

CS 293/EDUC 473

Discovery and Exploration in Educational Text Data

Does anyone have questions about the syllabus /
assignments?



Announcements

- Join the Ed forum! Follow Canvas announcement instructions.
- For projects - Office hours
 - Rose will have extra office hours **after today's class (4:30-5:50; in this room)**
 - Dora will also add extra office hours this week (**Thursday 3-4; CERAS 536**)
- Reading commentary & Discussion
 - First reading commentary due **Tuesday at 5pm** for (Liu & Cohen, 2021).
 - Discussant leaders will receive commentaries from Rose by Tuesday 5:15pm.
 - **We're looking for 1 volunteer** to join Jürgen as a discussant.
 - We'll assign students who have not yet signed up tomorrow morning.
- **HW1 due next Tuesday 11:59pm.** We will discuss the HW logistics at the end of today's lecture.
- [Sarah Johnson](#) from the Teaching Lab will join the beginning of Wednesday's class!

Today's class

- Course-related Q&A
- Empathy Mapping
- Short lecture about Data exploration
- Q & A about textbook paper (Lucy et al., 2021) led by Rose
- HW1 prep





Empathy Mapping Exercise

1. You'll be randomly assigned to groups of ~2-3
2. Pick 2 questions to think about for your persona
3. Share in a roundtable with the group

Personas (credit: ChatGPT)

Name	Age	Role	Setting	Background
Maria Alvarez	28	High school science teacher	Urban public school with limited resources	Recent graduate, passionate about climate change, struggles with classroom management.
David Chen	40	Middle school math teacher	Well-funded suburban school	15 years of experience, technologically savvy, trying to make math engaging for all students.
Beatrice Okoye	55	Elementary school principal	Inner-city school with a diverse student population	Former English teacher, focused on improving school literacy rates, faces budget cuts.
Aaron Singh	33	Special education teacher	Rural school district	Trained in inclusive education, struggles with lack of specialized resources, keen on fostering individualized learning.
Sarah El-Khoury	29	High school history teacher	International school in a major city	Expatriate, adjusting to a new curriculum, balancing cultural sensitivities.
Liam O'Connor	47	Physical education teacher	Large public school	Former semi-pro athlete, promotes fitness amidst rising student obesity rates, struggling with students' increasing screen time.
Aisha Patel	35	Elementary school music teacher	Charter school focused on the arts	Trained in classical Indian music, integrating global music traditions, facing potential program cuts.
Samuel Akoto	52	College professor in Anthropology	State university	Published author, adjusting to online teaching, keen on fostering critical thinking in a polarized society.
Elena Rodriguez	31	School counselor	Mixed-income middle school	Passionate about mental health, overwhelmed with increasing student caseload, seeking ways to provide support virtually.
Hiroshi Takahashi	43	ESL (English as a Second Language) teacher	Community college	Fluent in multiple languages, working with diverse age groups, challenges in bridging cultural gaps.

Empathize with your persona (8 mins)

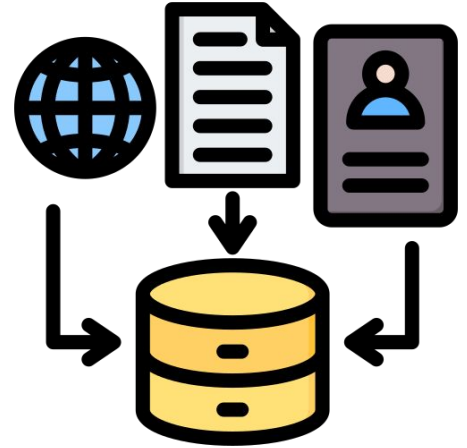
1. **Think & Feel:** What do you think they think and feel on a daily basis (personally and professionally)?
2. **Hear:** What does this persona hear from colleagues, students, parents, or administrators?
3. **See:** What does this persona see in their environment? This could be resources, lack of resources, technology.
4. **Say & Do:** What actions does this persona take? What do they usually say in their professional setting?
5. **Pain Points:** What challenges or obstacles does this persona face? Consider both emotional and practical aspects.
6. **Gains:** What are the aspirations, needs, or wants of this persona? What would make their job/life better or easier?

Roundtable (12 mins)

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”

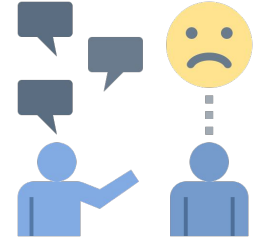
“Text analysis is unlikely to answer that question”

“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”

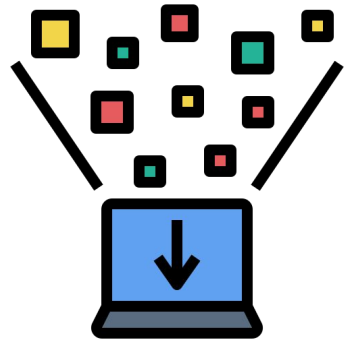
“That sounds like something teachers would love”

“That’s a really fundamental question”



Obtaining quantitative data

- **Collecting brand new quantitative data is beyond the scope of this course**
- You're welcome to bring your own data
- Use the the default course dataset: NCTE transcripts (see HW #1)
- 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](#)
- [NAEP Student Writing Data](#)

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

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?

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?

Guiding questions for conceptualization

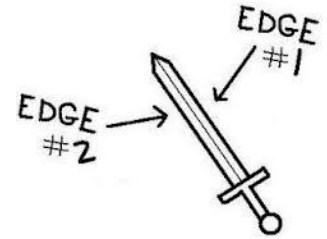
- What are the core concepts you are addressing? And are you being true to their core meaning?
- What are competing definitions? Which is best suited to the task and why?
- Does the systematized concept you've selected reflect an adequate understanding of the background concept?
- How do domain experts 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?

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?

Think about dual use

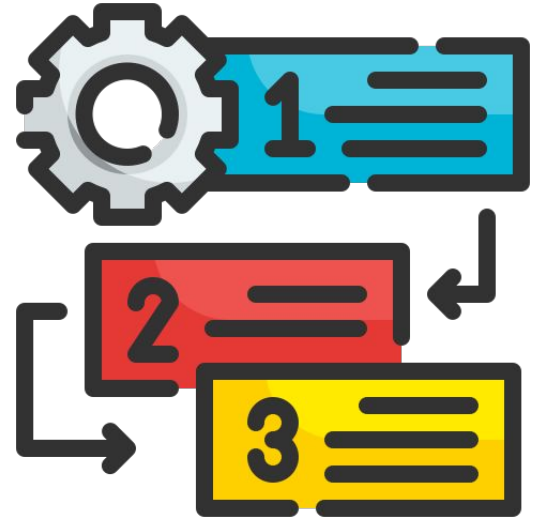
- What if the tool was used in any type of high stakes decision making?
- Could the tool be used to surveil teachers?
- Could it be used to punish students or teachers?
- Could the tool be used to harm vulnerable populations, aggravate biases and inequities?



Best practices for data exploration

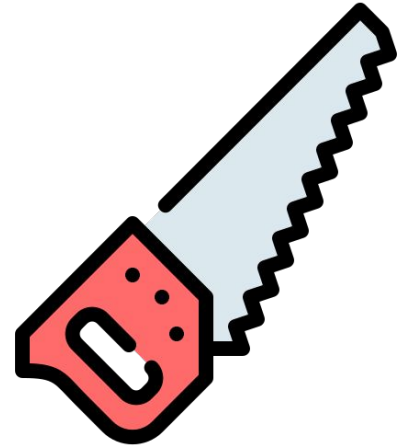
#1 Prioritize.

- 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.



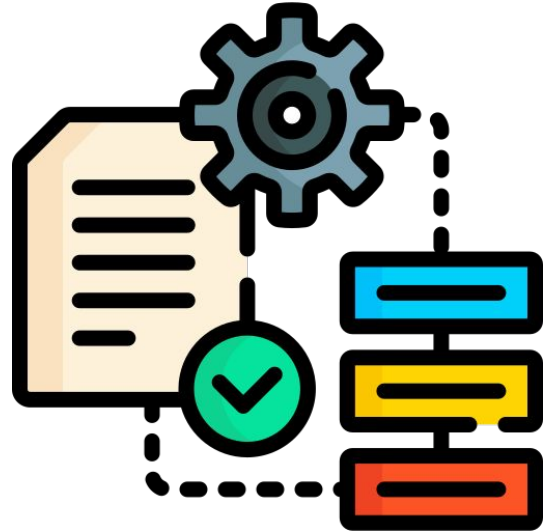
#3 Create visualizations.

- 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



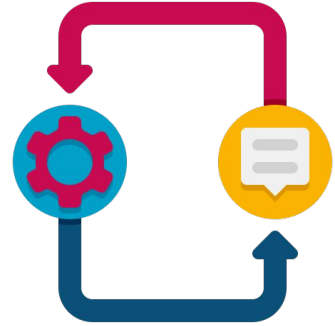
#4 Don't be hand-wavy about pre-processing.

- 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)

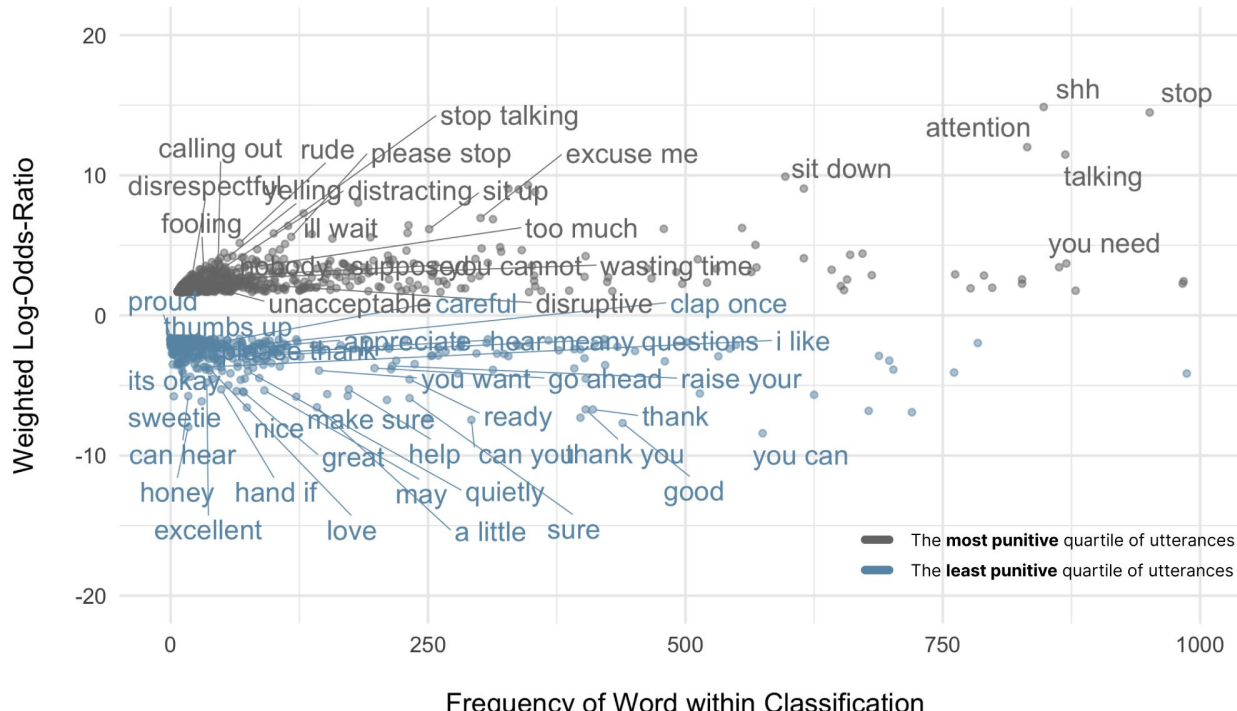
Z-scored log odds ratios (Section 4.3 in [Monroe et al., 2017](#))

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)
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).

Examples for using z-scored log odds ratios

- See this paper: [Demszky et al. \(2019\)](#) for an extensive use of this method (for words / phrases / topics / LIWC categories)
- [Tan & Demszky \(2023\)](#):



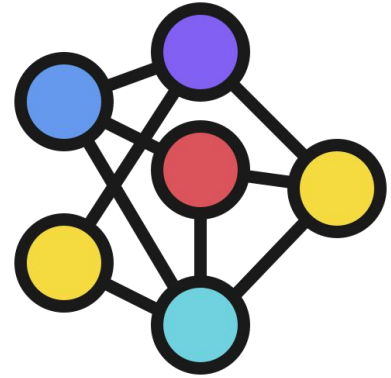
#7 Be scientific about debugging.

- 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](#)



Lucy Li



Tricia Bromley



Dan Jurafsky

Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas US History Textbooks

AEA Open
Volume 6, Issue 3, July-September 2020
© The Author(s) 2020, Article reuse Guidelines
<https://doi.org/10.1177/232858420940312>

SAGE
journals

Special Topic: Educational Data Science

Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks

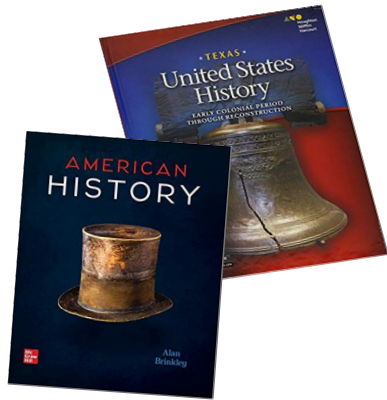
Li Lucy, Dorottya Denszky, Patricia Bromley, and Dan Jurafsky

Abstract
Cutting-edge data science techniques can shed new light on fundamental questions in educational research. We apply techniques from natural language processing (lexicons, word embeddings, topic models) to 15 U.S. history textbooks widely used in Texas between 2015 and 2017, studying their depiction of historically marginalized groups. We find that Latinx people are rarely discussed, and the most common famous figures are nearly all White men. Lexicon-based approaches show that Black people are described as performing actions associated with low agency and power. Word embeddings reveal that women tend to be discussed in the contexts of work and the home. Topic modeling highlights the higher prominence of political topics compared with social ones. We also find that more conservative counties tend to purchase textbooks with less representation of women and Black people. Building on a rich tradition of textbook analysis, we release our computational toolkit to support new research directions.

Keywords
artificial intelligence, case studies, content analysis, curriculum, data science, gender studies, history, natural language processing, race, textbooks, textual analysis

Motivation

Textbooks are the most widely used instructional tool around the world



Social & cultural
values



Traditional Methods

Coding protocols, e.g:

36. Fill in each cell of the matrix below using the following codes:

	Groups/ Issues	Rights
Citizens / citizenship		
Children, youth		
Women		
Elderly / Old Age		
Ethnic minorities / racism		
Indigenous groups		
Immigrants / Immigration or Refugees		
Workers / Labor		
Disabled, handicapped		
Gays, lesbians		
The poor / Poverty (nationally or in an international development context)		
Health		
Environment		
Education		
Language and/or culture		
Other. List:		

1 = mentioned

0 = not mentioned

0 = no mention

1 = one or two sentences

2 = at least a paragraph

3 = at least one subheading

4 = at least one chapter heading

5 = over half the chapters

Texas

- 5.4M K-12 students (2017), 2nd largest in US
- Major textbook market
- Large influence on textbooks in U.S.



Texas

The New York Times

How Texas Teaches History

The Washington Post

Education

What do students learn about slavery? It depends where they live.

★ THE TEXAS TRIBUNE

≡ MENU

Texas' Controversial Social Studies Textbooks Under Fire Again

The New York Times

Texas Mother Teaches Textbook Company a Lesson on Accuracy

Our Goal

Apply NLP to textbooks to answer questions that textbook researchers in education care about

Research Questions

RQ1

How much are different groups of people **mentioned**?

Research Questions

RQ1

How much are different groups of people **mentioned**?

RQ2

How are different groups and individuals **described**?

Research Questions

RQ1

How much are different groups of people **mentioned**?

RQ2

How are different groups and individuals **described**?

RQ3

Which **topics** are prominent and how do they relate to groups of people?

Research Questions

TODAY'S CLASS

RQ1

How much are different groups of people **mentioned**?

RQ2

How are different groups and individuals **described**?

RQ3

Which **topics** are prominent and how do they relate to groups of people?

American History Textbook Data (2015-17)

American History Textbook Data (2015-17)

Messy purchase data from
Texas districts

8th gr give me liberty	T.ISD
pearson US hsitory coloniz...	B.ISD
Pearson us history texas ed	B.ISD

American History Textbook Data (2015-17)

Messy purchase data from
Texas districts

8th gr give me liberty	T.ISD
pearson US hstory coloniz...	B.ISD
Pearson us history texas ed	B.ISD



manual clean-up &
disambiguation



book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100

American History Textbook Data (2015-17)

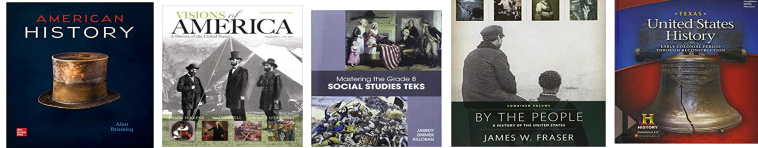
Messy purchase data from
Texas districts

8th gr give me liberty	T.ISD
pearson US hstory coloniz...	B.ISD
Pearson us history texas ed	B.ISD



book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100

keep **15** most widely purchased textbooks



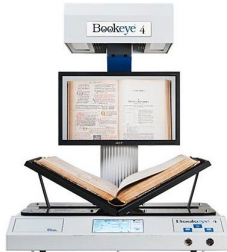
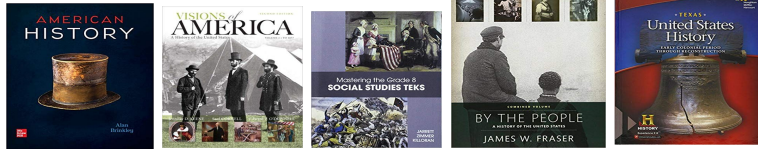
American History Textbook Data (2015-17)

Messy purchase data from
Texas districts

8th gr give me liberty	T.ISD
pearson US hstory coloniz...	B.ISD
Pearson us history texas ed	B.ISD



book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100

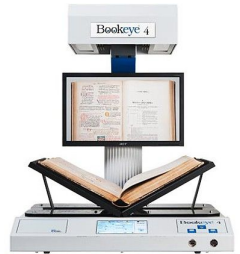


American History Textbook Data (2015-17)

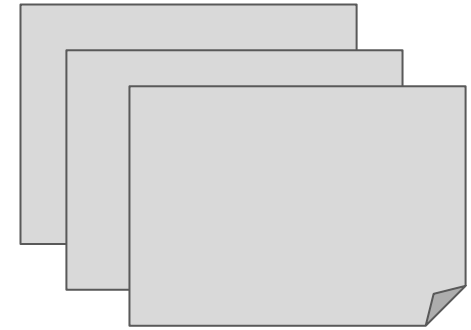
Messy purchase data from
Texas districts

8th gr give me liberty	T.ISD
pearson US hstory coloniz...	B.ISD
Pearson us history texas ed	B.ISD

book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100



OCR w/ **ABBYY FineReader®**



American History Textbook Data (2015-17)

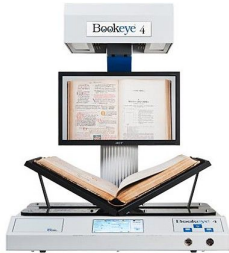
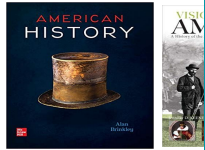
Messy purchase data from
Texas districts

8th gr give me liberty	T.ISD
pearson US hstory coloniz...	B.ISD
Pearson us his	



book	district	count
Am. Hist.	T.ISD	30
Give me lib	B.ISD	100

2 MONTHS OF WORK FOR TWO PEOPLE



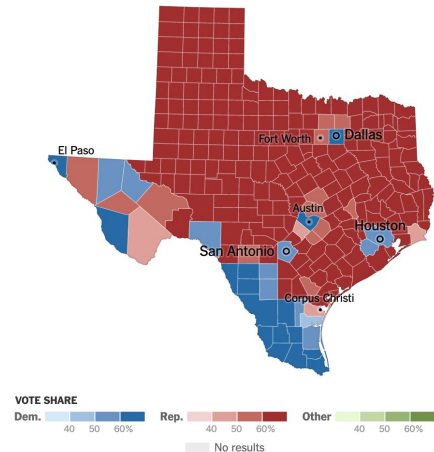
Demographic Data

- district-level student demographic data
 - the National Center for Education Statistics (NCES), for AY 2016-17



Demographic Data

- district-level student demographic data
 - the National Center for Education Statistics (NCES), for AY 2016-17
- county-level political leaning
 - two party vote shares in 2016 elections



source: [New York Times](#)

Example

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)



RQ1

How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

RQ1

How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. **Elizabeth Blackwell**, born in England, gained acceptance and fame as a physician. **Her** sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, **Lucy Stone**, took the revolutionary step of retaining **her** maiden name after marriage. **Stone** became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Coreference Resolution

RQ1

How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual **women** did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a **physician**. Her **sister-in-law** Antoinette Brown Blackwell became the first ordained woman **minister** in the United States; and another **sister-in-law**, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential **lecturer** on women's rights. (Brinkley, 2015: p. 330)

**Identifying people-related common nouns
(WordNet, 95% accuracy)**

RQ1

How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. **Elizabeth Blackwell** born in **England** gained acceptance and fame as a physician. Her sister-in-law **Antoinette Brown Blackwell** became the first ordained woman minister in the **United States**; and another sister-in-law, **Lucy Stone** took the revolutionary step of retaining her maiden name after marriage. **Stone** became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Named Entity Recognition

RQ1

Race/Ethnicity & Gender

Common nouns referring to individuals or groups

446 marked

1665 unmarked
engineer, family

Women
wife, mother
Men
son, boy

Black
black, slaves, africans
Latinx
mexican, latina
White
colonist, white, european
Other
immigrants, asian-americans

RQ1

Race/Ethnicity & Gender

Common nouns referring to individuals or groups

446 marked

1665 unmarked
engineer, family

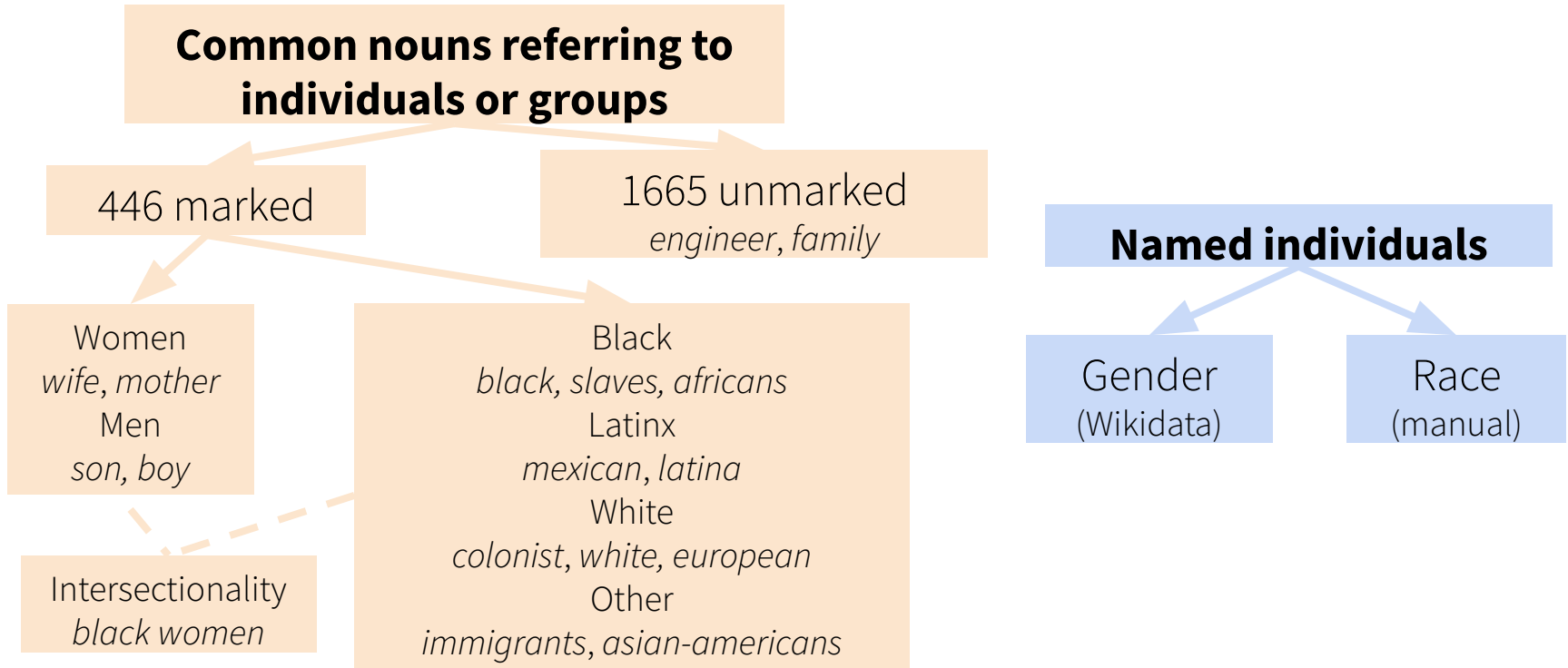
Women
wife, mother
Men
son, boy

Intersectionality
black women

Black
black, slaves, africans
Latinx
mexican, latina
White
colonist, white, european
Other
immigrants, asian-americans

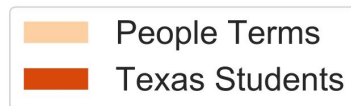
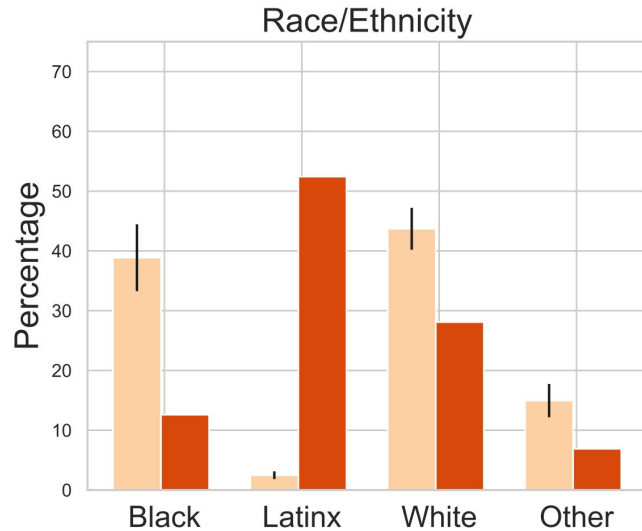
RQ1

Race/Ethnicity & Gender



RQ1

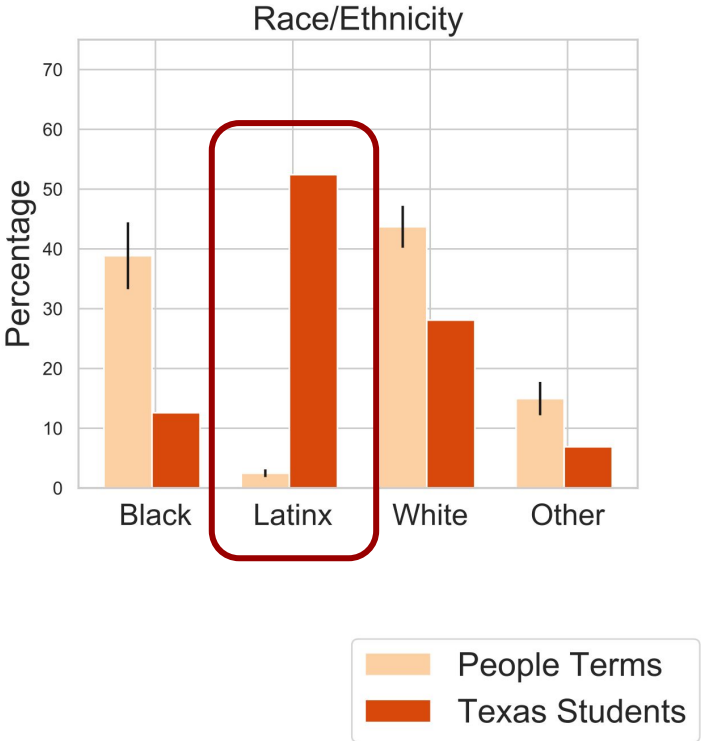
Comparing Student Demographics w/ Representation in Text



Common nouns referring to individuals or groups

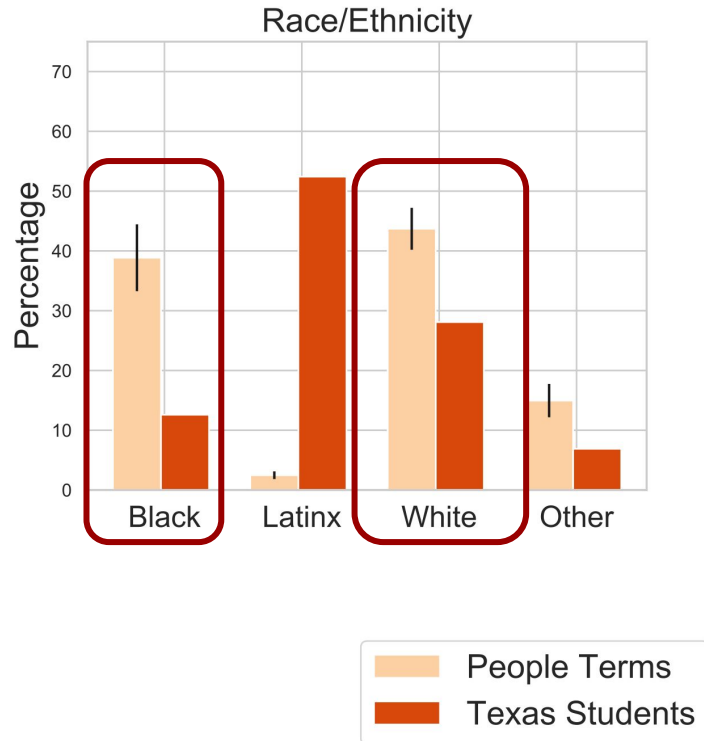
RQ1

Hispanic / Latinx Students are Disproportionately Underrepresented



RQ1

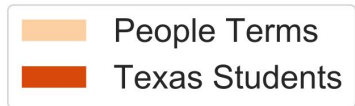
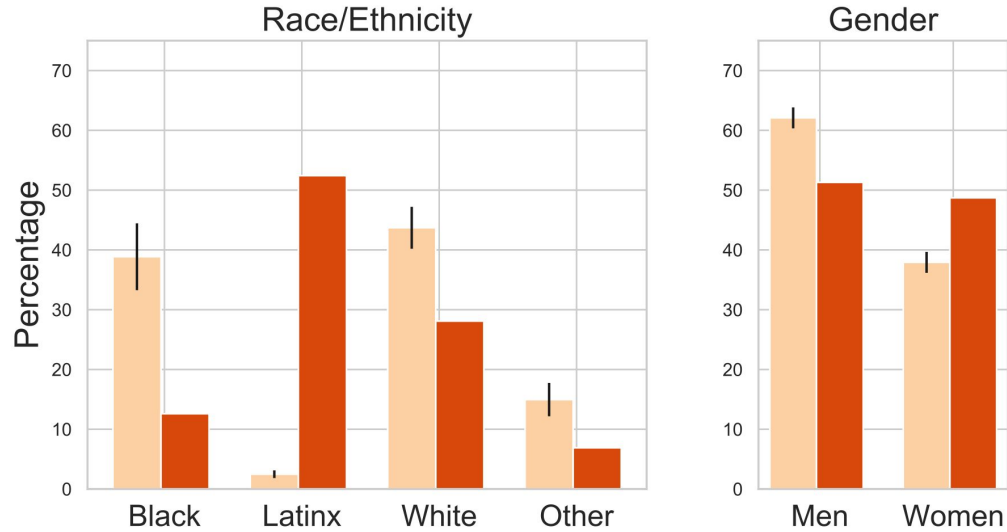
African Americans and White People are Mentioned Disproportionately More



white people are mentioned even more often than the plot shows (since this ethnicity is often unmarked)

RQ1

Men Are Mentioned Disproportionately More Often Than Women



RQ1

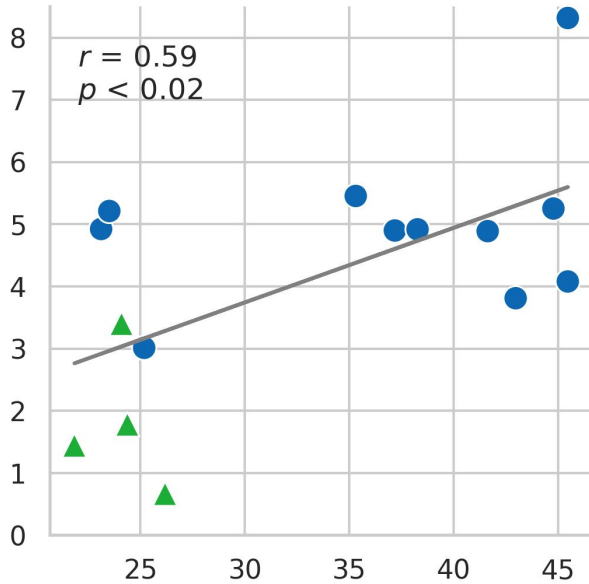
Top 50 Named People



RQ1

Books in More Democratic Counties Mention Black People and Women More

% of All People Terms



Median % of Democrats Across Counties
Where Textbook is Bought

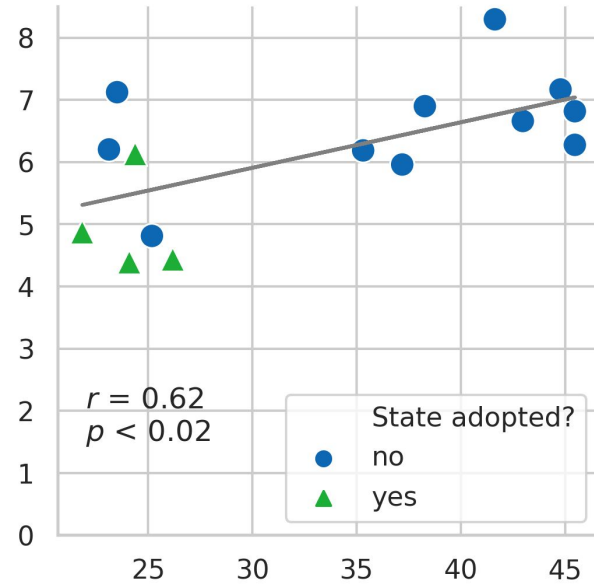
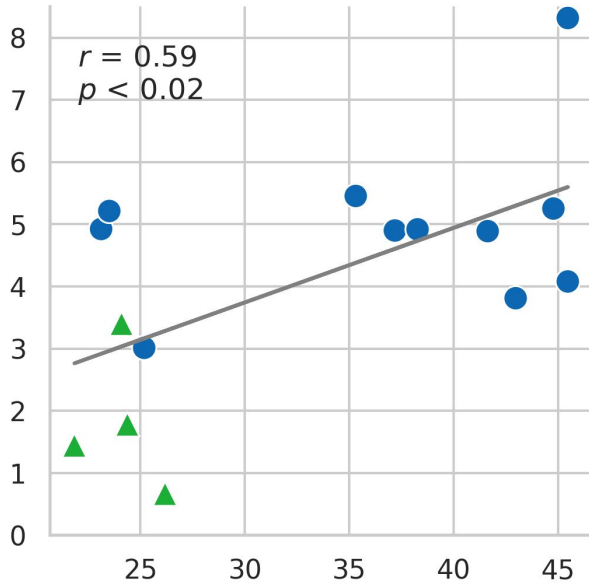
RQ1

Books in More Democratic Counties Mention Black People and Women More

Black People

Women

% of All People Terms



Median % of Democrats Across Counties
Where Textbook is Bought



RQ2

How Are Different Groups and Individuals **Described**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

RQ2

How Are Different Groups and Individuals **Described**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Dependency Parsing

RQ2

How Are Different Groups and Individuals Described?

Progress toward feminist goals was limited in the antebellum years, but in **subject** women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Dependency Parsing

RQ2

How Are Different Groups and Individuals **Described**?

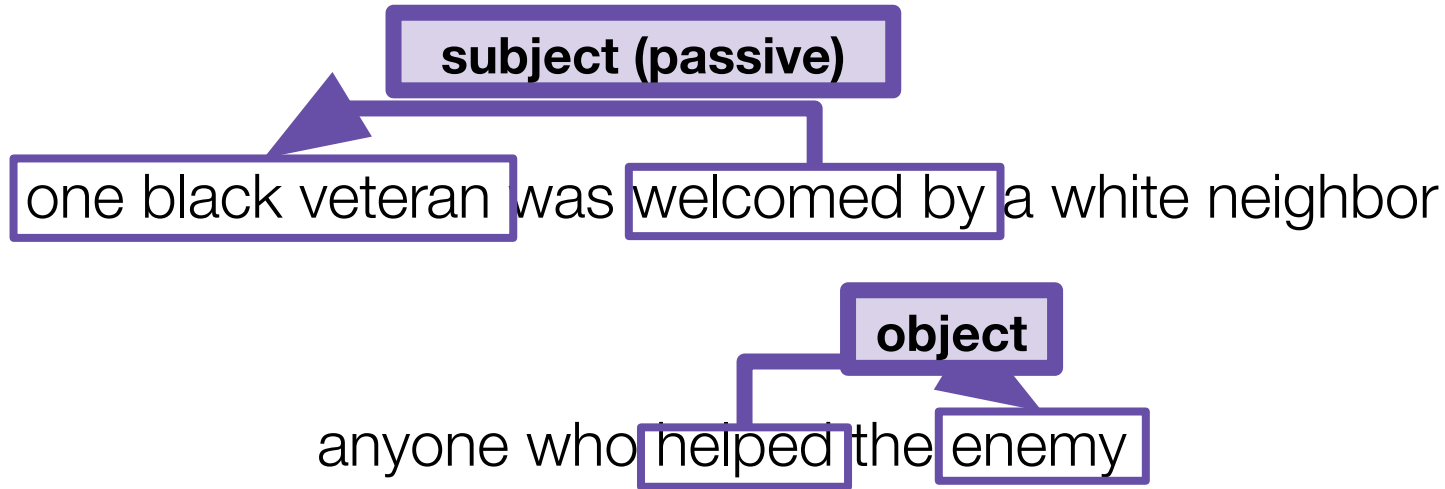
adj modifier

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Dependency Parsing

RQ2

How Are Different Groups and Individuals **Described?**



Dependency Parsing

RQ2

Lexicons

adjectives

NRC Valence, Arousal, Dominance lexicons (Mohammad, 2018)

amazing (↑ valence)

asleep (↓ arousal)

competitive (↑ dominance)

RQ2

Lexicons

adjectives

NRC Valence, Arousal, Dominance lexicons (Mohammad, 2018)

amazing (↑ valence)

asleep (↓ arousal)

competitive (↑ dominance)

verbs

Connotation frames (Rashkin et al, 2016; Sap et al., 2017)

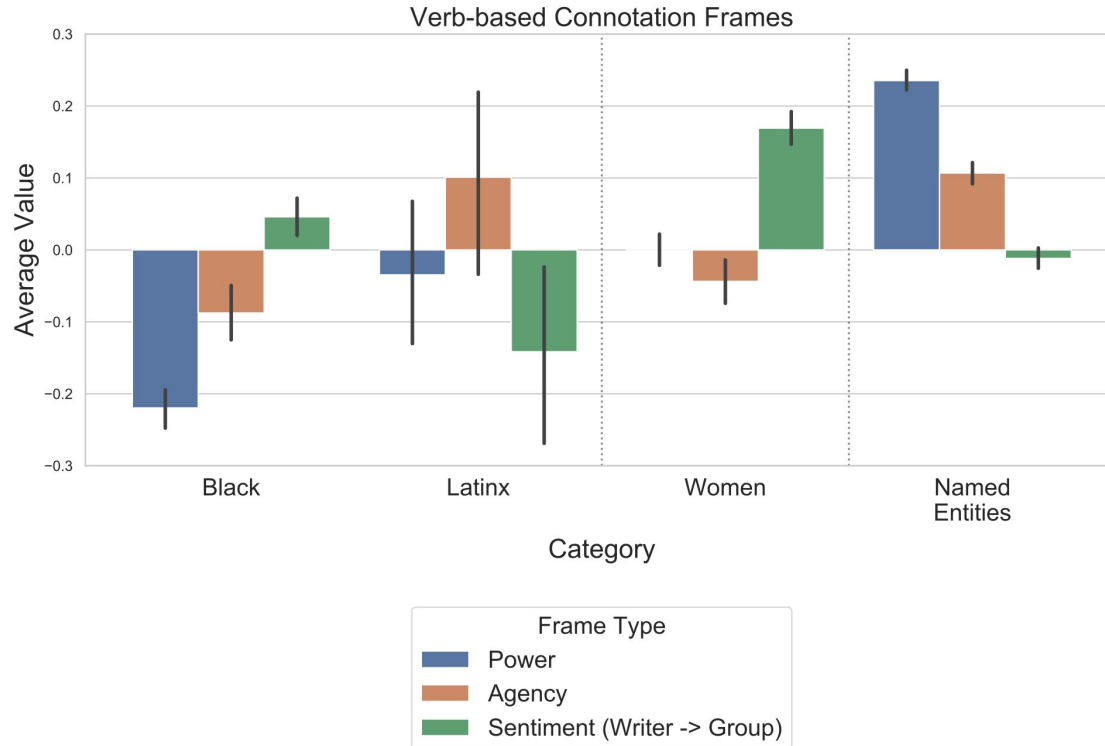
X (-1 agency) obeys

X (-1 power) applauds Y (+1 power)

X (↑ sentiment) suffered

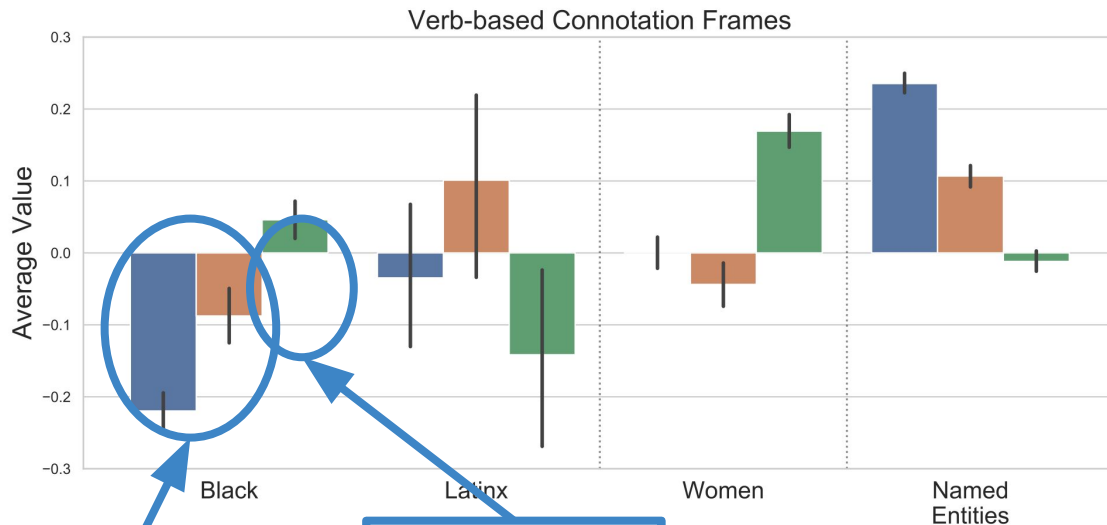
RQ2

Power & Agency



RQ2

Power & Agency



owned, barred

want, have

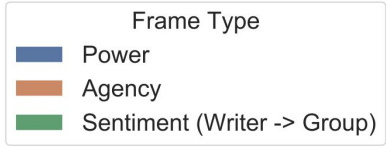
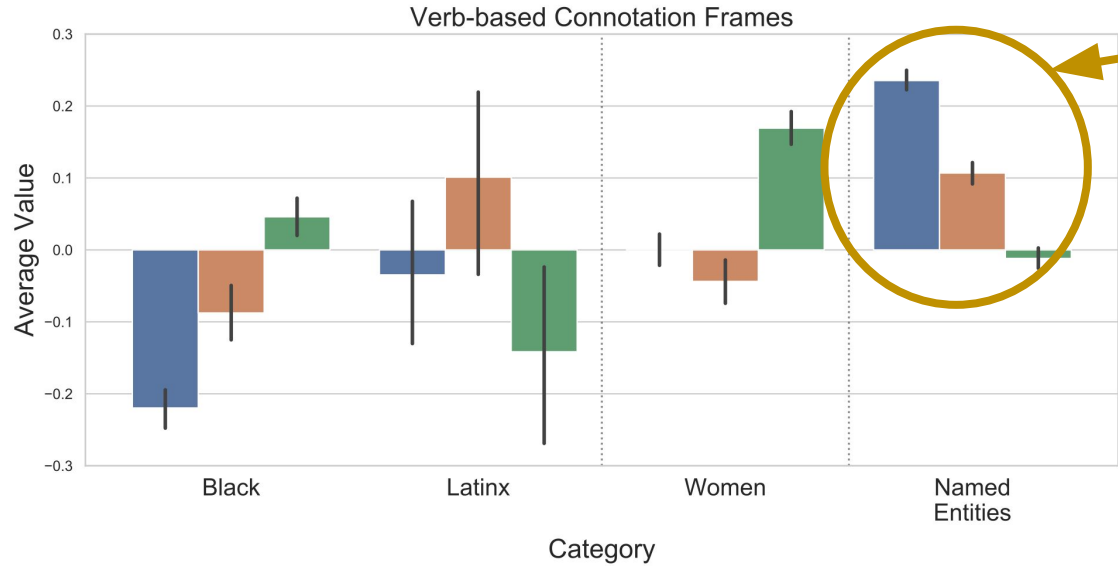
Frame Type

- Power
- Agency
- Sentiment (Writer -> Group)

RQ2

Power & Agency

veto, initiate



RQ2

Other Lexicon Findings

African Americans (↓ adjective dominance)

Ex: *slave, inferior*

Famous people (↑ adjective arousal)

Ex: *worried, victorious, furious*

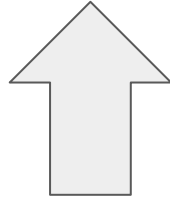
Women (↑ verb sentiment)

Ex: women *marry* or *help*

RQ2

GloVe Embeddings w/ Bootstrapping

- unigrams & bigrams (skip stopwords)
- GloVe training w/ bootstrapping (Antoniak & Mimno, 2018)
- mean cosine similarity across 50 runs, between:



Bootstrapping helps mitigate data sparsity!

Create samples of the data (e.g. sample 50 times with replacement), train model on each and aggregate results.

RQ2

GloVe Embeddings w/ Bootstrapping

- unigrams & bigrams (skip stopwords)
- GloVe training w/ bootstrapping (Antoniak & Mimno, 2018)
- mean cosine similarity across 50 runs, between:

man-related terms

(*man, men, male, he, his, him*)

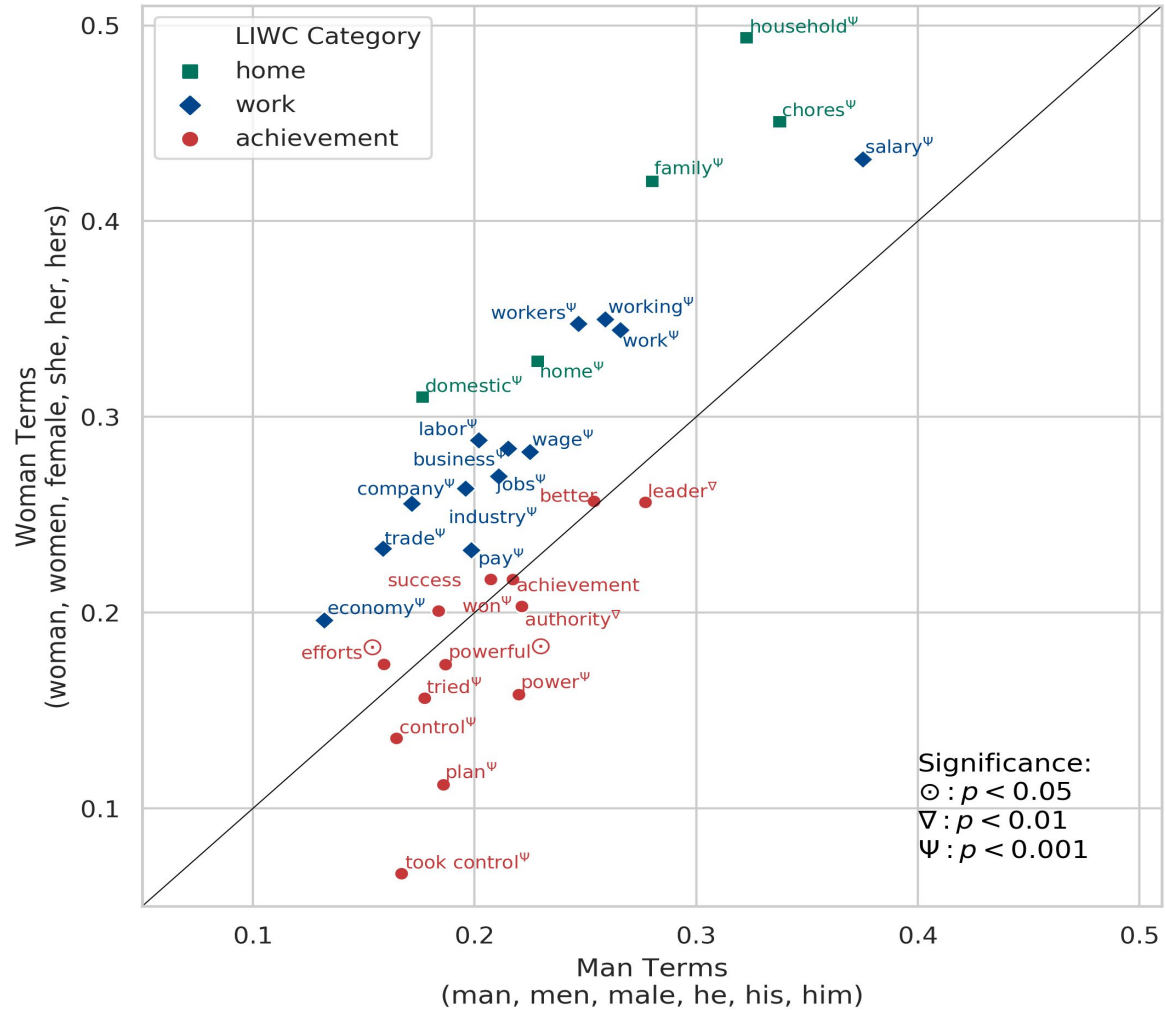
woman-related terms

(*woman, women, female, she, her, hers*)



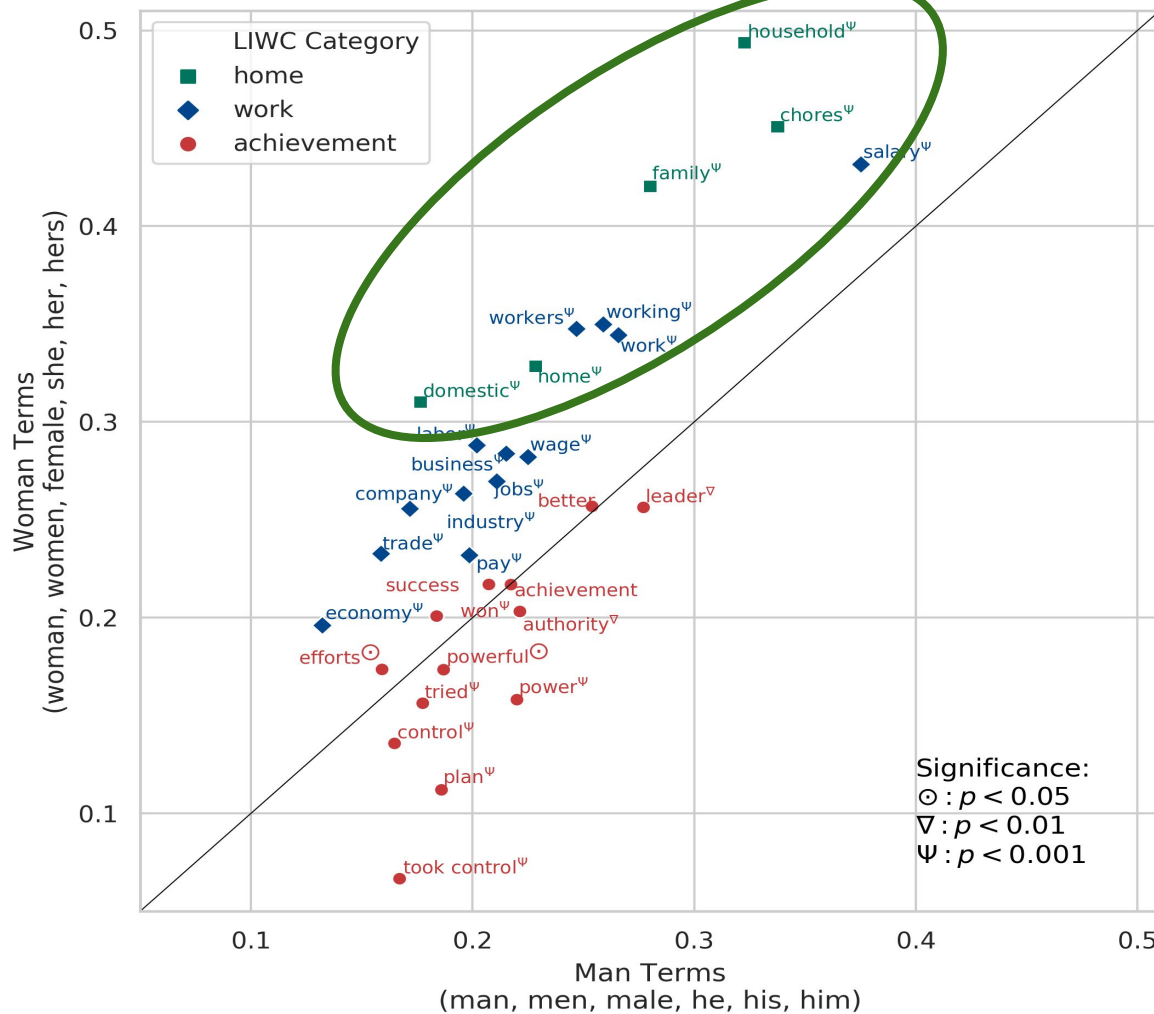
most frequent words in
home, **work** and
achievement LIWC
categories

RQ2



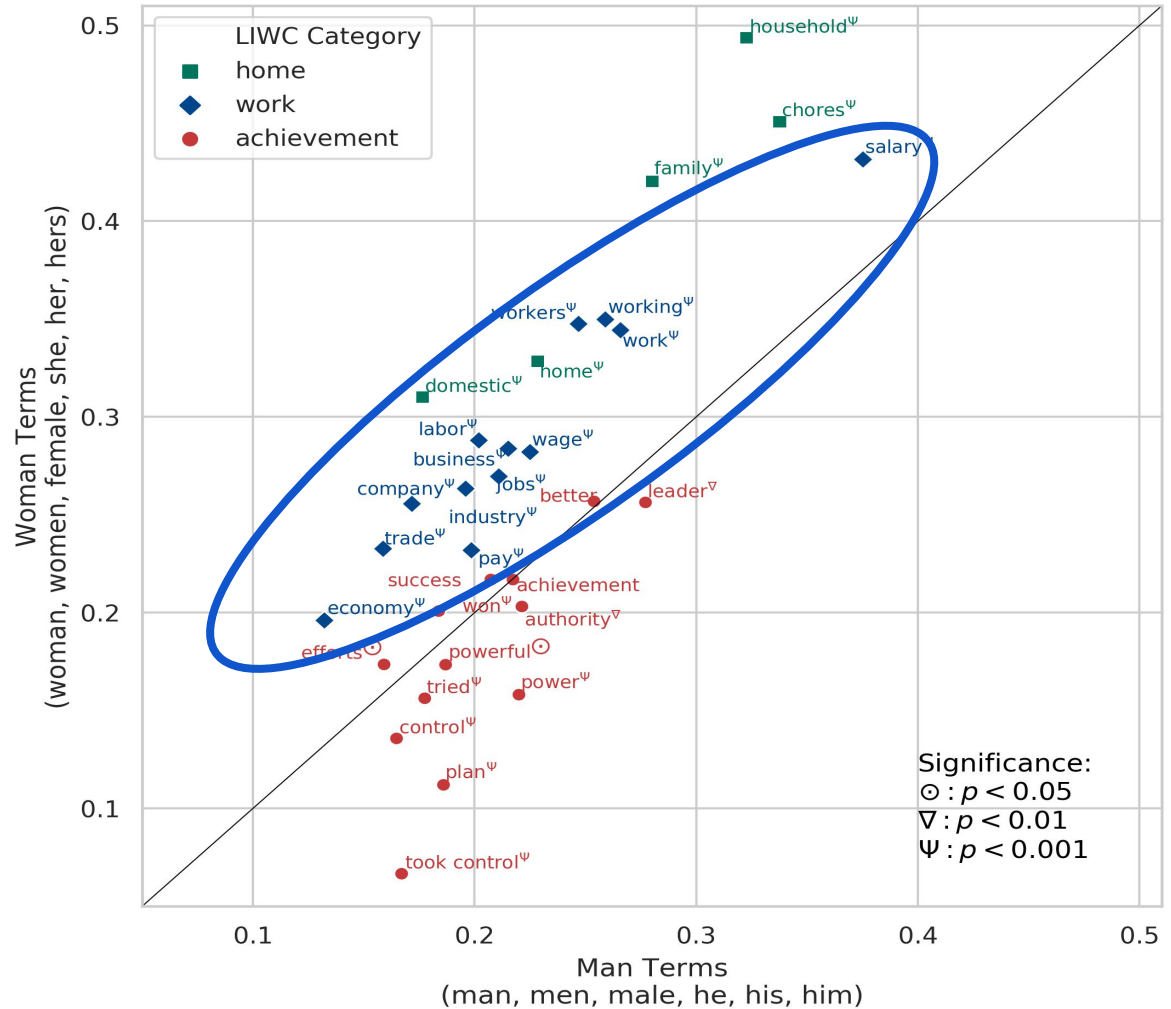
RQ2

Home related terms are more closely related to women, with very high significance



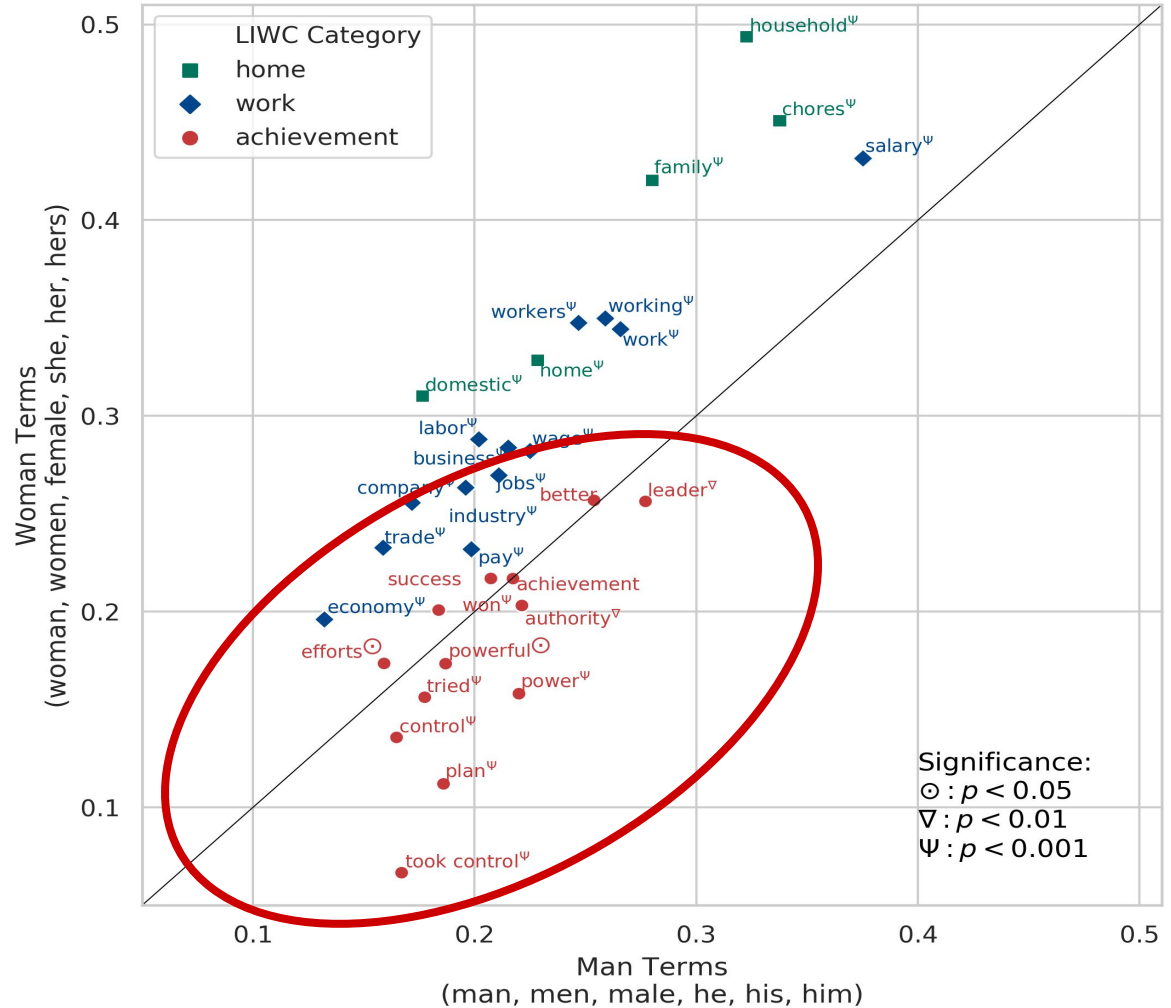
RQ2

Work related terms are more closely related to women, with very high significance



RQ2

Most **achievement** related terms are more closely related to men but not all





What questions do you have?

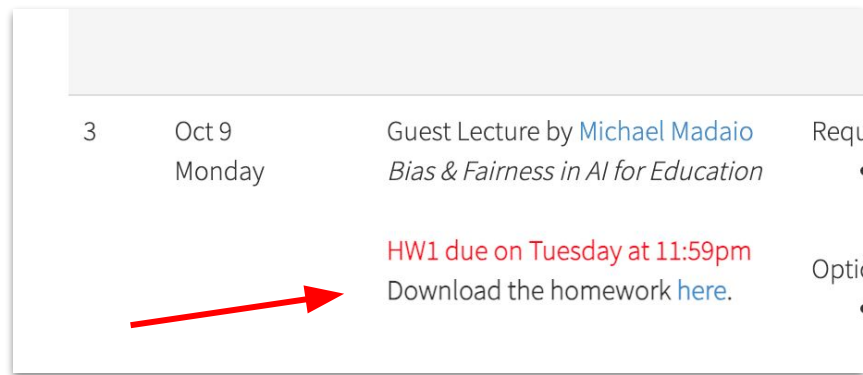
Homework #1 Setup

Instructions for HW1

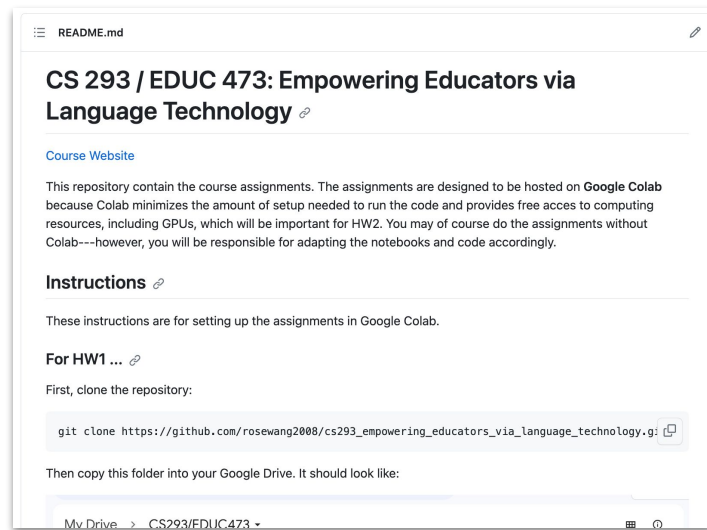
Homework assignments are intended to be hosted on Colab.

- Clone the repository.
- 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.



3 Oct 9 Monday Guest Lecture by [Michael Madaio](#) *Bias & Fairness in AI for Education* Requ
HW1 due on Tuesday at 11:59pm
Download the homework [here](#). Optic



README.md

CS 293 / EDUC 473: Empowering Educators via Language Technology

[Course Website](#)

This repository contain the course assignments. The assignments are designed to be hosted on **Google Colab** because Colab minimizes the amount of setup needed to run the code and provides free acces to computing resources, including GPUs, which will be important for HW2. You may of course do the assignments without Colab---however, you will be responsible for adapting the notebooks and code accordingly.

Instructions

These instructions are for setting up the assignments in Google Colab.

For HW1 ...

First, clone the repository:

```
git clone https://github.com/rosewang2008/cs293_empowering_educators_via_language_technology.g
```

Then copy this folder into your Google Drive. It should look like:

My Drive > CS293/EDUC473