CS 293/EDUC 473

Discovery & Exploration in Educational Text Data

Topic Modeling, Clustering, Grounded Exploration

Stanford
GRADUATE SCHOOL OF
EDUCATION

Quiz time

What are you most excited about?

Teacher Buddy!

Hands-on Project

Guest Visits

Build Tools that Teachers Actually Want

Discovery & Exploration in Educational Language Data

Student Simulations

Deploying/Testing Tools*

ambitious! only if your project is quite advanced already!

Generative AI for Tutor Support

What else you'd like us to cover?

Building Evals

Building Tools that Preserve Productive Struggle

Bias

Frontier Industry Tools

Multilingual Learners

Interactive Textbooks

Multilingual Models

What else you'd like us to cover?

Your final project?

Building Evals

Building Tools that Preserve Productive Struggle

Bias

Frontier Industry Tools

Multilingual Learners

Interactive Textbooks

Multilingual Models

What else you'd like us to cover?

Building Evals

Building Tools that Preserve Productive Struggle

Part of lecture(s), guest visits, readings

Bias

Frontier Industry Tools

Multilingual Learners

Interactive Textbooks

Multilingual Models

What else you'd like us to cover?

Building Evals

Building Tools that Preserve Productive Struggle

Bias

Frontier Industry Tools

Multilingual Learners

A bit beyond scope, but welcome to explore for final projects!

Interactive Textbooks

Multilingual Models

Today’s class

- Brief activity around teacher interview
- Debrief Dan Meyer visit; think ahead of TeachFX, LearnLM team visits
- Short lecture on topic modeling
- Reading discussion for Kubsch et al (2023) led by Ananya, Nick, Wanning

Group discussion on teacher interviews! (5 minutes)

- **[Context]** What did you learn about the day-to-day realities of teaching that you hadn’t considered before?
- **[Challenges]** What challenges or pain points did the educator describe?
- **[Opportunities]** How do these challenges relate to potential opportunities for NLP tools or interventions?
- **[Human-Centered Design]** How did the educator describe their experiences with technology in the classroom, and what did you learn about the design of tools for education?

Talk in small groups

- Did anything surprise you, or challenge your ideas from Dan’s talk/blog posts?
- Is there anything you disagree with?
- Thinking ahead of the visits by tech developers from TeachFX and Google, what might you want to ask them in light of Dan’s points?

Topic Modeling, Clustering, Grounded Exploration

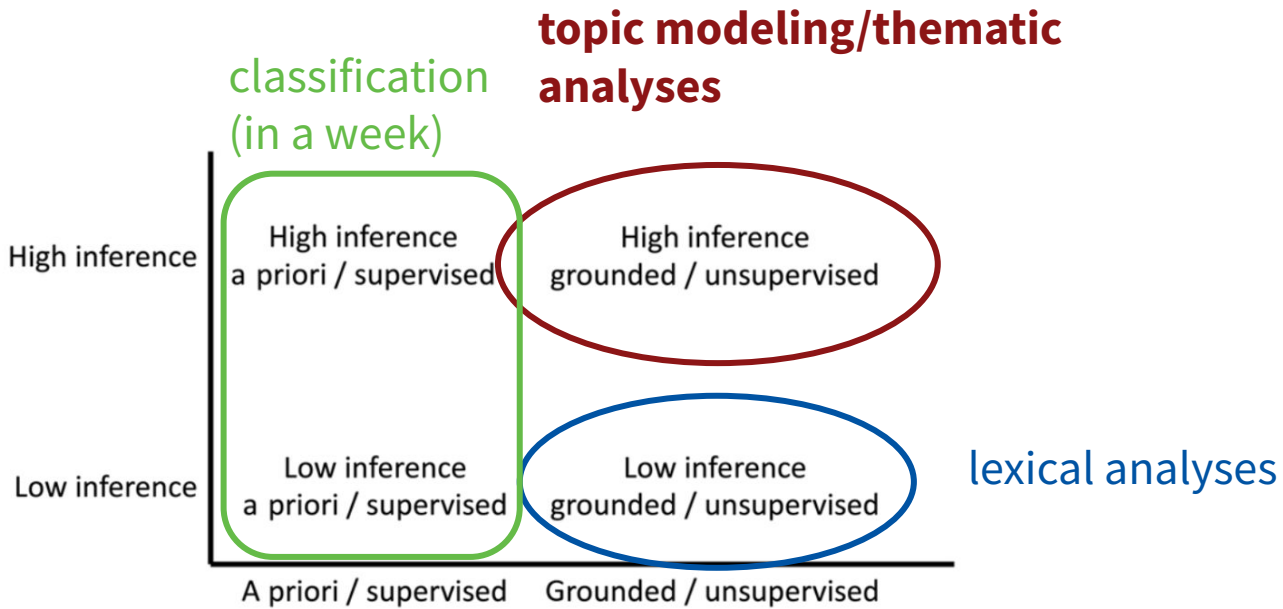


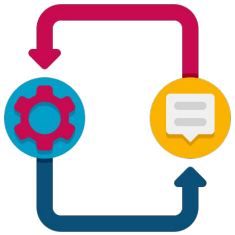
FIGURE 2 Epistemic functions and level of inference of four different kinds of tasks in assigning codes or numbers to data based on characteristics of the data and question

Kubsch et al. (2023)

When to use topic modeling & clustering?

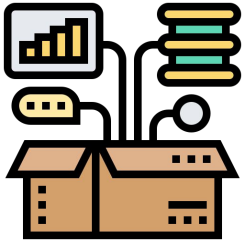
Discovery

Discovering **interesting or unexpected structures** that can be useful for hypothesis generation.



Synthesis

Comparing and synthesizing **qualitative** and **quantitative** analyses.

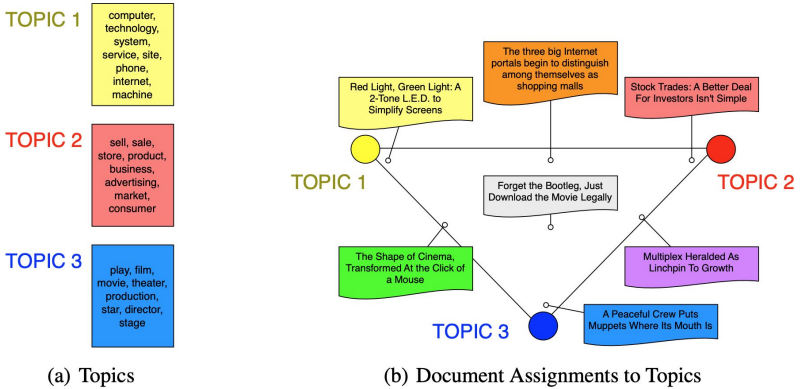


Featurization

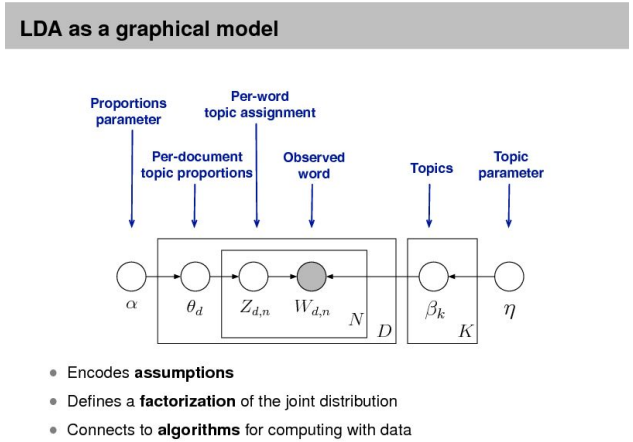
Using topics as **features** (representations) in downstream analyses, e.g. log odds, LLM-based analyses, classification.

Topic modeling

- There are several topic modeling approaches; in my experience, **LDA MALLET** (David Blei) works best for simple explorations
  - Original command line package
  - Python wrapper
- LDA outputs per-document topic proportions and per-topic word proportions
- Take a look at these slides here to learn more



(a) Topics (b) Document Assignments to Topics



Credit: Jason Yosinski

Pre-processing decisions for topic modeling **matter**

- Highest probability words within topics are often stop words (e.g. “I”, “not”, “will”) → **remove stopwords (?)**
  - Sometimes stopwords lead to meaningful distinctions (negations, use of 1st vs 3rd person)
- Forms of the same word may appear together as highest probability words (e.g. “like”, “liked”, “liking”) → **stem / lemmatize words (?)**
  - but: they may decrease stability (Schoefield & Mimno, 2016) and can lead to combining words with different meanings
- **Lowercasing?**
  - may collapse named entities with non-named entities; removes emphasis (e.g. all caps)
- (Denny & Spirling, 2017) introduce an R package **preText** to measure the impact of preprocessing on topics

Evaluating topic models: Reading tea leaves (Chang et al., 2019)

- Manual methods are the most reliable!
- Two evaluation tasks
  - **Word Intrusion**
  - **Topic Intrusion**
    - More challenging with longer texts
- Relatively quick once it's set up
  - For resource-efficiency, first select ~3 most promising parameter settings (num topics + preprocessing decisions) by eyeballing + automated metrics and compare those

Please select the word which is out of place or does not belong with the others.

◯ password ◯ help ◯ log ◉ woke ◯ account

Word Intrusion Task

Please select the group of words which is out of place or does not belong with the tweet.

Tweet: @SpotifyCares Keep getting Gateway error messages. Don't have time for this.

◉ lyrics; back; notifications; app; feature

◯ error; get; tried; message; says

Topic Intrusion Task

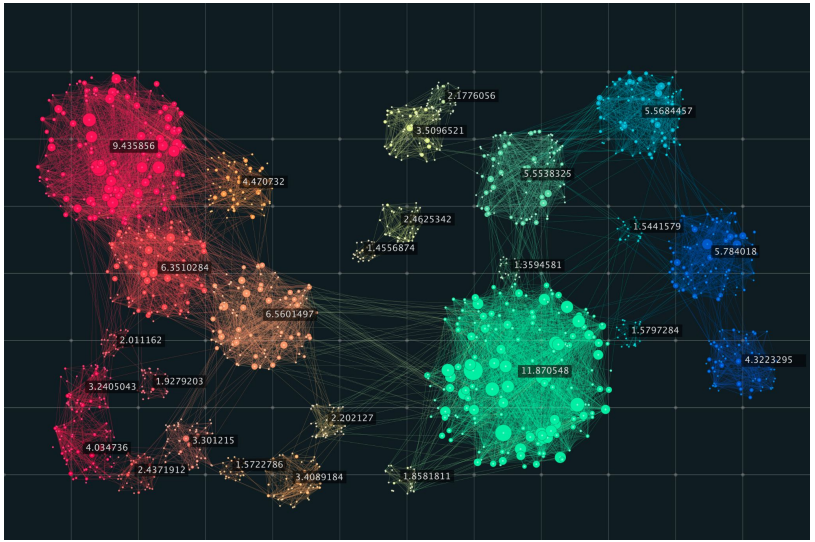
image credit: (Guzman et al., 2017)

Clustering

Group a set of data points into a number of clusters, so that

- Data points in the same cluster are similar to each other
- Data points in different clusters are dissimilar

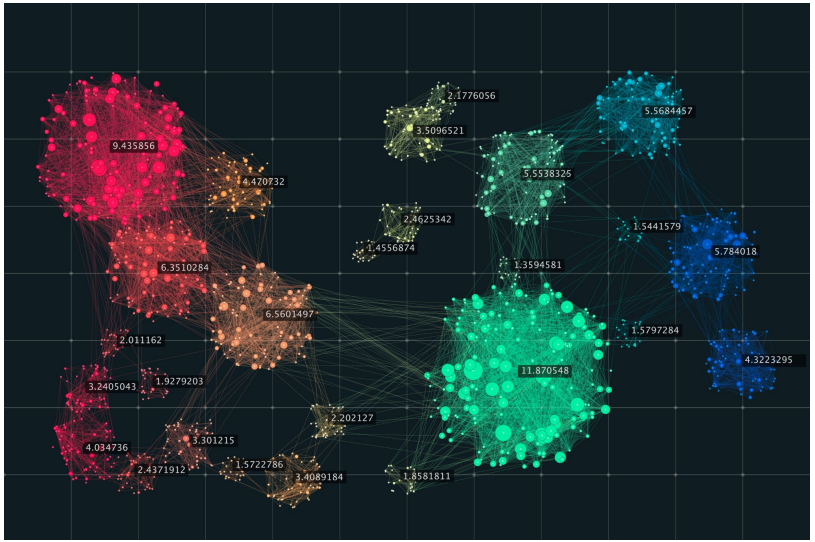
Use only X, not Y → unsupervised method



Slide credit: Diyi Yang

Cluster structures

- Partitioning a group of data point into K disjoint sets (K-means clustering)
- Assigning X to hierarchical structures (Hierarchical clustering)
- Assigning X to partial membership in K different sets (Graphic models, GMM)
- Learning a representation that puts similar data points closer to each other (Word embeddings w/ deep learning)



Slide credit: David Bamman & Diyi Yang

Word embedding based representations

- Simple but tough to beat baseline sentence embeddings
  - tf-idf weighted average of word embeddings
- Sentence transformers (see HW1)

Clustered **tutor responses** (Wang et al., 2023)

```
Cluster  11
['This is our first question.', 'Here is the next one.', 'Here is your first question.', 'This is our next question.',
"Let's get directed to the next question.", "Let's move on to the next question.", 'Here is our first practice question.
', "Let's get directed to the next question.", 'Let us get directed to the next question.', "Let's get directed to the
next question.", 'This is your next question.', 'Here is our first question.', 'This is your first independent question.
', 'Here you go for the next one.', 'Here is your next question.', 'This is our first question.', 'Here comes the next
```

```
Cluster  7
['That was a good try.', 'That is a very good start.', "That's good.", "That's good.", 'Very good.', 'Great!', "That's
fantastic.", 'Well done.', 'All the best.', 'Excellent.', 'Great going.', 'That was a nice try.', "Let's get started.",
'That was a great start.', 'Now try this one.', 'Awesome!', "Let's get started.", 'This is a very good start.', "That's
great.", "That's great!", 'That was a good start.', 'Great try.', 'That was a great try.', 'That was a good try.', 'Good
work!', 'That was a good try!', 'That was a good try.', "Let's get started.", "That's good.", 'Nice One!', "Let's get
started.", "That's good.", 'Good try.', "That's great.", 'Good job!', 'Lets get started!', 'That was a good start.',
```

What to think about when doing clustering?

- How to **represent** each data point?
- How to **calculate the similarity** between data points?
- What is the **number of clusters** to use?
- How can we **evaluate** the resulting clusters?

Slide credit: Diyi Yang

Large Language Models for thematic analyses

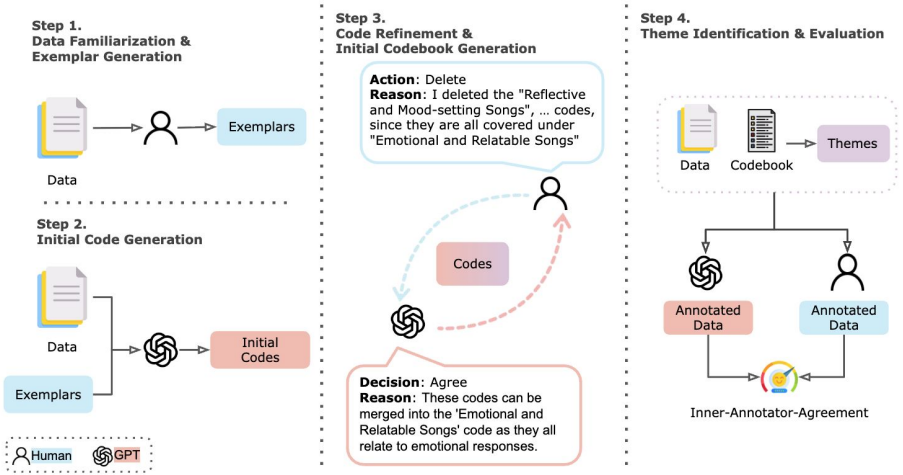
E.g. Thematic Analysis (Gamieldien et al., 2023, Dai et al., 2023), Topic Modeling (Wang et al., 2024, Kapoor et al., 2024)

Pros:

- Can do very well at producing highly interpretable topics with good coverage

Downside:

- Hallucinations
- Difficult to control level of specificity
- Black box process; the outputs require extensive evaluation to ensure they're accurately representing the data



[Dai et al, 2023]

Topic modeling can support downstream supervised learning

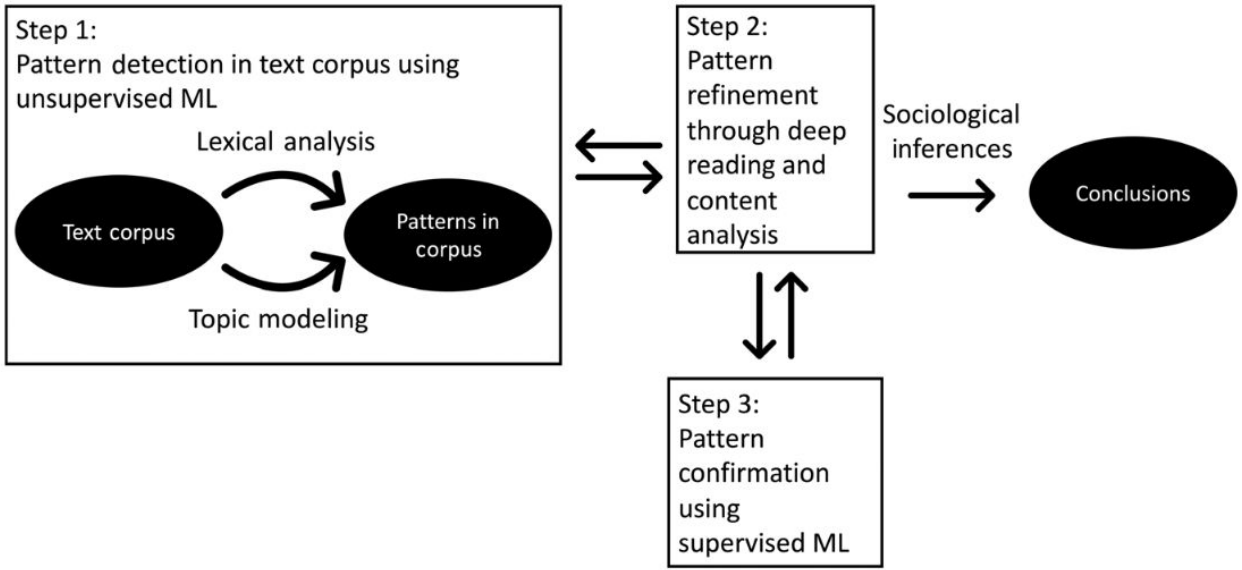


FIGURE 4 Analysis workflow from Nelson, 2020

Student takes from last year:
Topics/clusters are very hard to interpret!

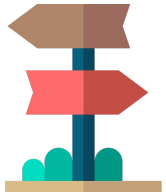
Cons of **LDA**

- word co-occurrence may not always equal thematic consistency
  - math topics “blend together”
  - classroom management + math-related themes co-occurring
  - function words, transcription issues add a lot of noise
- difficult to interpret topics
  - words are removed from context
  - need domain knowledge
- hard to steer
  - not clear how to identify right number of topics
  - processing data in a way that leads to cleaner topics is nontrivial

Cons of **clustering**

- very sensitive to surface-level lexical similarity
- very sensitive to length
- clusters become too fine-grained quickly (long tail…)

Our response (and some of yours, too…)



- processing the data and choosing the right model can make a big difference
  - try better models
  - subsetting the data, removing irrelevant words can help improve them
  - pay attention to document length!
- trade-off between output interpretability (LLMs win here) vs model interpretability (LDA, clustering win here)
- sometimes you *surface level similarity* may be what you want to capture (example later)

Reading Discussion for Kubsch et al. (2023)