

CS 2731 Introduction to Natural Language Processing

Session 27: Computational social science, digital humanities

Michael Miller Yoder

December 6, 2023



University of
Pittsburgh

School of Computing and Information

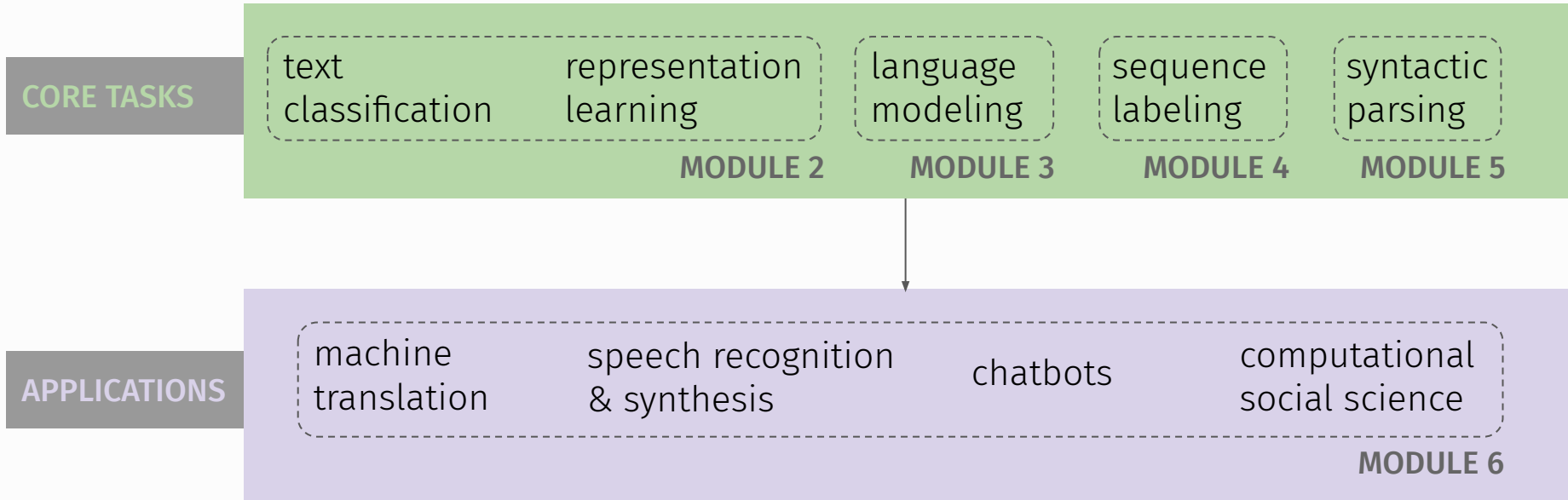
Course logistics

- Last regular class session!
- No class next Mon Dec 11
- Final project presentations next **Wed Dec 13, 2:30-4pm**
- Project report is **due next Thu Dec 14 at midnight**

Final project report rubric

Rubric category	Points
Clear motivation for the work is provided	5
Research questions and/or task definition is clear	10
Sufficient grounding in relevant related literature	15
Applicable dataset/s are chosen	5
Methods are relevant. For new approach contributions, multiple methods are compared. For dataset contributions, annotation methodology is explained	15
Results are provided. For new approach contributions, results from multiple methods (at least one baseline) are presented. For dataset contributions, this may be a single set of results from a simple classifier, or other results if discussed with the instructor	20
Discussion is provided of the results and/or the potential uses or contributions of any new datasets contributed	10
Limitations of your approach or dataset are sufficiently discussed	5
Ethical issues that may be raised by your system or dataset are sufficiently discussed	5
Project content total	90
Meets all formatting requirements. Is maximum 8 pages, not including references or group member task breakdown	15
Writing is clear	15
Writing total	30
Group member had a sufficient amount of workload in the project	15
Task and roles assigned to this group member were completed sufficiently	15
Individual contribution total	30
Grand total	150

Core tasks and applications of NLP



Overview of today's class session

- Language in social context
- Computational social science
 - Example project
- Digital humanities
 - Example project
- Time to complete OMETs
- Project time
 - Michael available to answer questions

Language is embedded in social context

What types of social contexts is
language used in?

What types of social contexts?













A professional advertisement for Alan Black. On the left is a headshot of Alan Black, a middle-aged man with a mustache, wearing a dark suit, light blue shirt, and yellow tie. To the right of the photo, the text reads:

Alan Black
has your back.

Below this text is a red rounded rectangle containing the following services in white text:

- DWI
- Criminal Defense
- Family Law
- Personal Injury

Euro quals	2:45 PM ET ESPN3	2:45 PM ET ESPN3	2:45 PM ET ESPN2/ESPN3	2:45 PM ET ESPN3	2:45 PM ET ESPN3
	 BEL  CYP	 NED  EST	 WAL  HUN	 GER  NIR	 POL  SVN

On-Line Homework Instructions for Physics 1250-1251

Homework will be submitted and graded via the online software package WebAssign.

ACCESSING WEBASSIGN:

Open Internet Explorer or Netscape Navigator or Mozilla Firefox (Some other browsers may have difficulty), and go to the WebAssign login page (<https://www.webassign.net/osu/student.html>). (The WebAssign login page at <https://www.webassign.net/login.html> will get you to the site above as well, but the OSU login site should be your primary site.)





What's happening?



Tweet



Odd Pittsburgh @OddPittsburgh · 59m

[#Pittsburgh](#) in 1930



City of Pittsburgh

Trends for you



Trending in United States



#DevinNunesIsAnIdiot

53.9K Tweets

Trending in United States



#AdviceForBoomers

4,684 Tweets

Trending in United States



Vindman

Trending with: Lt Col Vindman, Colonel Vindman, Col Vindman

Trending in United States



#2009v2019

3,035 Tweets

[Show more](#)

How Not to Plot Secret Foreign Policy: On a Cellphone and WhatsApp

U.S. officials expressed wonderment that Rudy Giuliani ran an “irregular channel” of Ukraine diplomacy over open cell lines and apps penetrated by the Russians.

2h ago [494 comments](#)



Rudolph W. Giuliani, President Trump's personal lawyer, makes a living selling cybersecurity advice.

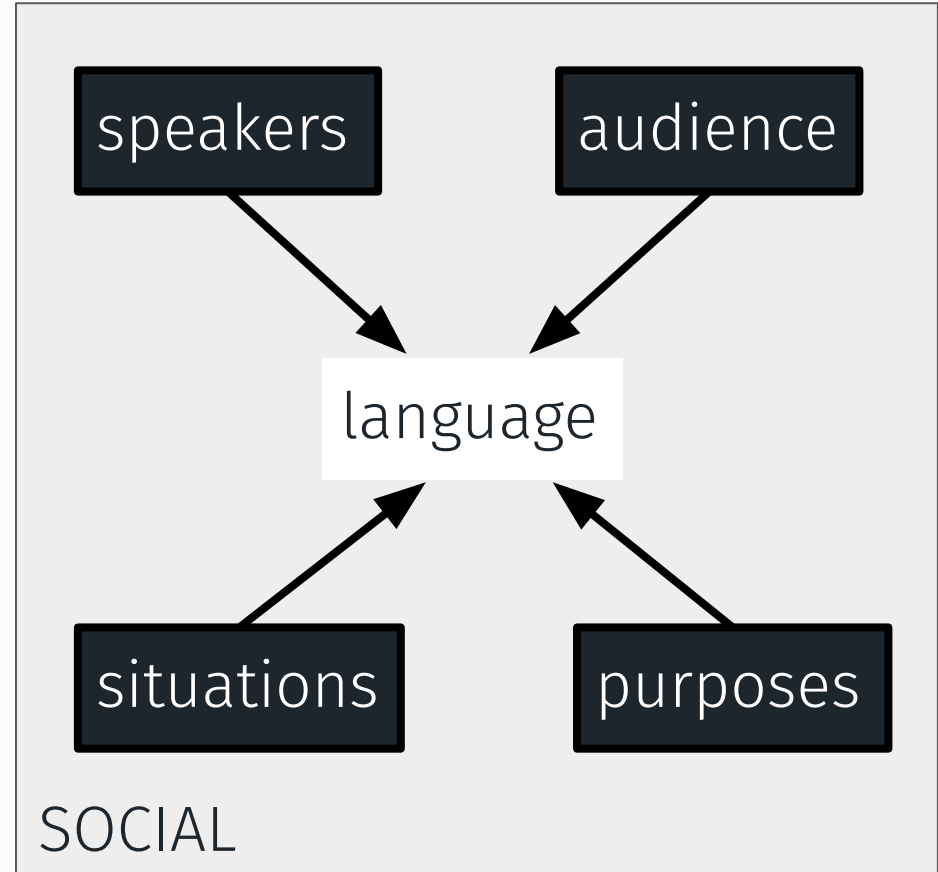
Doug Mills/The New York Times

Who is Kurt Volker, President Trump's former special envoy to Ukraine?

28m ago

Tim Morrison, a hawkish aide loyal to Mr. Trump, will also testify this afternoon.

42m ago



Computational social science

Computational social science

- Investigating (modeling, analyzing) social phenomena with computational tools [Cioffi-Revilla 2017]
- CSS goal: find out something about **people** (social science)
- NLP goal: build computational tools that can process or produce language
 - Social NLP: build tools that address language in social context
 - Hate speech detection, sarcasm and irony, dialects and language variation

Computational social science: methods and data

- Observational studies, not lab or survey studies



large datasets of social interaction

Computational social science example

Example: How fast does fake news spread? [Vosoughi et al. 2018]

y

=

$f(x)$



spread through a network



true/fake news

network analysis

NLP/text mining

Computational social science example

Example: Do police officers speak more respectfully to white drivers than black drivers in traffic stops? [Voigt et al. 2017]



Example project:
NLP + computational social science

Self-presentation and interaction on Tumblr [Yoder et al., WebSci 2020]

$$y = f(x)$$



social outcome

- Content propagation

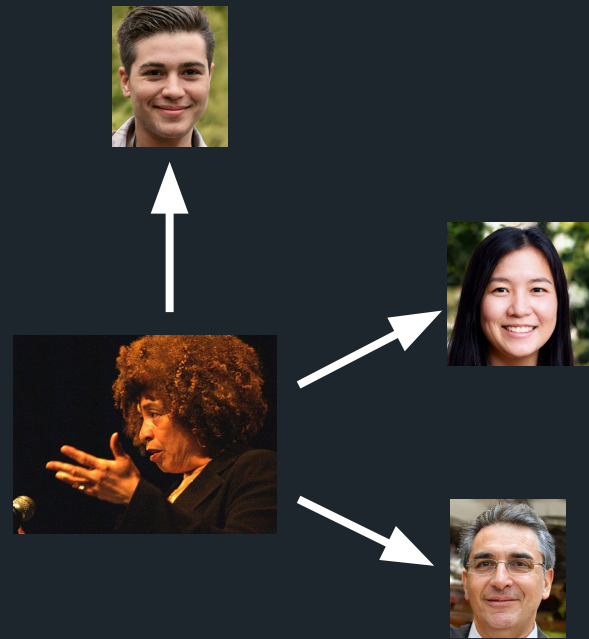


identity presentation

- Text blog descriptions
- Profile images

Motivation: identity + content propagation

- Content and network features predictors of content propagation [Zhang et al. 2014, Xie et al. 2017]
- Social media is also a place for identity construction: effects on propagation?
- Homophily: users more likely to have links if share attributes [Gong et al. 2018]



What effects of similarities and differences in **self-positioning** do we see on **content propagation** in Tumblr?



text self-descriptions

max | 23yo | she/they |
twerfs don't follow

profile images



control features

#untitled goose game #untitled goose simulator #hjonk #horrible goose
#press y to honk #memes #shitposts #nettle quacks #1k
#my first 1k post! #500 notes #100 notes

4,826 notes Oct 3rd, 2019



reblogging

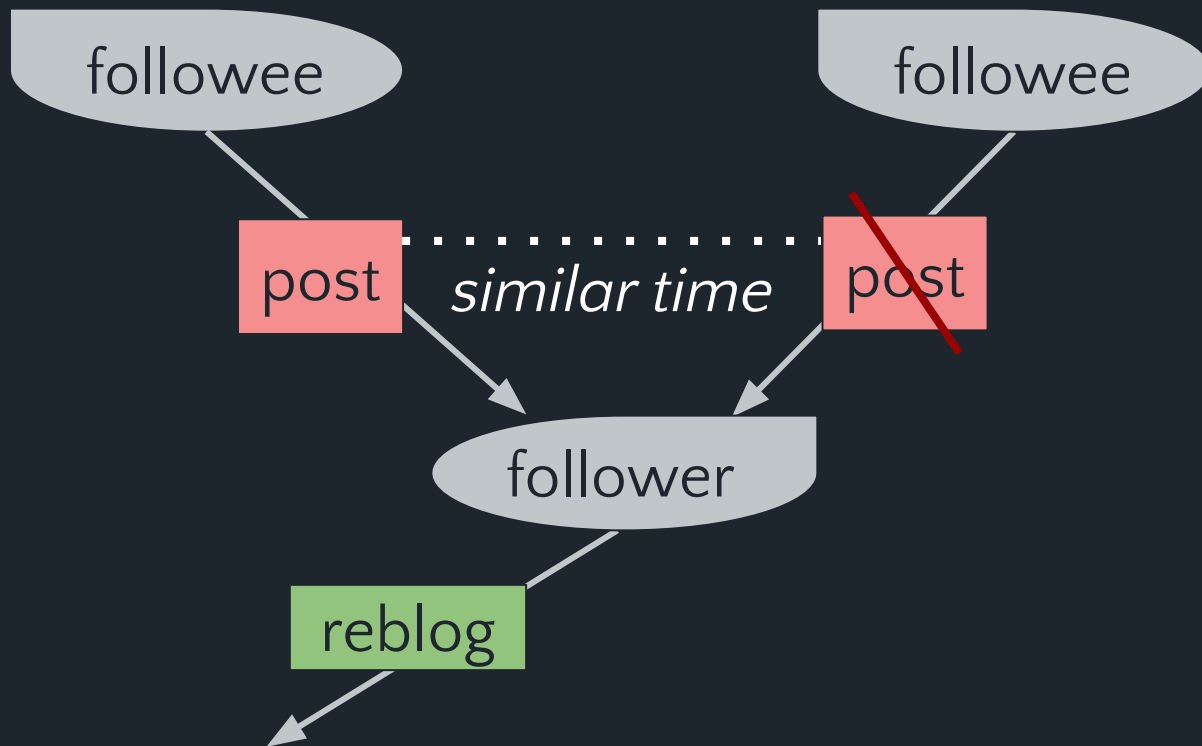
post



reblogged post

Reblog prediction

- Reblog "opportunity"
- Learning to rank pairwise formulation



Descriptions*

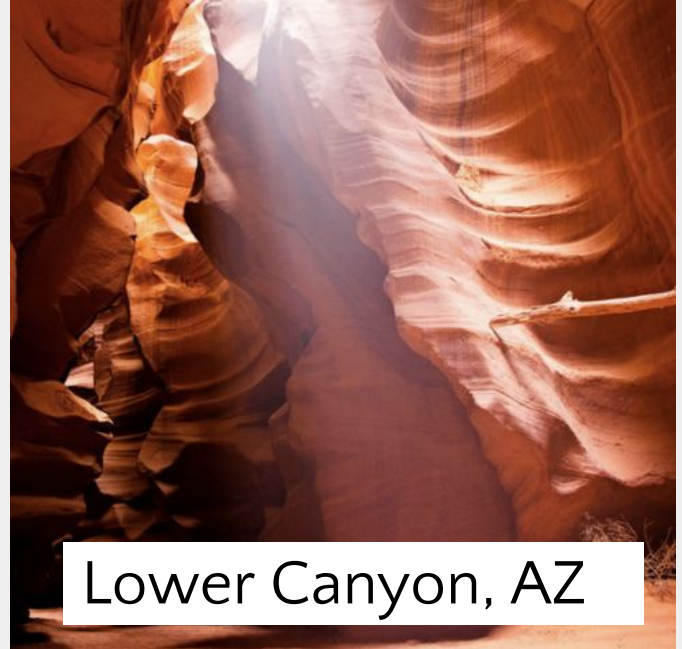
Follower:

Travel Enthusiast – Photography
– Web Design

Followed:

world traveler. | Wanderlust |
Landscape | Photography

Reblogged post



Lower Canyon, AZ

* changed for privacy

Data

Number of users	34,801
Number of reblog opportunities	712,670
Timeframe	June - Nov 2018

Identity features

22 male infj coffee 

FOLLOWED

they/them 29 leo infj

FOLLOWER

Identity features

match: age

22 male infj coffee 

FOLLOWED

they/them 29 leo infj

FOLLOWER

Identity features

match: personality type

22 male *infj* coffee 

FOLLOWED

they/them 29 leo *infj*

FOLLOWER

Identity features

mismatch: pronouns

X

22 male infj coffee 🌈

FOLLOWED

they/them 29 leo infj

FOLLOWER

Identity features

match: infj

22 male *infj* coffee 

FOLLOWED

they/them 29 leo *infj*

FOLLOWER

Identity features

followed: 22, follower: 29

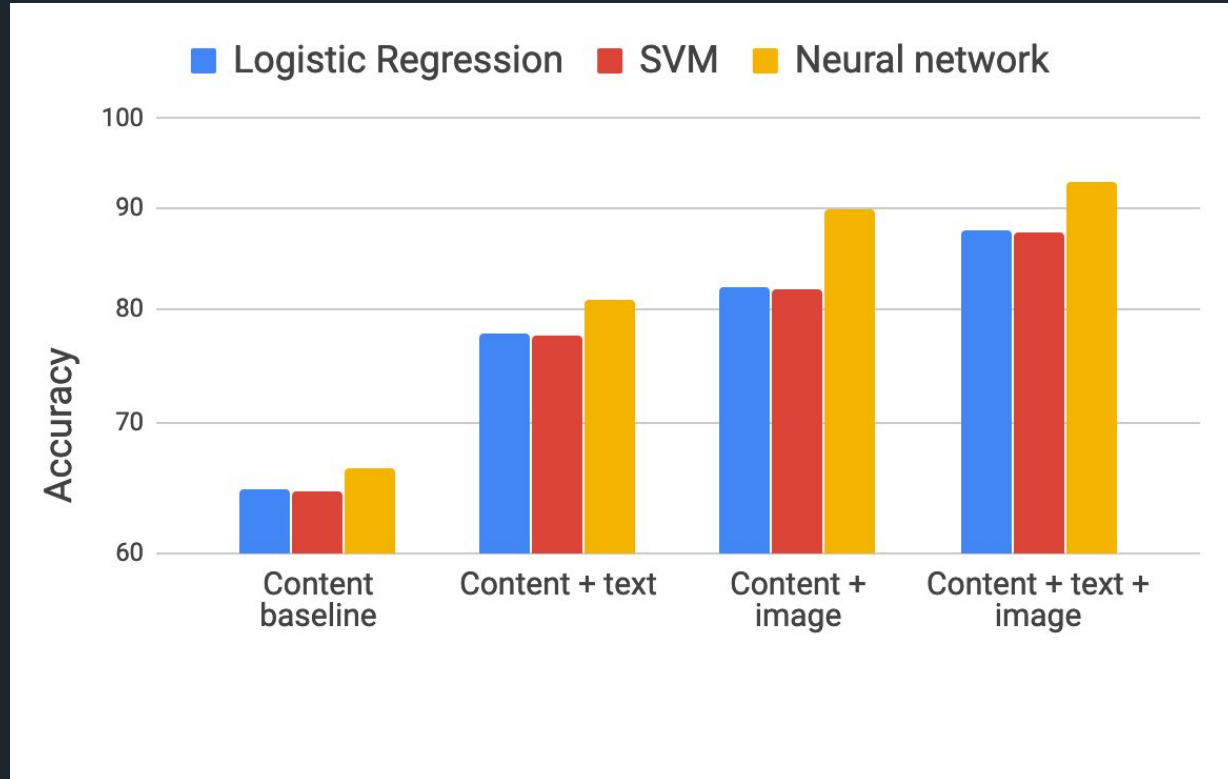
22 male infj coffee 

FOLLOWED

they/them 29 leo infj

FOLLOWER

Is there an effect?



Interpretation: text features

- Shared values/experiences: categories and label **matches** are positively associated with reblogging

any sexual orientation



any sexual orientation

indie



indie

What is the nature of this effect?

Features	Likelihood of reblogging
Follower: presents pronouns Followed: does not	↓
Both: <i>cis</i> or <i>cishet</i>	↑
Race/ethnicity label alignment	↑
Nationality label alignment	<i>none</i>

What is the nature of this effect?

Features	Likelihood of reblogging
Follower: <i>gaming</i> Followed: <i>manga</i>	↑
Follower: <i>memes</i> Followed: <i>history</i>	↓

Takeaways

Identity on

- Shared interests or shared values indicated that users were more likely to share each other's content

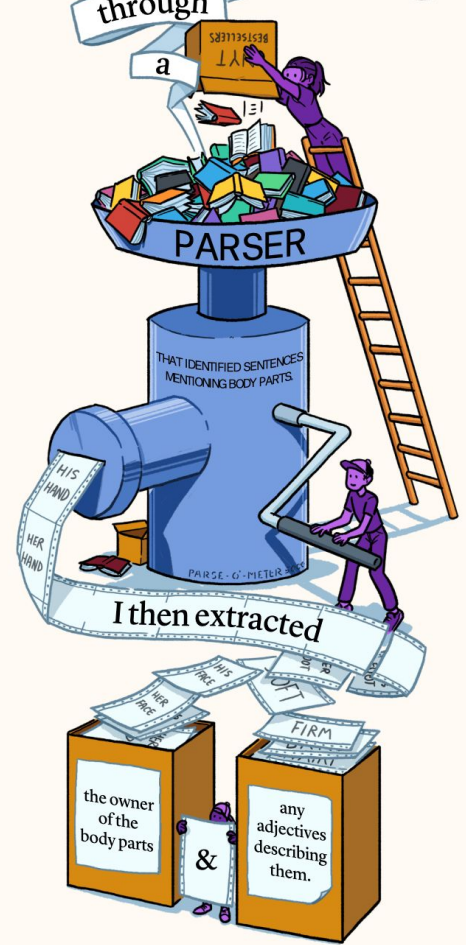
NLP+CSS takeaways

- The *choice* of self-presentation can have an effect
- Users' own terms matter: specific combinations of self-presented terms related to content sharing

Digital humanities

Digital humanities

- Analyzing works from the humanities with digital methods
- Computational literary analysis
- Example: Genre prediction
 - What novels don't fit genres?
 - How have genres changed over time? [Underwood 2016]
- NLP/text mining
 - Patterns in language data

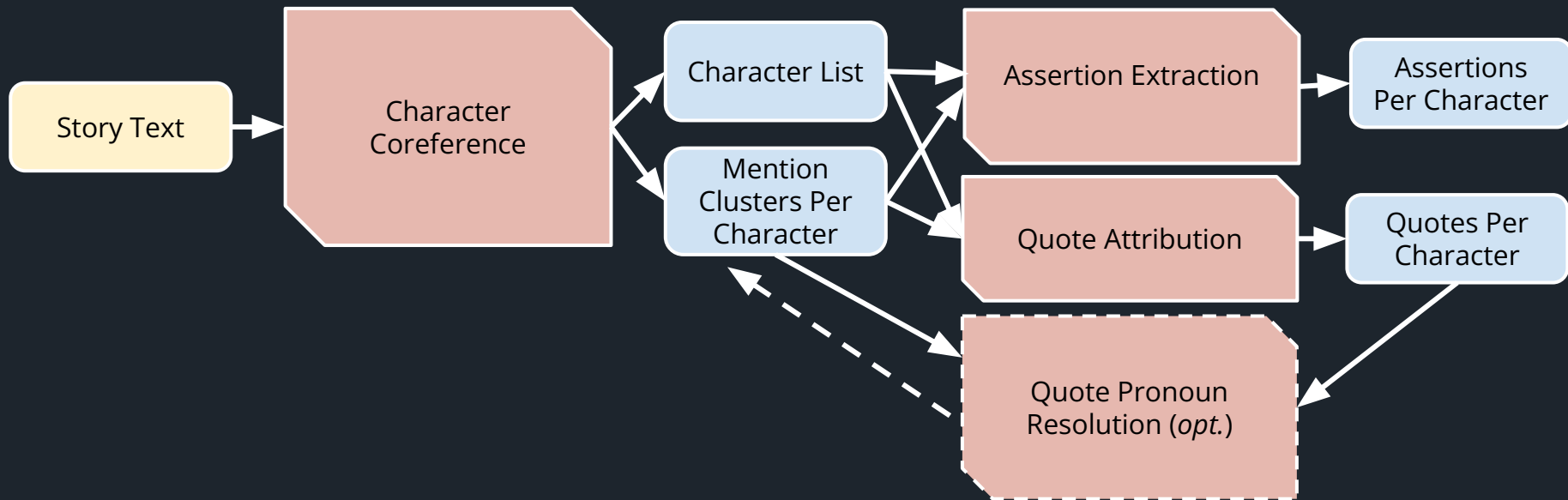


Example project:
NLP + digital humanities

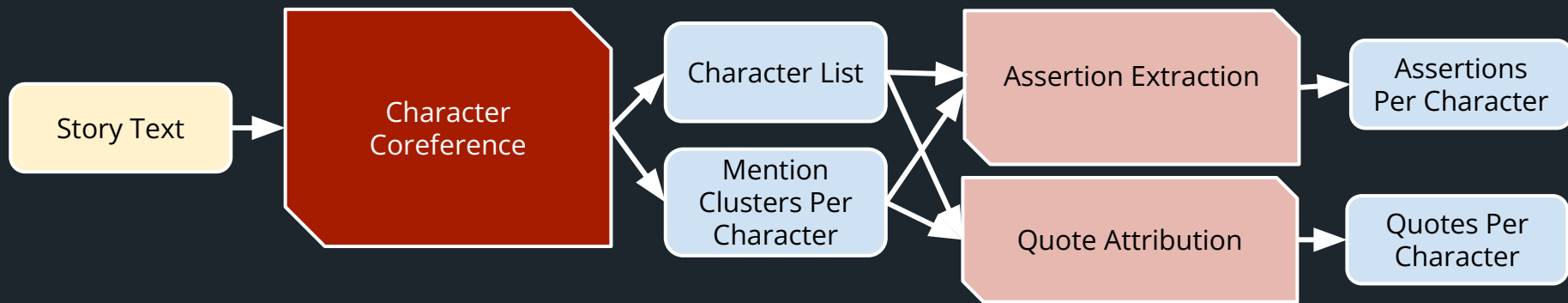
Fanfiction

- Stories based on existing media [Fiesler+ 2016]
- “Participatory culture” [Jenkins 2003]
- How to extract text that portrays particular characters?

Fanfiction NLP pipeline [Yoder et al., WNU 2021]



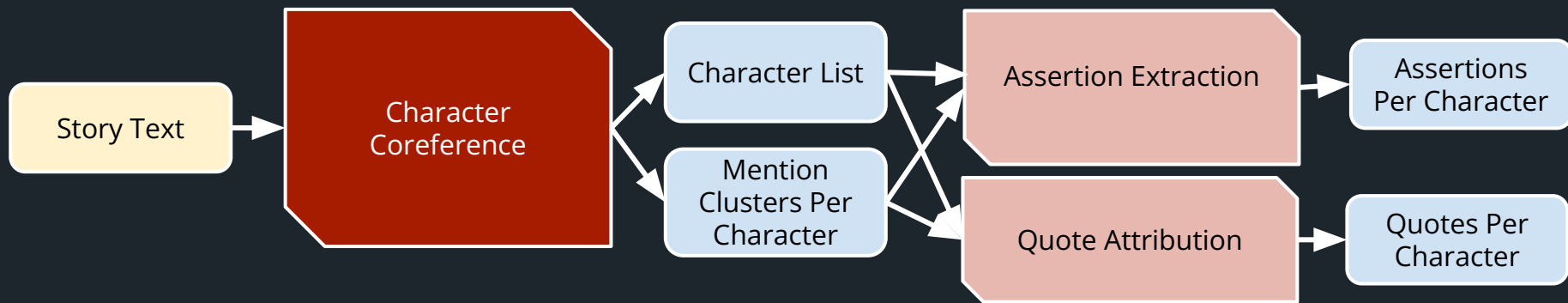
Character coreference



SpanBERT coreference
resolution [Joshi+ 2020]
fine-tuned on LitBank
[Bamman+ 2020]

“Why are **you** under a table, kid?”
Jake asked, then grabbed
Charlie’s walkman without
waiting for an answer.

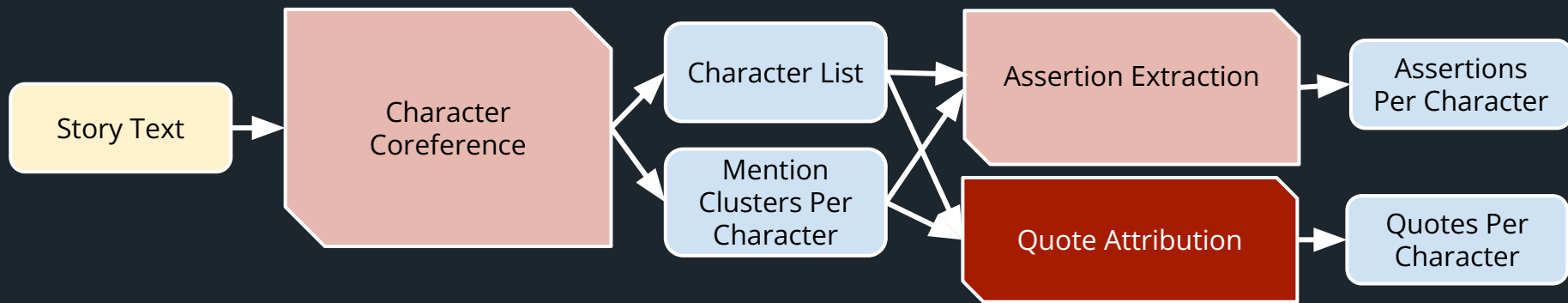
Character coreference



SpanBERT coreference resolution [Joshi+ 2020]
fine-tuned on LitBank [Bamman+ 2020]

Coreference resolution system	CoNLL F1
BookNLP [Bamman+ 2014]	38.5
BERT-base (LitBank fine-tune)	58.4
SpanBERT-base (LitBank fine-tune)	64.8
FanfictionNLP	71.4

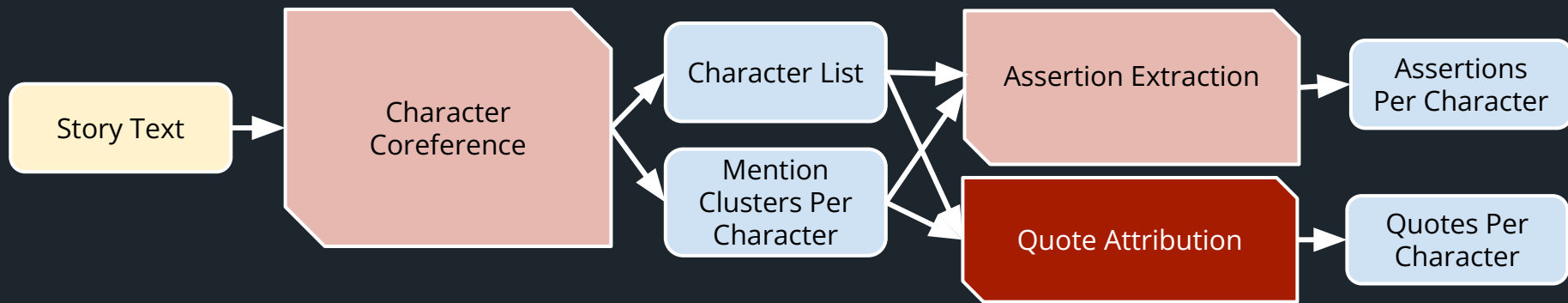
Quote attribution



Sieve-based
deterministic
approach
[Muzny+ 2017]

“Why are you under a table, kid?”
Jake asked, then grabbed Charlie’s
walkman without waiting for an
answer.

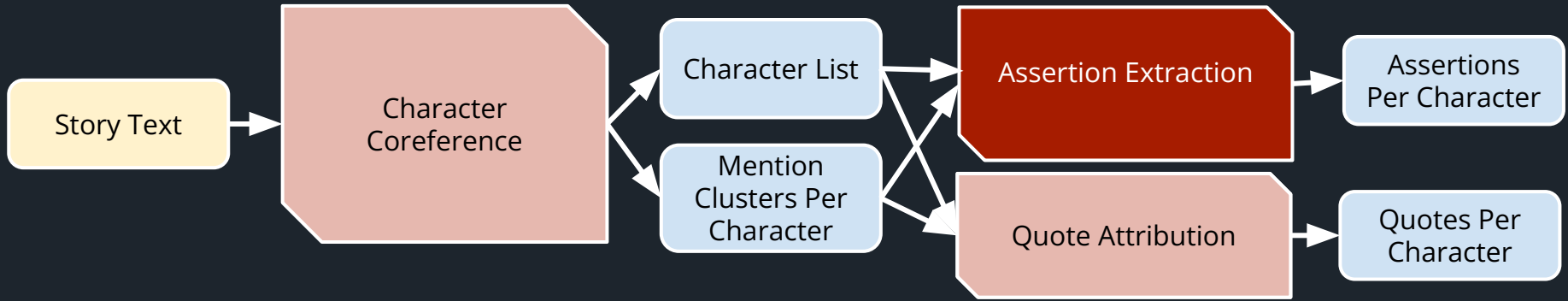
Quote attribution



Sieve-based
deterministic
approach
[Muzny+ 2017]

Quote attribution system	F1
BookNLP [Bamman+ 2014]	34.7
[He+ 2013]	53.6
FanfictionNLP [Muzny+ 2017]	67.8

Assertion extraction



- Select boundaries of spans based on word frequency changes [Hearst 1997]

“Why are you under a table, kid?”
Jake asked, then grabbed Charlie’s
walkman without waiting for an
answer.

Wrapping up

- Language is embedded in social context
- Computational social science studies people and societies with computational models of observational data
 - Often uses NLP for analyzing text
- Digital humanities uses NLP methods to study literary works and other humanities artifacts

Please fill out OMETs

- Course evaluations (OMETs) are open
- Will close this Sun Dec 10

<https://go.blueja.io/BEBlAj4xFEydvsaSR780YA>



See you next Wednesday

Thanks for a great semester!