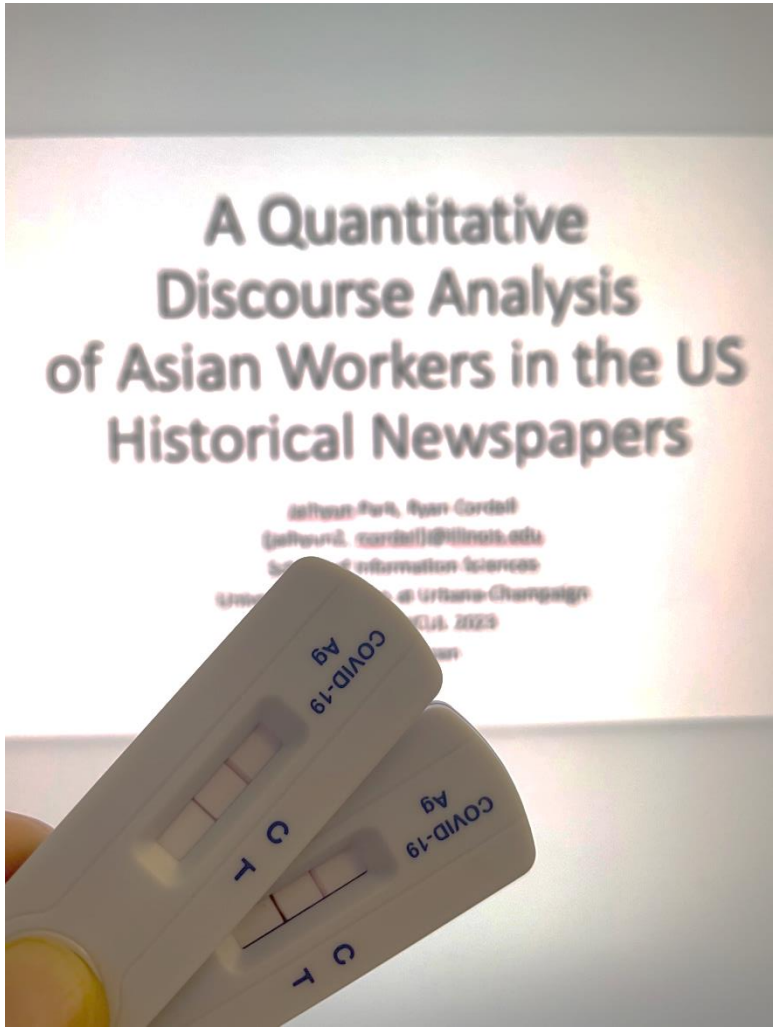
A Quantitative Discourse Analysis of Asian Workers in the US Historical Newspapers

Jaihyun Park, Ryan Cordell
{jaihyun2, rcordell}@illinois.edu
School of Information Sciences
University of Illinois at Urbana-Champaign
NLP4DH-IWCUL 2023
Tokyo, Japan

I was tested positive for COVID :(



- Warning!: This presentation contains examples of offensive language targeting marginalized populations. Feel free to leave the room while this recording is in play when you feel uncomfortable.

Motivation

- Digitization of historical texts has opened up new opportunities for researchers to explore a large-scale corpus with computational methods.

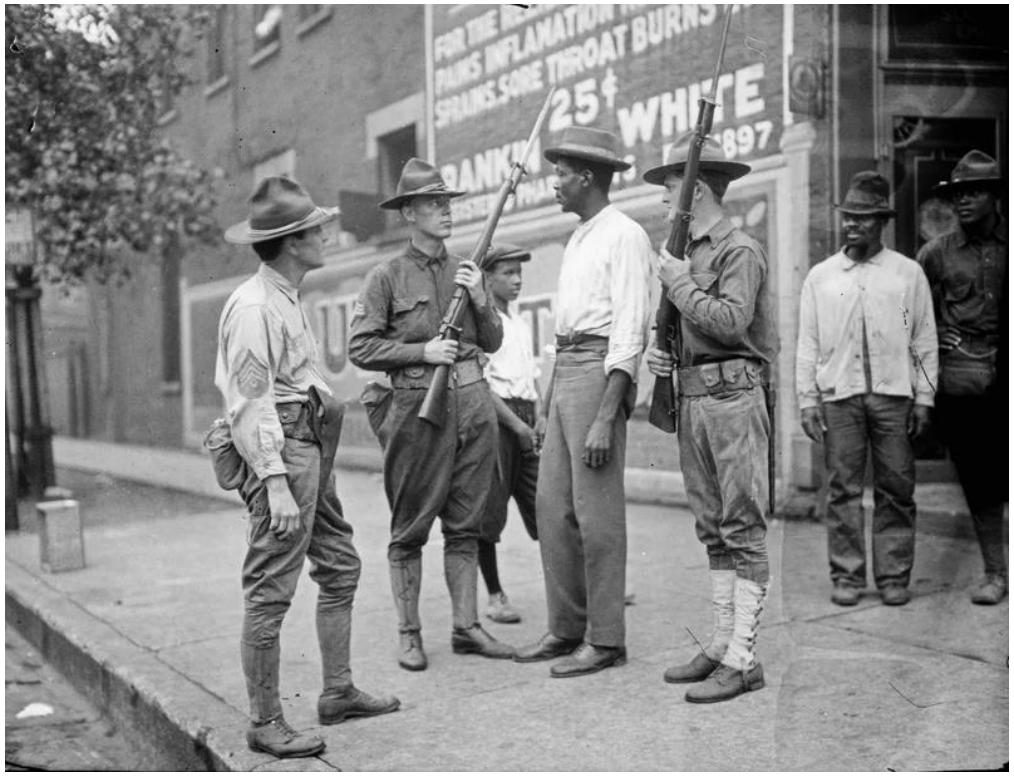
- Taking this advantage, many researchers in diverse fields, such as Sociology, History, English, and Information Science have applied NLP techniques to historical texts such as books (Parulian et al., 2022), newspapers (Smith et al., 2013), and/or congressional records (Lin and Peng, 2022; Guldi, 2019).

- In *Digital Humanities*, there have been studies primarily focused on the race problem in the US.

- Soni et al. (2021) used diachronic word embedding to trace the semantic change of the word from African American newspapers corpus.
- Franzosi et al. (2012) performed NER on 19C newspapers to detect locations of lynchings in Georgia.

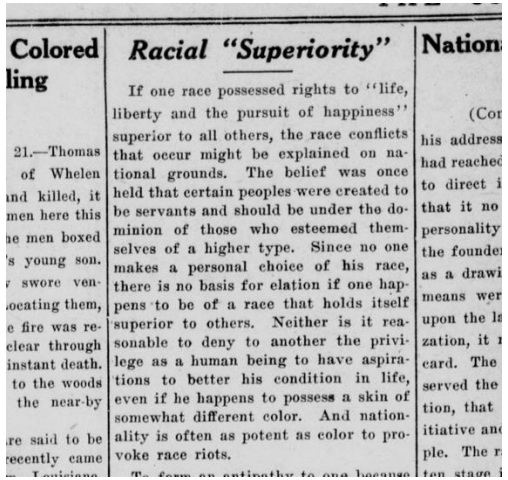
- These studies intersect the problem of historical racism and NLP research.

Historical Event



Armed National Guards and African American men during the race riot in Chicago, 1919 (Red Summer)

Digitized History – Image & Text



Racial "Superiority"

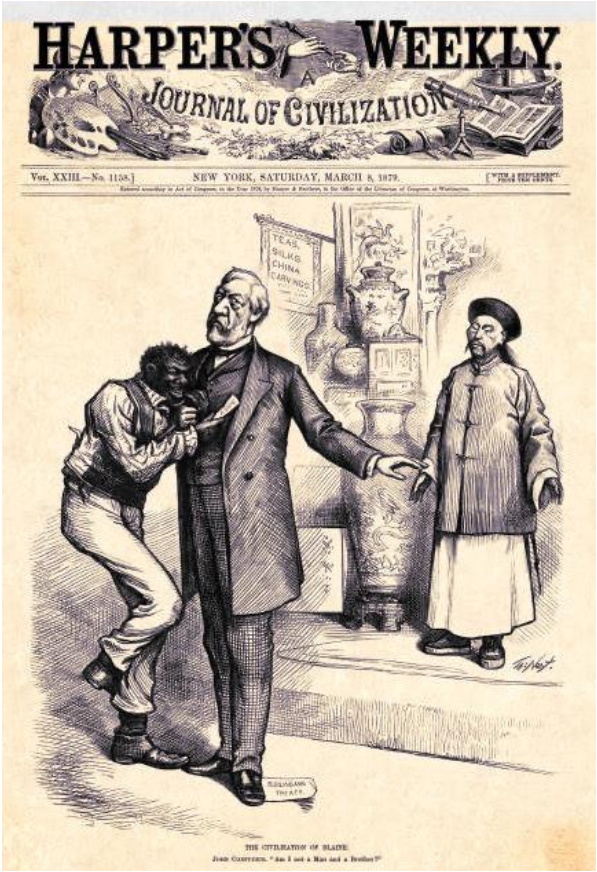
If one race possessed rights to "life, liberty and the pursuit of happiness" superior to all others, the race conflicts that occur might be explained on national grounds. The belief was once held that certain peoples were created to be servants and should be under the dominion of those who esteemed themselves of a higher type. Since no one makes a personal choice of his race, there is no basis for elation if one happens to be of a race that holds itself superior to others. Neither is it reasonable to deny to another the privilege as a human being to have aspirations to better his condition in life, even if he happens to possess a skin of somewhat different color. And nationality is often as potent as color to provoke race riots.

https://chroniclingamerica.loc.gov/lccn/sn86056950/1919-08-21/ed-1/seq-7/

Racial "Superiority"
If one race possessed rights to "life, liberty and the pursuit of happiness" superior to all others, the race conflicts that occur might be explained on na tional grounds. The belief was once held that certain peoples were created to be servants and should be under the do minion of those who esteemed them selves of a higher type. Since no one makes a personal choice of his race, there is no basis for elation if one hap pens to be of a race that holds itself superior to others. Neither is it rea sonable to deny to another the privi lege as a human being to have aspira tions to better his condition in life, even if lie happens to possess a skin of somewhat different color. And nation ality is often as potent as color to pro voke race riots.

https://chroniclingamerica.loc.gov/lccn/sn86056950/1919-08-21/ed-1/seq-7/ocr/

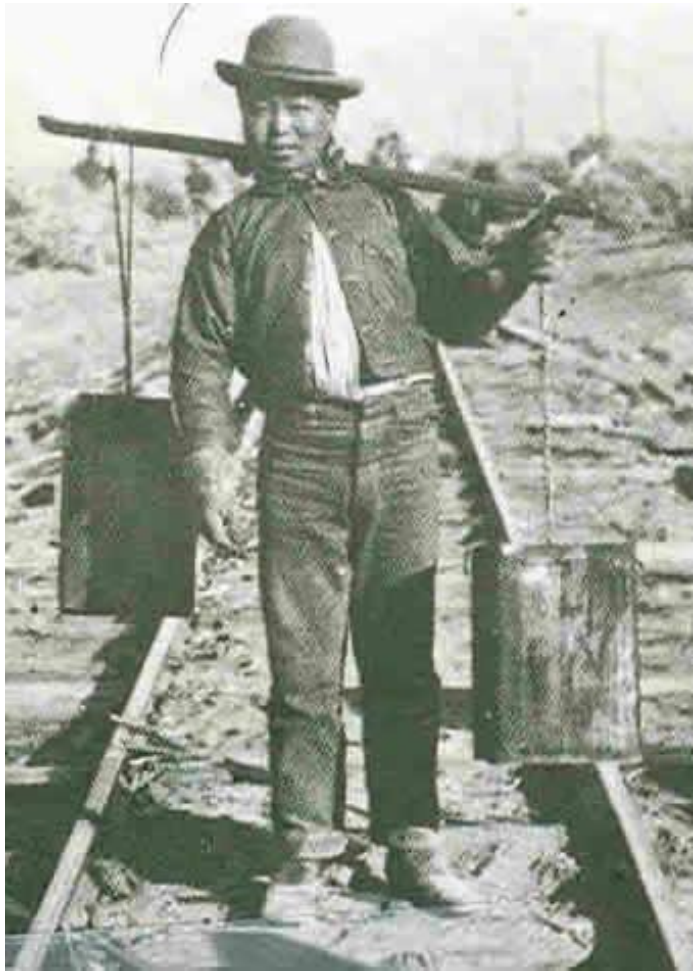
https://longislandwins.com/columns/immigrants-civil-war/ban-chinese-proposed-frederick-douglass-spoke-3/



https://thomasnastcartoons.com/2014/04/08/civilization-of-blaine-8-march-1879/

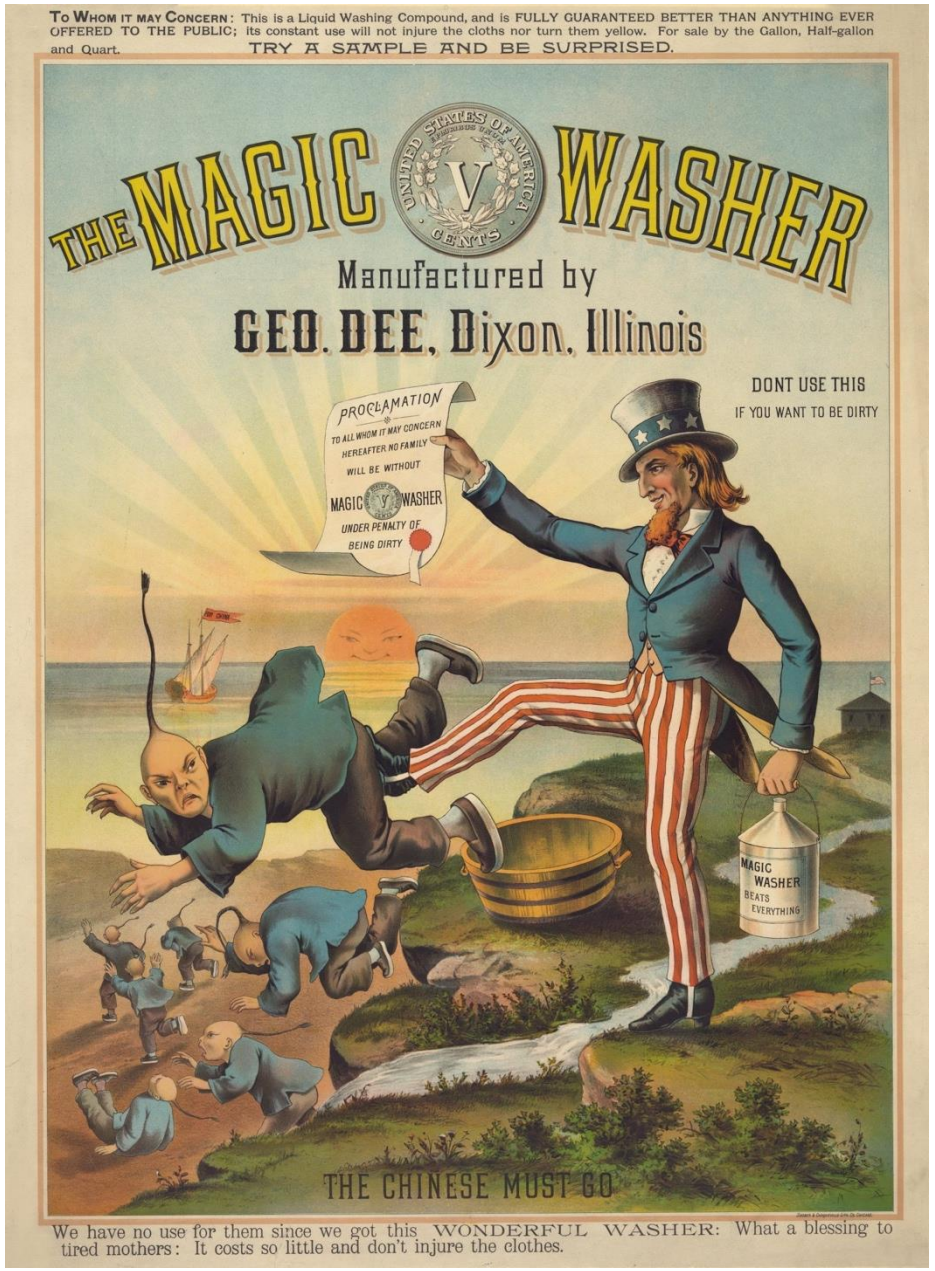
- However, historical racism toward Asians in the US is understudied compared to the major race tension between White and Black.

Background



https://www.dol.gov/general/aboutdol/hallofhonor/2014_railroad

- Coolies: “Pejorative European usage” to describe “an unskilled labourer or porter usually in or from the Far East hired for low or subsistence wages (Britannica).
- Chinese workers arrived in the US and perceived to be patient, tractable, obedient, industrious and frugal compared to African slaves (Jung, 2006)



https://www.calmigration.org/exclusion

- Chinese Exclusion Act of 1882: The first American gatekeeping of immigration and defined the desirability (and "Whiteness") of immigrant groups (Lee, 2002).

- The problem of coolie exemplifies the extension of colonial and capitalist exploitation beyond Africa and sugarcoated the extended system as indentured migrant contract workers (Van Rossum, 2016).

- RQ 1. How different are the semantic meaning of “coolie” in each State?
- RQ 2. What are the words over-represented in the newspapers between then-Confederate States and then-Union States?
- RQ 3. What “coolie” stories are reprinted and what are their characteristics?

Methodology

- Corpus: Chronicling America
- Data collection: 124,511 newspaper pages with the search term “coolie”
- Searched the exact string match for “coolie.”
- Pseudo-sentence creation: ten tokens before and after where “coolie” appeared.
- Data pre-processing: removed punctuations, non-alphabet tokens, stopwords (NLTK package). Lemmatiziation (Spacy).
- Catch possible OCR errors: included top 200 similar tokens based on FastText embedding (Bojanowski et al., 2017) (e.g., “coolieize” (0.8654), “oroolie (0.8630), “roolie” (0.8541)).

- Final data for analysis: 125,253 text

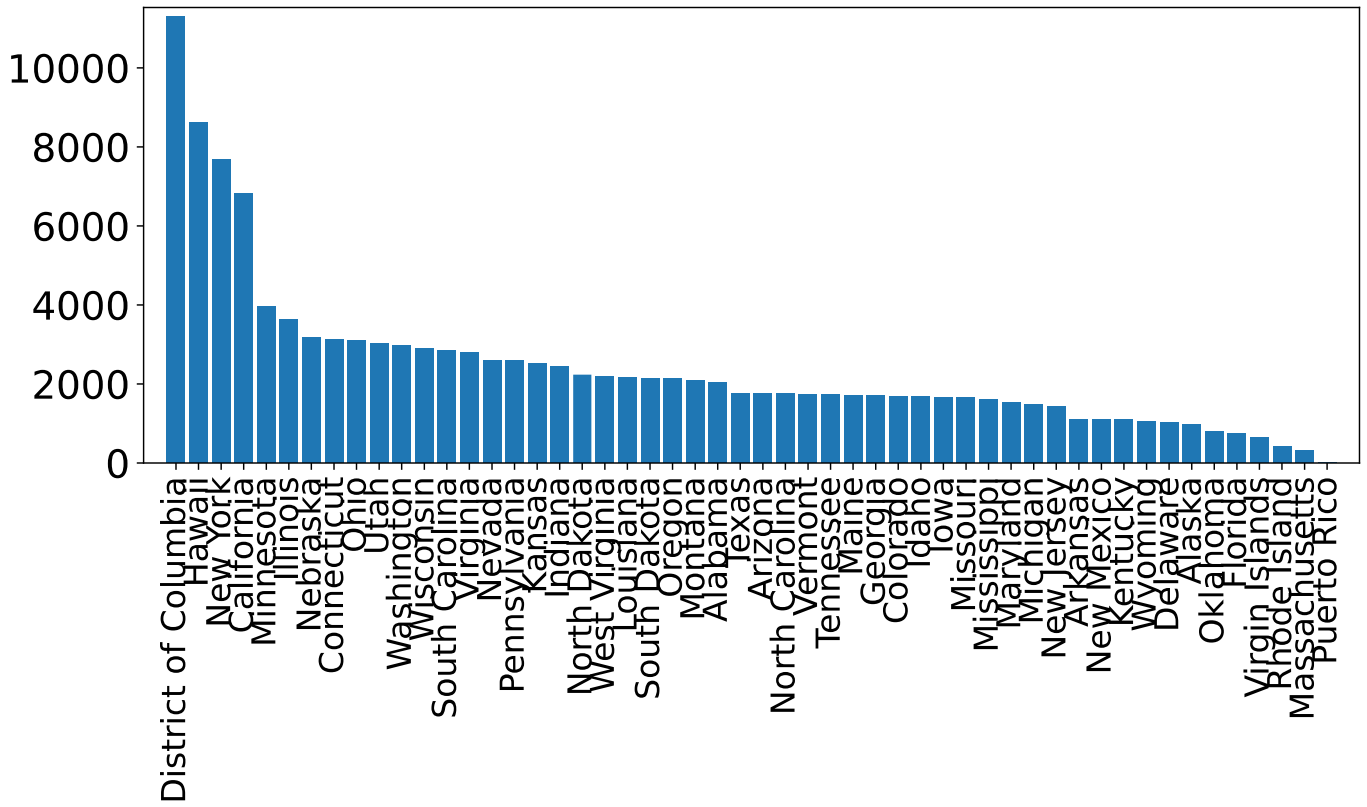


Figure 1: The count of text data containing the word “coolie” by State

- The count of text data is not evenly distributed due to the different digitization processes of the newspapers.

- DC (n=11,302)
- HI (n=8,613)
- NY (n=7,671)

- PR (n=15)
- MA (n=305)
- VI (n=656)

RQ1. Word embedding

- Word2vec model with a minimum word count of 5 and window size of 5.
- To represent the State-level embedding, the embedding vector of the word “coolie” was averaged and cosine similarity was calculated across the States.

RQ2. Statistically over-represented words

- Then-Confederate States: Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia.
- Then-Union States: Maine, New York, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island, Pennsylvania, New Jersey, Ohio, Indiana, Illinois, Kansas, Michigan, Minnesota, Wisconsin, Iowa, California, Nevada, Oregon, Delaware, Maryland, and West Virginia.
- VI and PR are excluded

RQ2. Statistically over-represented words

$$\delta_w^{(i-j)} = \log \frac{y_w^i + a_w}{n^i + a_0 + y_w^i - a_w} - \log \frac{y_w^i + a_w}{n^i + a_0 - y_w^i - a_w}$$

- 15,000 most frequent words were selected and Z-score was calculated.
- When n^i is the total number of words in corpus i, y_w^i is the number of times word i appeared in corpus i, a_0 is the size of the corpus a, and a_w is the frequency of word w in corpus a (Kwak et al., 2020).

RQ 3. Text reprint detection

- We used n-gram document representations to detect text reprints within errorful OCR-derived text.
- We processed the corpus with a 5-gram chunking using NLTK whitespace tokenizer and further made a judgement that the text has been reprinted when there were more than three matches of 5-grams across the corpus.
- [“demolish”, “part”, “build”, “injure”, “two”, “coolie”, “police”, “investigation”, “latter”, “case”, “lead”]
- [“demolish”, “part”, “build”, “jure”, ”two”, “latter”, “case”, “lead”]

RQ1. Results: Comparing the meaning of coolie

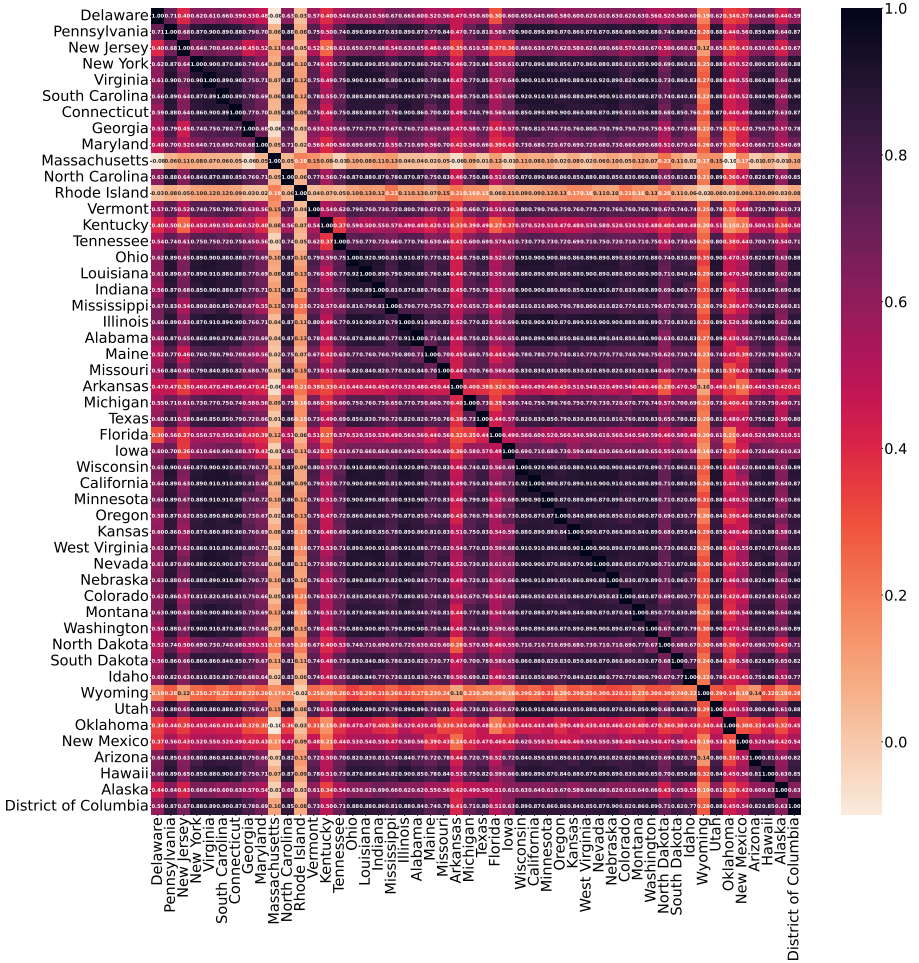


Figure 2:The heatmap of cosine similarity comparison across the average embedding vector of the word “coolie” in each State.

- MA and RI showed average cosine similarity of 0.08 and 0.12 while average cosine similarity across the entire States was 0.65.
- MA vs. OK (-0.10) / MA vs. ND (0.23)
- RI vs. DE (-0.03) / RI vs. MS (0.23)
- Some then-Confederate States showed lower cosine similarity (e.g., AR (0.43), FL (0.48), TN (0.64))
- AR vs. MA (-0.06) / AR vs. CO (0.54)
- FL vs. RI (0.06) / FL vs NV, UT (0.61)
- TN vs. MA (-0.03) / TN vs. UT (0.80).

The highest five States				
Illinois	California	Wisconsin	Virginia	Nevada
labor (0.9998)	country (0.9995)	chinese (0.9998)	chinese (0.9998)	chinese (0.9997)
chinese (0.9998)	bill (0.9995)	labor (0.9998)	man (0.9997)	club (0.9997)
wage (0.9997)	upon (0.9995)	two (0.9998)	trade (0.9997)	labor (0.9997)
two (0.9997)	well (0.9995)	one (0.9998)	work (0.9997)	make (0.9997)
one (0.9997)	go (0.9995)	time (0.9998)	one (0.9997)	use (0.9996)
day (0.9997)	stop (0.9995)	carry (0.9997)	three (0.9997)	say (0.9996)
china (0.9997)	many (0.9995)	take (0.9997)	number (0.9997)	importation (0.9996)
man (0.9997)	american (0.9995)	make (0.9997)	make (0.9997)	man (0.9996)
pay (0.9997)	con (0.9995)	japanese (0.9997)	importation (0.9997)	trade (0.9996)
say (0.9997)	would (0.9995)	would (0.9997)	two (0.9997)	day (0.9996)

Table 1: Top 10 most similar words to the word “coolie” in the top 5 States that showed the most similar meaning of the averaged word ”coolie”

- The word “coolie” was used in the context of labor, China, and wage.
- Labor (e.g., “labor”, “work”): IL, WI, VA, NV.
- “Chinese”: IL, WI, VA, NV / “Japanese”: WI.
- Wage (e.g., “wage”, ”pay”): IL
- Common words: “make”, “say”, “man”, numbers (e.g., “one”, “two”)

The lowest five States				
Massachusetts	Rhode Island	Wyoming	Oklahoma	Arkansas
among (0.3579)	order (0.4116)	chinese (0.9969)	chinese (0.9821)	labor (0.9985)
call (0.3201)	india (0.3972)	labor (0.9969)	shoulder (0.9815)	chinese (0.9984)
know (0.2832)	woman (0.3922)	would (0.9955)	japanese (0.9776)	mongolian (0.9978)
prohibit (0.2295)	great (0.3890)	japanese (0.9952)	pay (0.9748)	one (0.9978)
time (0.2232)	take (0.3761)	six (0.9950)	also (0.9725)	thousand (0.9977)
report (0.2221)	ship (0.3745)	one (0.9949)	labor (0.9706)	japanese (0.9975)
get (0.2209)	law (0.3432)	bring (0.9948)	carry (0.9632)	tolerate (0.9975)
arrive (0.2176)	united (0.3260)	say (0.9946)	home (0.9632)	revivial (0.9974)
come (0.2131)	carry (0.3234)	work (0.9944)	get (0.9630)	may (0.9974)
two (0.2104)	many (0.3122)	two (0.9941)	work (0.9612)	carry (0.9974)

Table 2: Top 10 most similar words to the word “coolie” in the top 5 States that showed the most dissimilar meaning of the averaged word ”coolie”

- MA and RI don’t show labor, ethnicity, and wage-related words.
- WY has 10 similar words in the highest top five States.
- OK has unique words: “shoulder”, ”home”
- AR has unique words: “Mongolian”, “tolerate”, “revival”

RQ 2. Results: Over-represented words

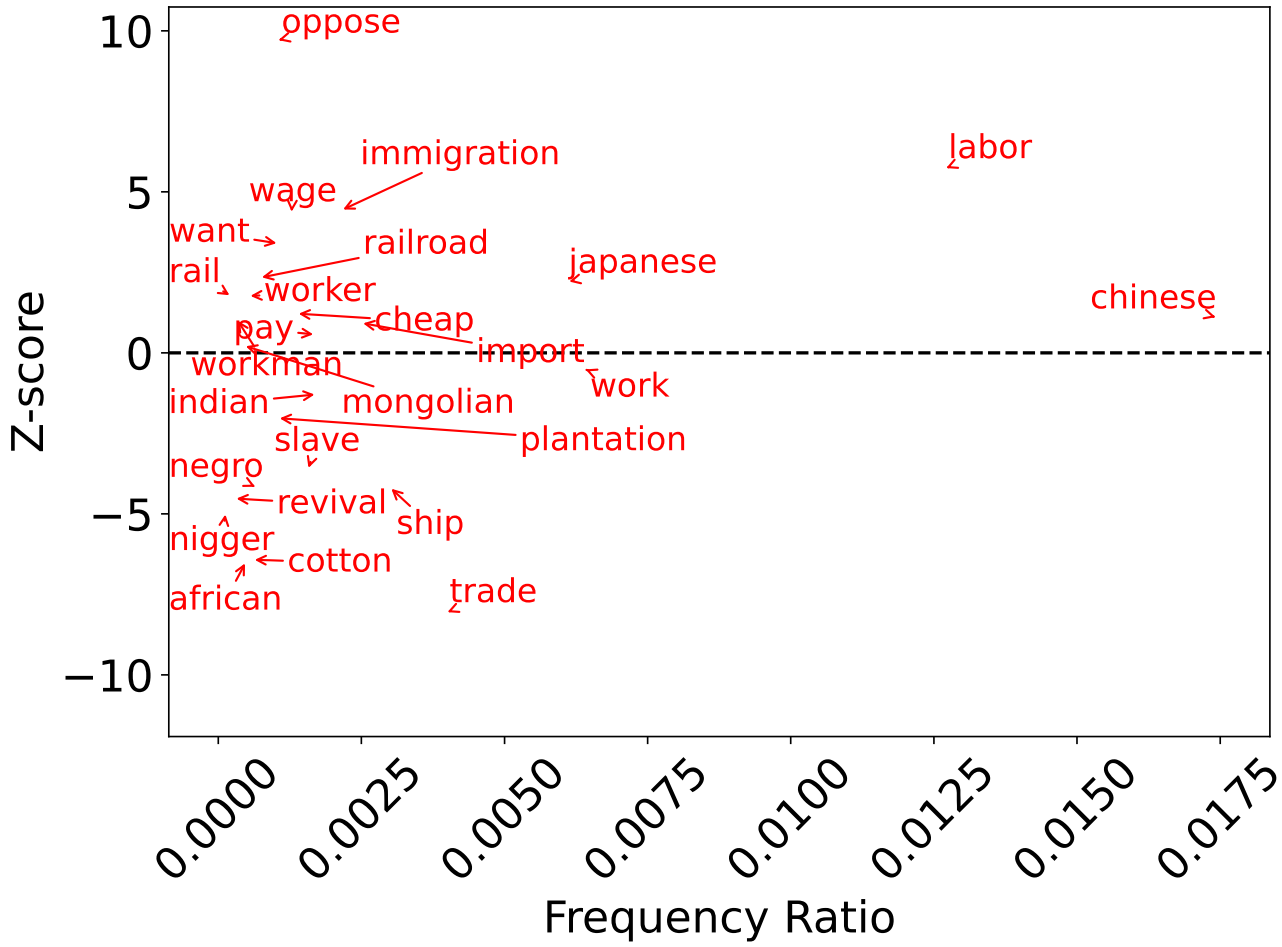


Figure 3: The Z-score of words in then-Confederate and then-Union newspapers

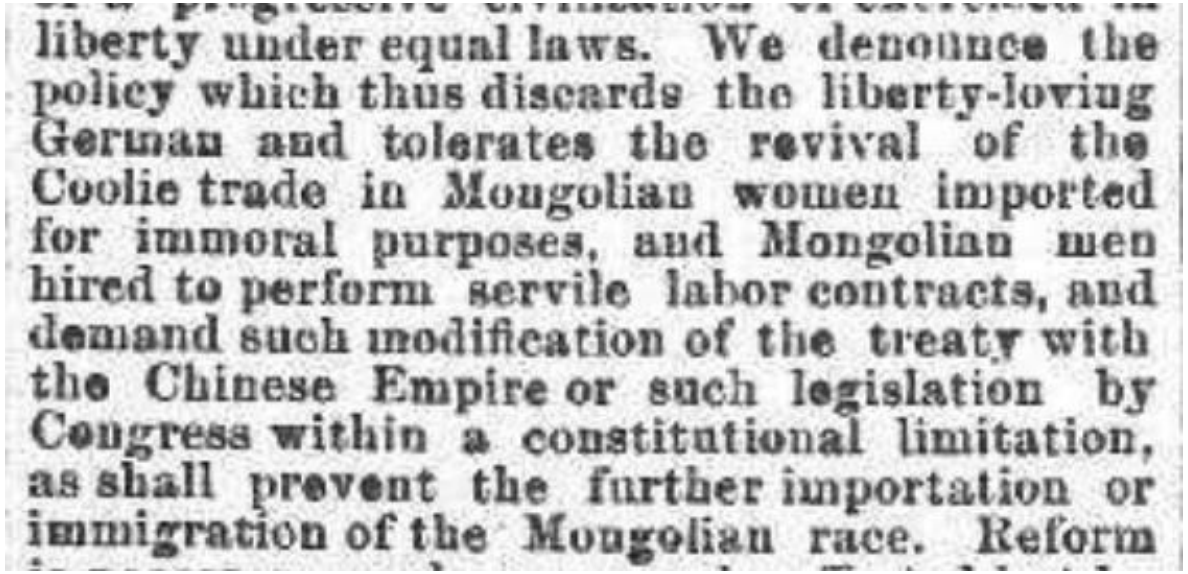
- Over-represented words in then-Confederate newspapers: “trade” (-8.05), “ship” (-4.12), “slave” (-3.704), “negro” (-4.16), “nigger” (-4.88), “african” (-6.41), “plantation” (-2.02), “cotton” (-6.42)

- Over-represented words in then-Union newspapers: “labor” (5.73), “wage” (4.38), “cheap” (1.22), “pay” (0.55), “rail” (1.71), “railroad” (2.31)

RQ 3. Results: Reprint network of coolie stories

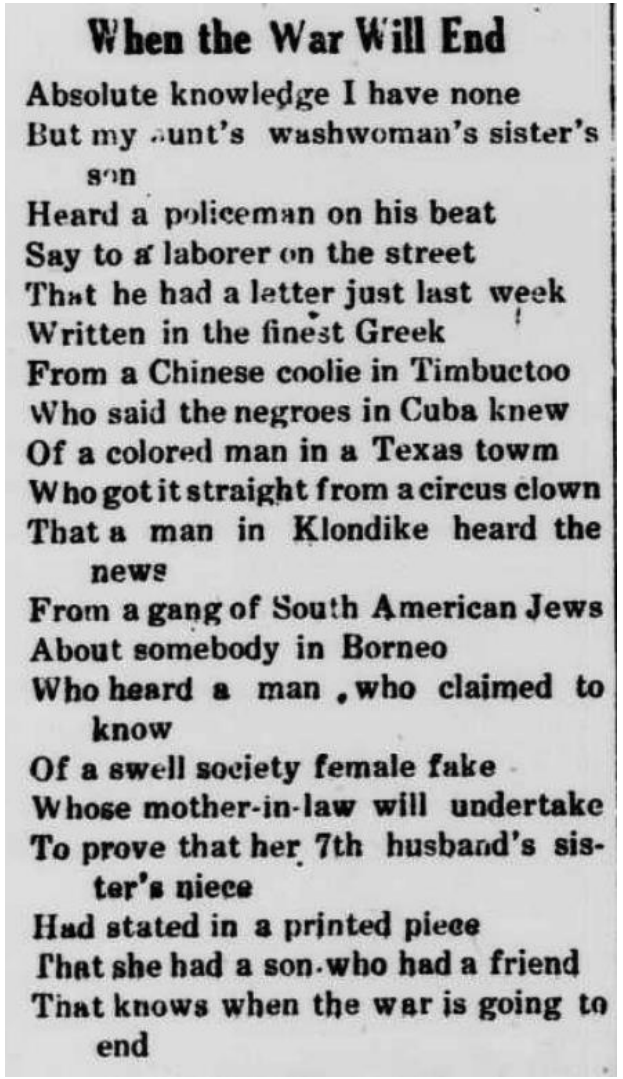
Figure 4: The reprint network of “coolie” stories in the newspapers

- The network of text reprints about coolie stories shows a high average clustering coefficient (0.9905) due to the presence of reprints spread to multiple States.



liberty under equal laws. We denounce the policy which thus discards the liberty-loving German and tolerates the revival of the Coolie trade in Mongolian women imported for immoral purposes, and Mongolian men hired to perform servile labor contracts, and demand such modification of the treaty with the Chinese Empire or such legislation by Congress within a constitutional limitation, as shall prevent the further importation or immigration of the Mongolian race. Reform

Figure 5: The text containing "coolie" in *The Opelousas courier* published on July 8th, 1876



When the War Will End

Absolute knowledge I have none
But my aunt's washwoman's sister's son
Heard a policeman on his beat
Say to a laborer on the street
That he had a letter just last week
Written in the finest Greek
From a Chinese coolie in Timbuctoo
Who said the negroes in Cuba knew
Of a colored man in a Texas towm
Who got it straight from a circus clown
That a man in Klondike heard the news
From a gang of South American Jews
About somebody in Borneo
Who heard a man, who claimed to know
Of a swell society female fake
Whose mother-in-law will undertake
To prove that her 7th husband's sister's niece
Had stated in a printed piece
That she had a son who had a friend
That knows when the war is going to end

Figure 5: The text containing "coolie" in *Middletown transcript* published on April 13th, 1918

The Opelousas courier (Democratic Party National Convention)
: 97 reprints
-Asians are inferior to "liberty-loving" Germans

Middletown transcript (poem)
: 78 reprints
-An emphasis on the exoticism of the places and people who are not White can Perceive this as micro aggression as they are marginally represented in the US population.

Conclusion

- MA, RI, WO, OK, and AR showed the most dissimilar meaning of the average meaning of “coolie.”
- IL, CA, WI, VI, and NV showed the most similar meaning of the average meaning of “coolie.”
- We found that the discourse of coolie in the then-Confederate newspapers was accompanied by words related to African American slavery as well as the workforce is the most needed.
- We found discriminating expressions toward Asian workers in political statements and poems were most circulated and they showed stereotypes of Asian workers in the United States history.

Limitation and future work

- Data-inherited limitation: (1) OCR errors, (2) a small number of digitized available data (e.g., MA (n=305) and RI (n=418) showed dissimilar meanings of the word “coolie” could have been attributed to the comparably small number of data)

- Geographic representation: Grouping then-Union and then-Confederate States might not be the best way to group the States as it doesn’t reflect the Westward expansion of the US.

- The impact of Chinese Exclusion Act of 1882

Thank you very much! The code is available!

- https://github.com/park-jay/coolie

- jaihyun@Illinois.edu
- X (Twitter): @91jpark19