CS 293/EDUC 473

Discovery and Exploration in Educational Text Data

Stanford
GRADUATE SCHOOL OF
EDUCATION

Does anyone have questions about the syllabus / assignments?

📣 Announcements

- Join the Ed forum!
- For projects - Office hours
  - Mei will have office hours **after today's class (4:20-5:30; in this room)**
  - Dora updated OH: Mondays 9:30-11:00. send email if this time doesn't work for your project team
- First reading commentary due **Sunday at 5pm** for Lucy et al (2020).
- Complete teacher interview, discussant signup and get to know you survey **by next Wednesday**.
- **HW1 due Tuesday Jan 21 at 11:59pm**. We will discuss the HW logistics at the end of today's lecture.

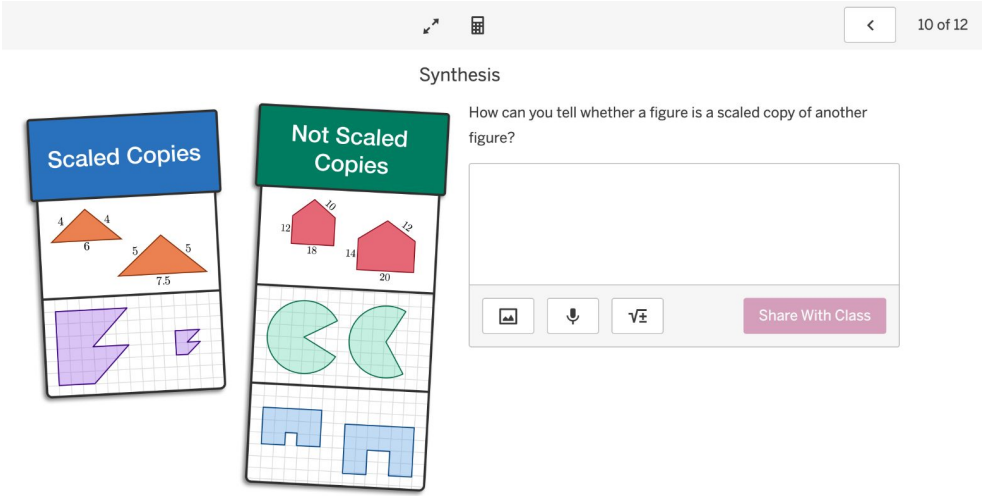
Today’s class

- More info on default project
- Watching a classroom observation video clip
- Small group discussion
- Lecture on data exploration
- HW1 prep

Default project using Amplify Data

Great opportunity for understanding **students' learning trajectories in response to dynamic tasks**! From the technical perspective, it can teach you:

1) how to analyze messy student language data
2) how to study dynamic (time-based) data
3) optionally: multimodal data analyses



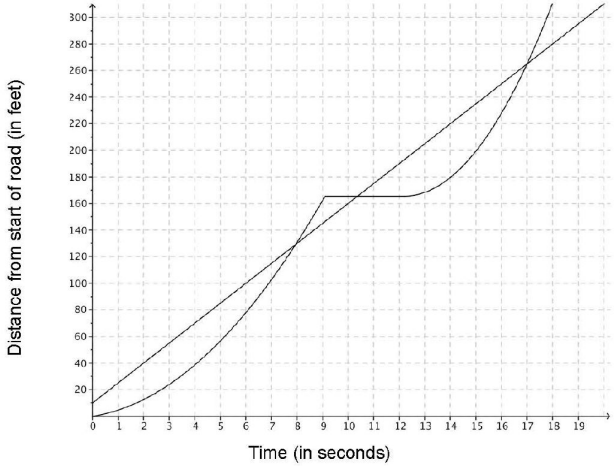
If you're interested, please complete CITI training (IRB NonMedical Research) **asap** and email me with your certificate!

Classroom observation video

Link

Bike and Truck

A bicycle traveling at a steady rate and a truck are moving along a road in the same direction. The graph below shows their positions as a function of time. Let B(t) represent the bicycle's distance and K(t) represent the truck's distance.



1. Label the graphs appropriately with B(t) and K(t). Explain how you made your decision.

Discuss in groups

1. What do you notice?
2. What do you wonder?
3. How might you imagine an NLP tool might support this teacher?

If you had access to thousands of transcripts like this, how would you begin analyzing them?

Example features from Liu & Cohen (2021)

TABLE 2

Computer-Generated Metrics on Teacher Practices

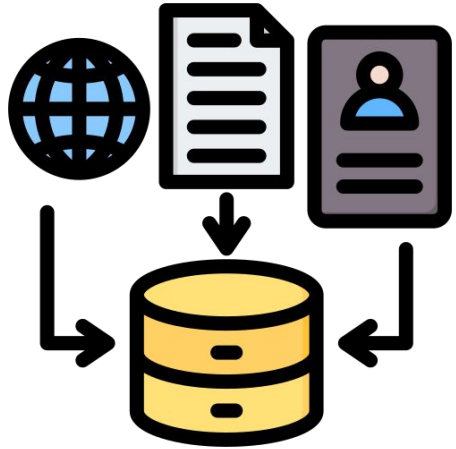
Variable	*M*	*SD*
Turn-taking		
Turns per minute	4.50	(2.08)
Proportion of time teacher talks	85.22	(10.90)
Average words per minute	115.45	(24.70)
Targeting (teacher)		
"You" (%)	4.76	(1.55)
"I" (%)	2.51	(1.17)
Analytic and social language (teacher)		
Analytic thinking	38.20	(12.39)
Social words (%)	13.71	(2.22)
Language coordination		
Language style matching (0–1)	0.80	(0.10)
Questioning (teacher)		
Open-ended questions per minute	0.22	(0.12)
Allocation of time between academic content and routine (teacher)		
Routine language (%)	10.63	(5.91)

Note. All statistics are calculated at the level of teacher video. Analytic thinking is a composite score that is converted to percentiles.

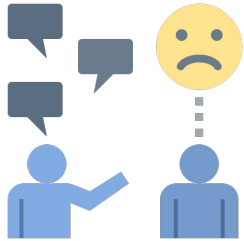
Data Exploration

What counts as data?

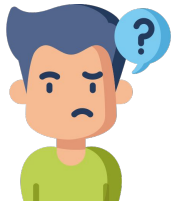
- Qualitative data
  - Conversations & interviews with stakeholders (teachers, students, peers, researchers)
  - Results and insights presented by related work
  - Surveys
- Quantitative data
  - Text data from the target domain, e.g. classroom transcripts, text books, lesson plans
  - Metadata associated with the text, e.g. demographic information, learning outcome data, satisfaction ratings



Things you learn from qualitative information gathering



“We already think we know that”
“That’s too naive”
“that doesn’t reflect social reality”
“Text analysis is unlikely to answer that question”



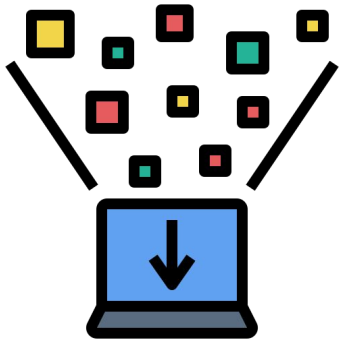
“Two major camps in the field would give different answers to that question”

“We tried to look at that back in the 1960s but we didn’t have the technology”
“That sounds like something teachers would love”
“That’s a really fundamental question”

How we do things with words (Nguyen et al., 2020)

Obtaining quantitative data

- **Collecting brand new quantitative data is beyond the scope of this course**
- You’re welcome to bring your own data
- Use default course data (Amplify or NCTE data)
- Use other existing datasets



Challenges in obtaining & using educational text data

- Privacy
- Noisiness
  - Transcription errors
  - Data often requires lots and lots of cleaning
- Sparsity
  - Especially for self-reported data
- Bridging Expertise (Pedagogical and Technical)
- Operationalizing Ambiguous Concepts

Existing data sources

Text corpora:

- Open Syllabus Project (old Github version)
- Teacher-student Chatroom Corpus
- Coursera Forum Dataset
- CIMA
- TalkMoves Dataset
- DRYAD dataset
- PERSUADE dataset of student writing
- KhanAcademy Math Tutoring Dataset
- DLP Catalog of Math-Teaching Related Datasets

Pedagogical resources:

- Middle school math misconceptions
- Achieve The Core
- TLE dataset of recordings
- MQI video library

Browse the following conference proceedings: BEA, LAK, EDM for more datasets

For your review

Building a solution requires **having the right data & understanding that data**

- ➔ What are the ethical or legal considerations for using this data?
- ➔ Do you have access to your desired sample (size)?
- ➔ What is the distribution of the data? Is it representative of your target population?
- ➔ How clean is the data?
- ➔ How do variables in the data relate to each other?
- ➔ Do you have the required outcomes for answering the research question / estimating impact?
- ➔ What hypotheses or assumptions can be made based on initial exploration?

For your review

Identifying a research question

Think about these questions as you'll need to address them in the Project Rationale.

- ➔ Who is waiting for the solution / answer to your question? What would this solution / knowing the answer would change, both in your field of study and in the wider world?
- ➔ Are these questions answerable with text?
- ➔ Why is computational text analysis necessary or valuable for this solution / answering this question?
- ➔ Do you have access to the data that will support these research questions?
- ➔ What are the ethical implications of your research / solution? Who will be affected by decisions made based on your solution / results?

How we do things with words (Nguyen et al., 2020)

For your review

Guiding questions for conceptualization

- ➔ What are the core concepts you are addressing? And are you being true to their core meaning?
- ➔ What are competing definitions? Which is best suited to the task and why?
- ➔ Does the systematized concept you've selected reflect an adequate understanding of the background concept?
- ➔ How do domain experts approach the topic? Does your research connect to this wider context? Have you considered relevant methods and theories in other domains?
- ➔ Is it possible to speak of "ground truth" for the concept(s) in question?

How we do things with words (Nguyen et al., 2020)

For your review

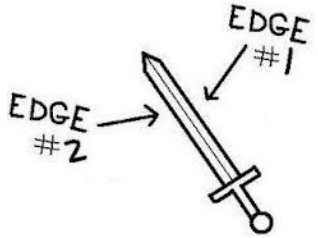
Guiding questions for data exploration

- ➔ Are sources representative? Are they disproportionately of one form? Are all relevant time windows covered? Does the data represent all relevant groups, including those often marginalized?
- ➔ When metadata is available: Are there errors, inconsistencies, biases, or missing information? Is this quality of metadata consistent across the dataset, or are some parts better or worse?
- ➔ When labels are available: How were the labels created? Do the labels actually mean what you are using them to represent?
- ➔ If you are filtering, subsampling, or selecting from the original data, is the remaining subset representative? Can you describe how selective removal alters the data and the interpretation of the data? Are you losing anything that might be valuable at a later stage?
- ➔ Who created the data, and do they have agency over its use? Should this data be used for research? How does respect for document creators affect how you conduct and share your research?

How we do things with words (Nguyen et al., 2020)

Think about dual use

- What if the tool was used in any type of high stakes decision making?
- Could the tool be used to surveil teachers?
- Could it be used to punish students or teachers?
- Could the tool be used to harm vulnerable populations, aggravate biases and inequities?

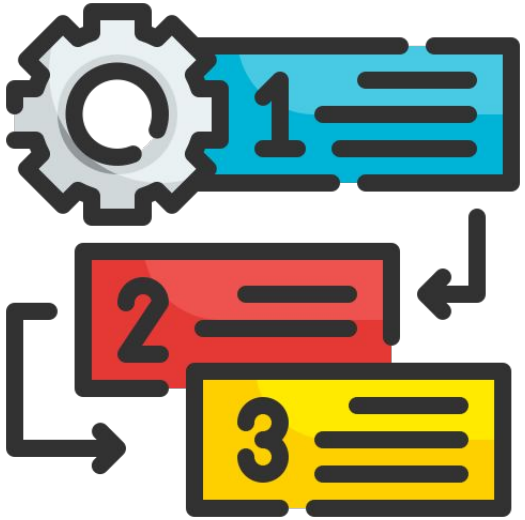


How we do things with words (Nguyen et al., 2020)

Best practices for data exploration

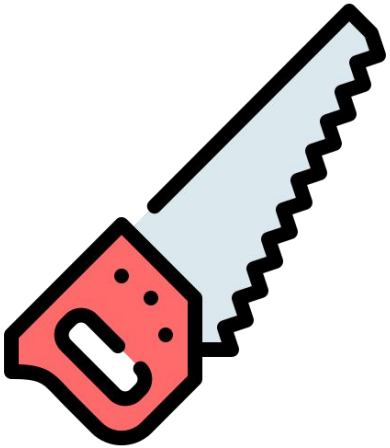
#1 **Prioritize.**

- It’s easy to get lost in the weeds — pause to **take a step back** and keep your core question / goal in mind
- Each separate measure requires a lot of resources to develop and validate (e.g. each pedagogical practice).
- Focus on constructs that seem to be the **highest leverage** and can serve as proxies for other important constructs (e.g. via lit review)



#2 **Don’t be afraid of doing a LOT of manual work.**

- Automating tasks is sometimes more work and less precise than doing the clean-up manually
- When you need data for validation, it’s usually best to clean it manually to avoid circularity
- Close reading and qualitative coding is often the best way to understand the contents of your data and do error analyses. Don’t be afraid to do that a lot even if you’re not an expert.

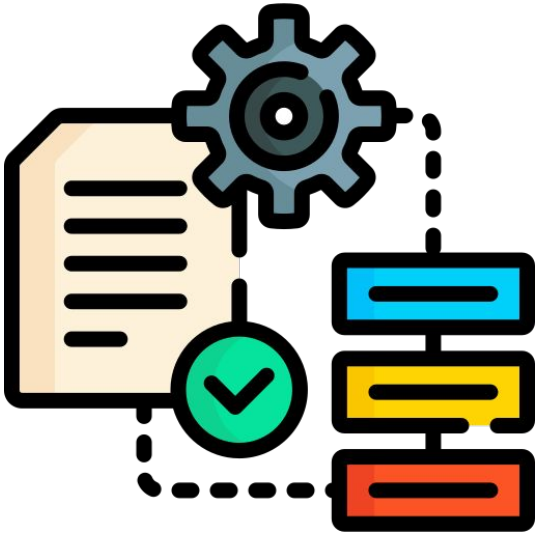


#3 **Create visualizations.**

- Visualizations are oftentimes best way to understand the distribution of your data
- Graphs and plots are usually much better at explaining patterns than regressions
- Visualizations can help a lot with debugging, too

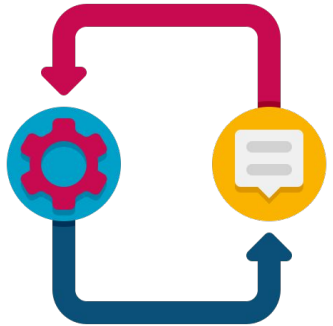
#4 **Don’t be hand-wavy about pre-processing.**

- Preprocessing decisions (e.g. removing stopwords) can make a big difference (see HW1)
- Explore how pre-processing decisions affect your data
- Often it’s helpful to report the core analyses with different pre-processing decisions in your results (e.g. in supplement)



#5 **Reinforce the feedback loop.**

- Use insights from your explorations to finetune your goals / research questions
- Seek feedback from educators after you've looked at the data
    - Best is to show them some data and ask what they observe

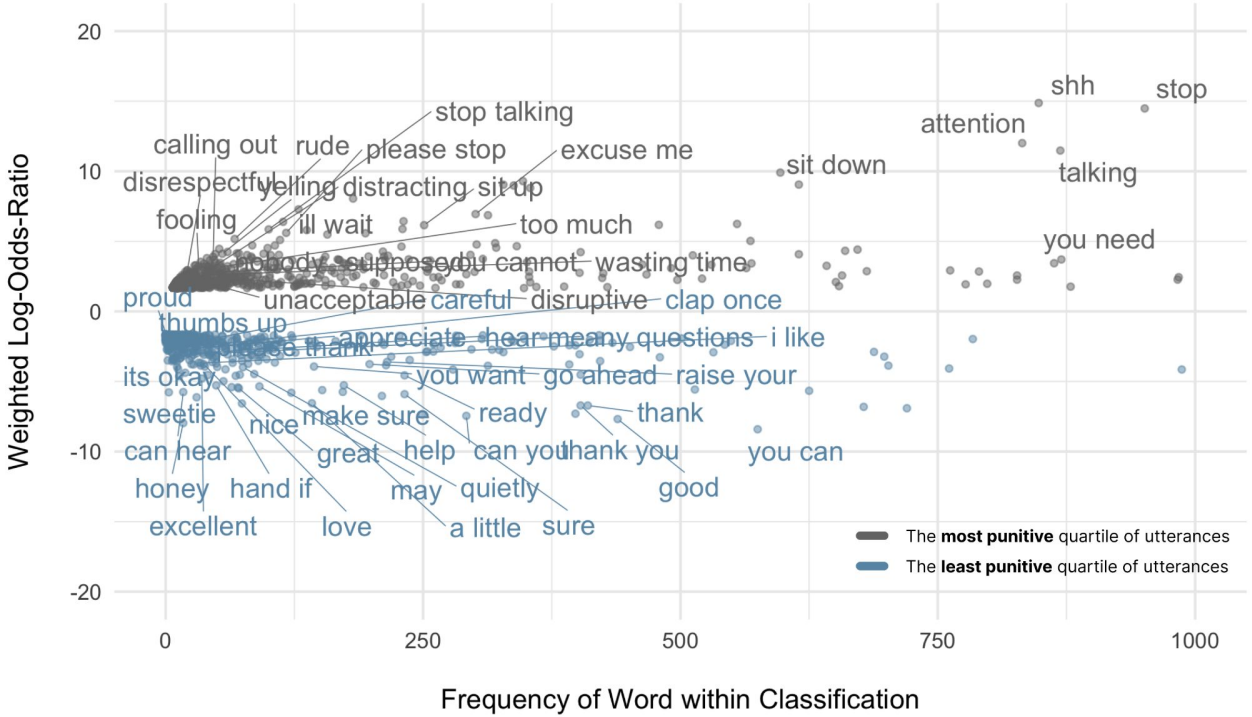


#6 **Start with the most interpretable methods first.**

- It’s hard to “debug” and interpret methods with multiple layers of abstraction
- **Words** often explain a lot of the variance — lexical analysis is often the best starting point

My favorite data exploration method!!! Script included in HW1 (credit to Dan J)

Z-scored log odds ratios (Section 4.3 in Monroe et al., 2017)



(Tan & Demszky, 2023)

Z-scored log odds ratios

It answers **how word usage is different along a particular dimension?**

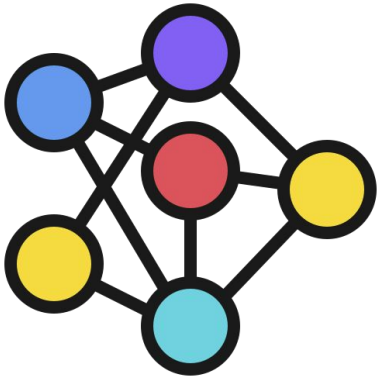
1. **Sample two different groups from the data** (e.g. teacher utterances from transcripts with high ratings for instruction quality vs ones with low ratings) + create a third group (**prior**) that includes your entire dataset (e.g. all teacher utterances in data)
2. **Obtain counts** for words / phrases in the data (you can use it to count any other feature, too, e.g. lexical categories);
    a. see Demszky et al. (2019) for the use of the log odds method on different features (LIWC, topics)
3. **Compute the z-scored log odds ratios** for each word / phrase. Positive values indicate association with group 1 and negative values indicate association with group 1. Magnitude represents number of standard deviations (discard those with < 1).

#7 **Be scientific about debugging.**

- If you (don't) observe something, that can be due to many things
  - Noise in the data
  - Imprecise definitions of your construct
  - Imprecise measures
    - e.g. lexicons or ML classifiers are not perfect
- Be systematic to explore all the possible causes of both positive and negative results

#8 **Triangulate several data sources as much as possible**

- E.g. self-reported data + student outcome data + language features
- Doing so can help
  - provide a more nuanced understanding of relationships in your data
  - validate measures
  - corroborate hypotheses
  - debug issues



Tools for data exploration

- **General Python packages (ChatGPT may be your best friend :))**
  - pandas (data wrangling)
  - statsmodels (regressions, although it's better to do them in R/Stata)
  - scipy (e.g. for t-tests, correlations)
  - seaborn (for visualization)
- **Lexical analysis:**
  - log odds method to identify words/phrases that distinguish two groups (see homework)
  - lexicons (e.g. NRC valence arousal dominance lexicon, concreteness lexicon)
- **Unsupervised methods (next class)**

Relevant resources

- Dirk Hovy's Github repository
- Introduction to Cultural Analytics by Melanie Walsh
- DLATK command line text analysis tool
- Text analysis tutorial on scikit-learn
- Computational text analysis course by Adam Poliak
- NLP + CSS tutorials by Katie Keith and Ian Stewart
- StatQuest Youtube Channel
- Computational and Inferential Thinking

Homework #1 Setup

Instructions for HW1

Homework assignments are intended to be hosted on Colab.

- Go to Canvas.
- Move to GDrive.
- Work on HW1 through Colab!
- Upload Colab/notebook as **PDF** to Canvas.

If you have any questions, please post on Ed Discussions! I'll try to promptly respond to them.