parti, candid, voter, cabinet, politician, polit
Religion	said, islam, religi, right, church, ban, law, countri, women, practic	islam, religi, religion, secular, veil, circumcis, fundamentalist, church, genit, koran

Label	Probability Keywords	FREX Keywords
Business	said, work, compani, year, percent, job, busi, worker, million, market	compani, bank, industri, factori, employ, market, employe, busi, corpor, manag
Sports	team, women, game, play, world, said, olymp, sport, player, first	game, olymp, sport, player, soccer, athlet, coach, team, medal, championship
Fashion	black, dress, one, cloth, wear, design, street, fashion, citi, white	restaur, jacket, shirt, color, skirt, blue, worn, cloth, fashion, pant
Arts	film, book, show, art, work, stori, life, one, play, write	film, artist, novel, art, museum, theater, movi, charact, fiction, reader
Women's Rights & Gender Equality	women, men, femal, law, right, chang, male, equal, mani, issu	equal, male, gender, femal, discrimin, men, women, law, status, chang
Politics	polit, minist, govern, elect, parti, presid, said, vote, leader, prime	elect, vote, minist, prime, parti, candid, voter, cabinet, politician, polit
Religion	said, islam, religi, right, church, ban, law, countri, women, practic	islam, religi, religion, secular, veil, circumcis, fundamentalist, church, genit, koran

Expected Topic Proportion Across Region



Today: Topic model American news coverage of women abroad
Goal: represent each article as a mixture of topics:

- Describe each topic.
- Measure proportion of each article addressing each topic.

Method: Latent Dirichlet Allocation (LDA); Structural Topic Modeling (STM)
Game Plan:

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