Edu-ConvoKit
An Open-Source Library for Education Conversation Data

Rose E. Wang

rewang@cs.stanford.edu, Stanford University

There's a lot of insight about how students learn in education conversations. But it's painful to process and analyze.

Edu-ConvoKit is an open-source library that handles all of that for you: pre-processing, annotation, and analysis.

```
pip install edu-convokit
```

```
https://github.com/stanfordnlp/edu-convokit
```

Why Edu-ConvoKit?

Motivation. Language is central to educational interactions, ranging from classroom instruction to tutoring sessions.

Challenges. No centralized tool & high learning curve for performing computational analysis! From interviews:

- "What's the best way to de-identify the data?"
- "I want an easily accessible collection of language tools that can detect insightful things."

Edu-ConvoKit

Edu-Convokit is a modular and end-to-end pipeline for:

Pre-Processing. The `PreProcessor` entity modifies the text, such as through deidentification. This is important to maintain the confidentiality of participants, while preserving the context of interaction.

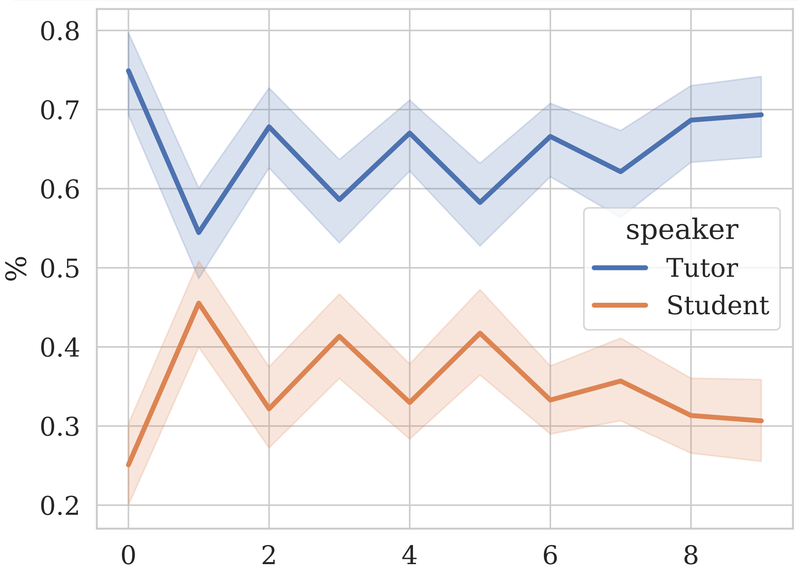
```
# Original data
>> print(df)
    text
0   My name is Alice Wang.
1   Hey Johnson, this is John.
>> processor = TextPreProcessor()
>> df = processor.anonymize_known_names(
            df=df,
            text_column="text",
            # from e.g., classroom roster
            names=["Alice Wang", "John Paul", "Johnson P"],
            replacement_names=["[T]", "[S1]", "[S2]"])
# Processed data
>> print(df)
    text
0   My name is [T].
1   Hey [S2], this is [S1].
```

Annotation. The `Annotator` entity annotates features at an utterance-level. It currently supports several measures: talk time, math density, student math reasoning, etc.

```
>> annotator = Annotator()
>> df = annotator.get_talktime(df=df, text_column="text",
output_column="talktime", analysis_unit="words")
>> print(df)
     speaker   text                                                talktime
…
47   Student   Miles, and then at B, it stops for they stop for […]    48
48   Tutor     Cool. that's I understand how you're thinking […]       27
49   Student   Cause the graph is, it says distance. This is fifty […] 56
50   Tutor     Okay and C to D, I'm sorry, D to E is Kirby. Does […]   31
…
```

Analysis. The `Analyzer` entity support common analyses in education conversation research, such as the temporal analysis:

```
>> analyzer = TemporalAnalyzer(DATA_DIR)
>> analyzer.plot_statistics(speaker_column="speaker", feature_column="talktime",
  num_bins=10)
```



Resources

You can find all these resources on `https://github.com/stanfordnlp/edu-convokit`!

Need a step-by-step walkthrough of the features?. We have Colab tutorials of each entity.

Need examples of Edu-ConvoKit applied to real datasets? We have Colab notebooks where we applied Edu-ConvoKit to NCTE (classroom), TalkMoves (classroom) and Amber (tutoring) datasets.

Need references for Edu-ConvoKit's features and applications? We have a paper database on our GitHub repository.

Need library documentation? Yep, we gotchu too: `https://edu-convokit.readthedocs.io/`.

Can find a feature that you need? Contribute to Edu-ConvoKit or reach out!