Summer Institute in Computational Social Science at NDSU 2023

Topic modeling with R

Shuning Lu (Ph.D.)
Assistant Professor
Department of Communication
North Dakota State University

Email: shuning.lu@ndsu.edu
Twitter: @shuning_lu

Today’s agenda

Morning:

- Overview of topic modeling
- Topic modeling tutorial
- Group project check-in

Lunch

Afternoon:

- Work on group projects

What is topic modeling?

a method for unsupervised classification of text documents to find natural groups of items

Users do not need to supervise the model. Instead, the model works on its own to discover patterns and information that was previously undetected.

Family of unsupervised machine learning

Clustering: split the dataset into groups based on their similarities or differences ← Group assignment based on your indicated interest

Association rules: identify sets of items which often occur together

Dimensionality reduction: reduce the number of features in a dataset

Anomaly detection: identify unusual data points

Why unsupervised machine learning?

Quick and easy way to start (don’t need training set or annotation of data)

Find patterns in (un)structured data

Can see what human minds cannot see

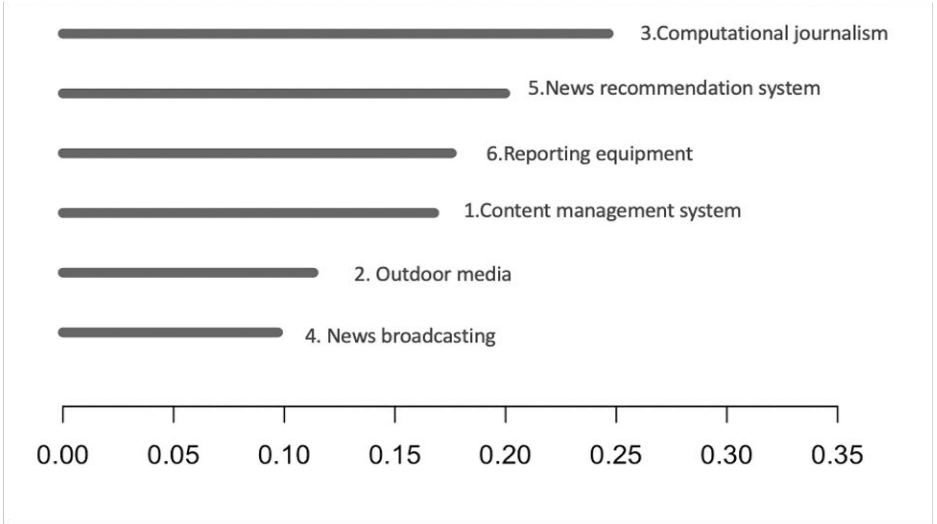
Especially useful for exploratory research

Applications of topic modeling

Any document collections

What is the document of interest in your discipline?

News technology innovation in China using patent data



Lu, S. (2022). News technology innovation as a field: A structural topic modeling analysis of patent data in mainland China. Communication and Society, 59, 147–175. (in Chinese)

Cons of unsupervised topic modeling

High risk of inaccurate results

- Domain
- Time

Human intervention to validate output

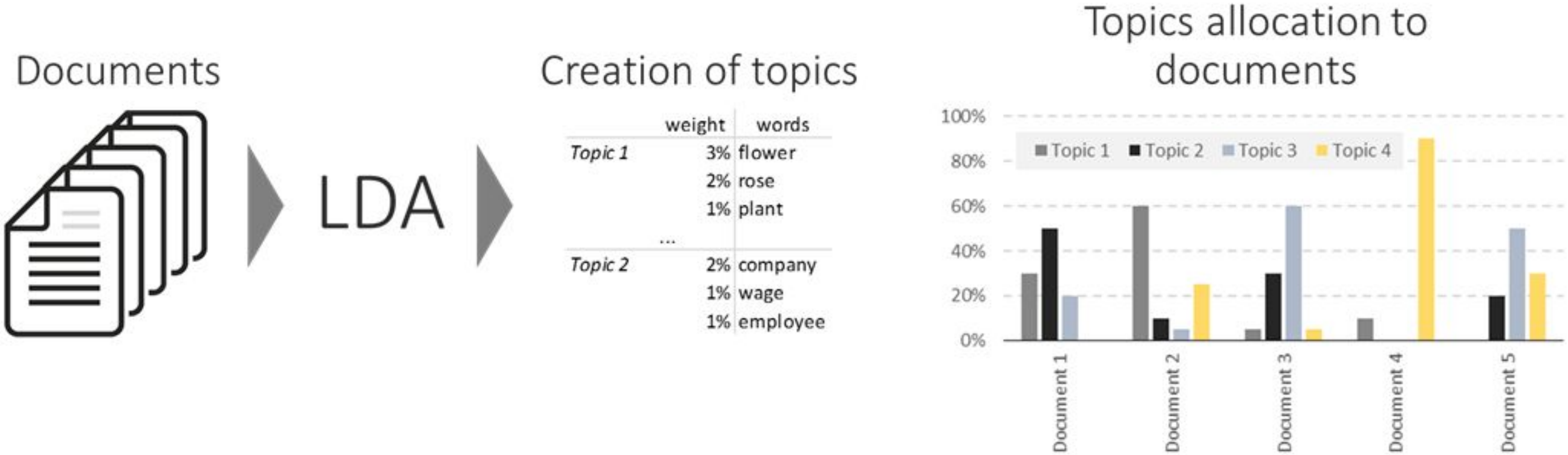
Lack of transparency into the basis on which data was clustered

Latent Dirichlet allocation (LDA)

Assumptions:

1. Every document is a mixture of topics
   - “Patent 1 is 90% topic A and 10% topic B, while Patent 2 is 30% topic A and 70% topic B.”
2. Every topic is a mixture of words
   - Computational journalism: Event, data, text, model, classification, analysis, topic, extraction
   - Reporting equipment: fixation, pole, board, interview light, support, utility, installation, installation

Latent Dirichlet allocation (LDA)



	weight	words
Topic 1	3%	flower
	2%	rose
	1%	plant
...		
Topic 2	2%	company
	1%	wage
	1%	employee

https://towardsdatascience.com/the-complete-guide-for-topics-extraction-in-python-a6aaa6cedbbc

What does LDA tell us?

Topic distribution: The distribution of topics in the document collection

Topic content: The distribution of words in each topic

Could we go beyond description of topics and words?

What if the topics are correlated with one another?

The correlational topic modeling (ctm) solution

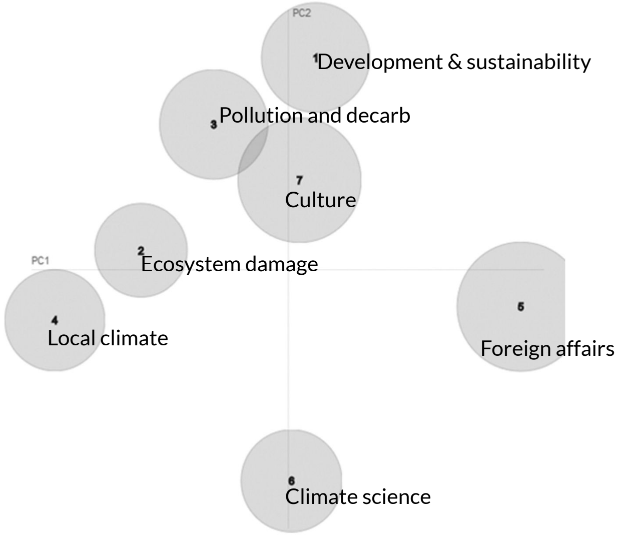


Figure 2. Jensen-Shannon Divergences between topics, visualized via multidimensional scaling.

Rabitz, F., Telešienė, A., & Zolubienė, E. (2021). Topic modelling the news media representation of climate change. Environmental Sociology, 7(3), 214-224.

Causal inference in topic modeling

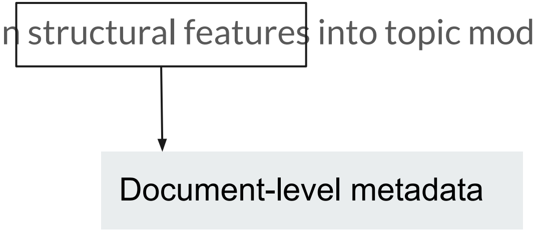
Does the distribution of topics change across time?

Will Republicans and Democrats view immigration issue differently?

How does partisans (R &D)’s views on immigration differ across age?

Structural topic modeling (stm)

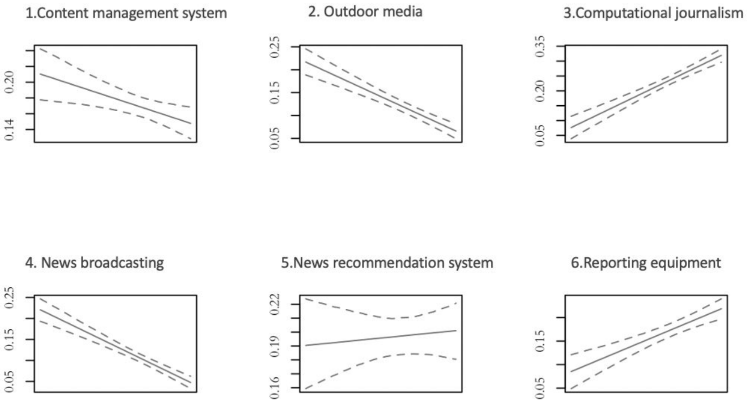
Bring in structural features into topic modeling



Document-level metadata

What is the document-level metadata related to your initial research idea?

News technology innovation in China using patent data



Lu, S. (2022). News technology innovation as a field: A structural topic modeling analysis of patent data in mainland China. Communication and Society, 59, 147–175.

Topic modeling workflow

Collecting/loading data

Pre-processing data

Model building (lda)

(Correlation of topics, ctm; Estimate effects, stm)

Interpretation ←→ fine-tuning

Visualization and summary

Tool time

Code along!