Text as Data Workshop

Lily Fesler

Center for Education Policy Analysis

lfesler@stanford.edu

May 29, 2019

Goals for Today

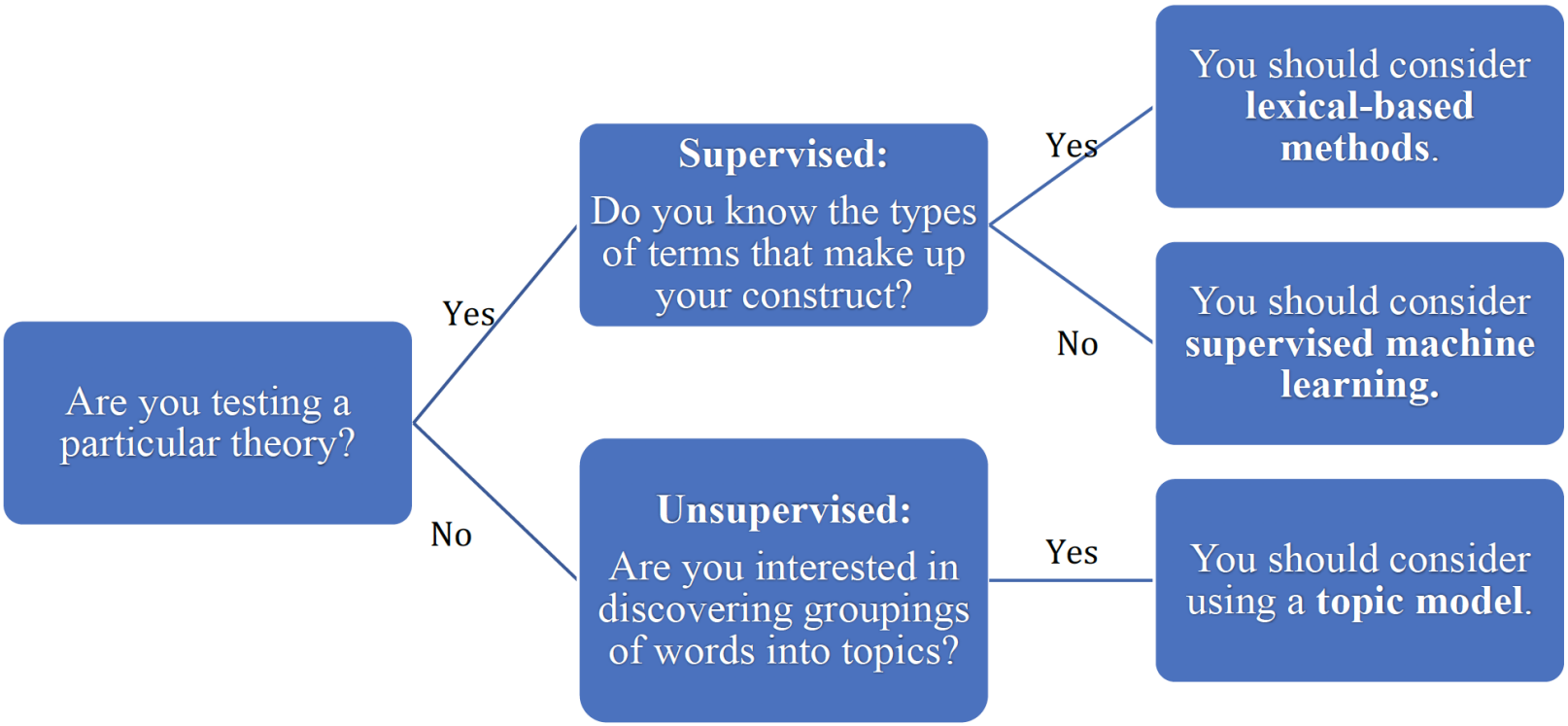
- Determine which method is most appropriate for a given research question (focused on categorization)
- Develop intuition for how dictionary methods, supervised machine learning, and topic modeling work
- Play around with some of these methods in R

Why analyze text quantitatively?

- There is an increasing amount of text data available for educational contexts: online discussion boards, written essays, policy documents, transcribed classroom audio, text messages, etc.
- There are also an increasing number of tools that have been developed in CS, computational linguistics and the social sciences to analyze this data (Gentzkow, Kelly & Taddy, 2017; Grimmer & Stewart, 2013)

Choosing a Method

Choosing a Method for Categorization



Which method?

- How frequently does teacher observation feedback fit into each of 7 predetermined domains? (Gegenheimer et al., 2018)
- How do teachers perceive their role in closing achievement gaps? (using teacher application essay data) (Penner, 2019)
- How often do teachers: leave teaching 1) involuntarily, 2) to avoid a bad job, 3) to approach a better job, or 4) for another reason? (Sajjadiani et al., 2019)
- How do low- and high-achieving students differ in how they talk about their long-term goals? (Beattie, Laliberte & Oreopoulos, 2018)

Converting Text into Data

- Most methods treat each document as a 'bag-of-words'
- Researchers typically convert the text to lowercase, remove all punctuation and numbers
- Researchers often also remove 'stopwords' (functional words like 'the,' 'a,' and 'she') but this varies based on analysis
- Researchers also often stem or lemmatize their words (remove the suffix or return the word to its root form)
- To readingtext.R!

Lexical-based methods

Lexical-based methods

- Researchers can come up with their own word lists based on a subset of documents (Baker, Bloom & Davis, 2016)
- Researchers can also use external word lists (dictionaries), like Linguistic Inquiry and Word Count (LIWC)
- Always validate dictionary methods (using a confusion matrix)! Words can have different connotations in different contexts
- Back to readingtext.R!

Supervised Machine Learning

Supervised machine learning

- Supervised machine learning (SML) relies on researchers to hand code documents to train a model
- SML (for categorization) is ideal when researchers can easily code a subset of documents but coding all of the documents would be too time consuming
- Can include many methods, like LASSO, ridge, support vector machine (SVM), random forests, and neural nets
- Also need to assess performance with SML using a confusion matrix or a similar method

Topic Models

Topic models

- Topic models are unsupervised in that the researcher does not choose the categories before estimation
- Topic models use the terms in the documents (which are observable) to estimate the topics being discussed (which are unobservable) by determining which words tend to co-occur. Each topic is defined as a distribution over the words in a vocabulary.
- Mixed membership models allow each document to contain multiple topics
- Structural topic models allow topics to vary by specified covariates.
- After estimation, researchers label the topics themselves.
- To stm_vignette.R!

Thank you!

Lily Fesler
lfesler@stanford.edu