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 OH: Tuesdays 10:30 - 11:30am, Thursdays 2-3pm -- ANKO lobby
  - Dora OH: Thursdays 4-4:40pm, Fridays 3-4pm
- First official reading quiz next Tuesday
- Complete discussant signup and get to know you survey **by Friday.**
- **A0 due Monday, Jan 19 at 5pm**. Let us know if you are stuck with teacher recruitment!
- Mei will introduce the datasets in today's lecture!

Today’s class

- Practice reading quiz
- More info on course datasets and A1
- Watching a classroom observation video clip
- Small group discussion
- Lecture on data exploration

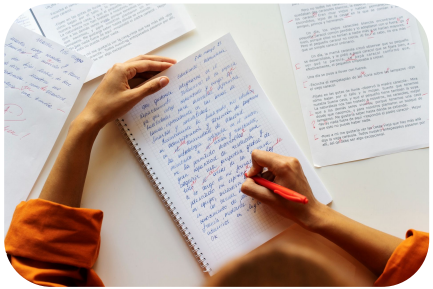
Did you read today's readings? Let's start with a practice quiz!

- Go to Canvas
- You should see a practice quiz among the Quizzes
- 5 minutes
- Not graded!

Four Language Datasets

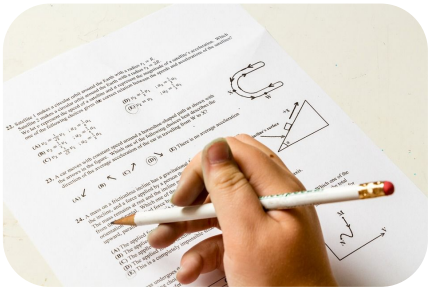
Classroom Observation Transcripts

- Grades 4-5
- Math
- N = 1660
- LOTS of metadata!
- Models of instructional language



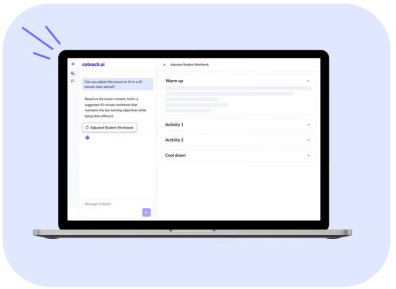
Persuasive Essays from Students

- Grades 6-12
- ELA
- N = 14K
- Several *like datasets!
- Mostly AES, fairness audit



Math Problems from Big Curricula

- Grades K-12
- Math
- N = 21K
- With 18K images!
- CCSS tagging, item generation



Teacher-AI Lesson Planning Chat Logs

- Grades K-12
- Math
- N = 10K threads
- With LLM responses!

Assignment Objectives

- Use NLP to understand and support teaching and learning
    - Inform both education theory
    - Inform modeling / edtech development
    - Teach each other what we discover
- E2E research project
    - With a landmark dataset
    - Driven by inquiry, exploration, analysis
    - Practitioner involvement at every stage of ML pipeline
- Build the components of your final project

4 (Cool??) Assignments

A0 Orientation
(Intro + Data)

- Pick a dataset
- Find a team
- Recruit a teacher buddy
- Start thinking :)

A1 Exploration
(Intro + Data)

- Data cleaning
- Exploratory analyses
- Lexical / unsupervised methods
- Case memos

Data Collection

- Develop a new method or model via new annotated data
- Validate / relate new model with existing ones
- Analyze in context

Modeling + Validation

Measurement

- Create a small validation set
- Apply an existing method or model
- Analyze in context

Design

- Co-design a tool with your teacher buddy

Which track do I choose?

- Based on your construct of interest
  - **Model Track:** complex phenomena not previously (well) modeled
  - **Tool Track:** validated or commonly understood phenomena ready for intervention
  - Cross this bridge when you get there!
- Based on skills you'd like to practice
  - **Model Track:** data annotation / modeling / validation
  - **Tool Track:** UX design / software development

A word on ChatGPT

- Research is an enterprise in communication.
  - We all have a “built-in, shockproof, sh*t detector” (Hemingway).
- Coding?
  - Go for it. But make sure you understand and can justify/explain the methods you use.
- Writing?
  - ChatGPT writing is maximally academic-pretentious and minimally informative.
  - Writing is “a repetitive, obsessive, iterative application of preference: watch the needle, adjust the prose, watch the needle, adjust the prose… When we [do this], this manifests to the reader as evidence of care. (We might say that a reader is able to intuit the many less-cared-for versions of a sentence behind the one the writer let stand.)” (George Saunders).
  - You are graded on how we receive your writing; show us the care :)

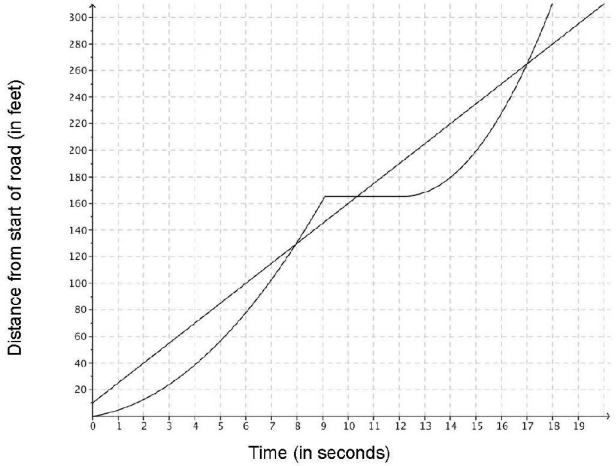
How might we support teaching by analyzing discourse in the classroom?

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

- What do you notice?
- What do you wonder?

- What are some moments of great teaching? Which teaching “actions” can you observe from the teachers’ words?
- What are some great moments of learning?

- How might an NLP tool support this teacher?
- If you had access to thousands of transcripts like this, how might you begin analyzing them?

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

Example features from Liu & Cohen (2021) → reading for next Tuesday!

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.