



ENTITY AND RELATION EXTRACTION

Sandeep Soni

04/09/2024

WORDS

They can fish

SEQUENCE LABELING

- Given a textual sequence, categorize each token in the sequence

$$f([x_1, x_2, \dots, x_n]) \rightarrow [y_1, y_2, \dots, y_n]$$

- Examples:
 - Part of speech tagging
 - Named entity recognition

NAMED ENTITIES

Apple **ORG** was founded on April 1, 1976 **DATE**, by Steve Jobs **PERSON** in California **GPE**.

- Named entities are proper nouns that reference an entity that can be grounded
- Named entity recognition (NER) is the task of identifying spans of text that are named entities and categorizing them

BIO NOTATION

B	Beginning of a named entity
I	Inside of a named entity
O	Outside of the named entity

B-ORG O O O B-DATE I-DATE O B-PER I-PER O B-LOC

Apple was founded on April 1976, by Steve Jobs in California

- If there are n categories, then there are $2n+1$ BIO tags

NER TAGS

Person	Proper nouns refer to people or characters (e.g., Harry Potter, Sandeep Soni)	PER
Location	Refer to natural geographical regions (e.g., mountains) or structures built by humans (e.g., building)	LOC
GPE	Geo political entities such as cities, states, countries,etc (e.g., London)	GPE
Organization	Organization names, company names, sports teams, etc (e.g., The United Nations)	ORG

SPACY TAGS

CARDINAL

DATE

EVENT

FAC

GPE

LANGUAGE

LAW

LOC

MONEY

NORP

ORDINAL

ORG

PERCENT

PERSON

PRODUCT

QUANTITY

TIME

WORK_OF_ART

PERSON:	People, including fictional.
NORP:	Nationalities or religious or political groups.
FAC:	Buildings, airports, highways, bridges, etc.
ORG:	Companies, agencies, institutions, etc.
GPE:	Countries, cities, states.
LOC:	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:	Objects, vehicles, foods, etc. (Not services.)
EVENT:	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART:	Titles of books, songs, etc.
LAW:	Named documents made into laws.
LANGUAGE:	Any named language.
DATE:	Absolute or relative dates or periods.
TIME:	Times smaller than a day.
PERCENT:	Percentage, including "%".
MONEY:	Monetary values, including unit.
QUANTITY:	Measurements, as of weight or distance.
ORDINAL:	"first", "second", etc.
CARDINAL:	Numerals that do not fall under another type.

FINE GRAINED NER

- Expanding the number of categories for named entities gives us more ways to distinguish between them

person	doctor engineer monarch musician politician religious_leader soldier terrorist	organization		terrorist_organization government_agency government political_party educational_department military news_agency	
location	body_of_water city country county province railway road bridge	product	camera mobile_phone computer software game spacecraft instrument train	art film play	written_work newspaper music
				event	military_conflict attack natural_disaster election sports_event protest terrorist_attack
building	time color award educational_degree title symptom disease drug body_part living_thing animal food	chemical_thing biological_thing medical_treatment disease symptom drug body_part living_thing animal food	website broadcast_network broadcast_program tv_channel currency stock_exchange algorithm programming_language transit_system transit_line		
airport dam hospital hotel library power_station restaurant sports_facility theater					

DOMAIN SPECIFIC NER

- The categories are highly dependent on the domain of interest
- Of course, the **black King** is out in the open, but the **white Bishop** is hemmed in by the **pawns**.

Tag	Meaning
Hu	Human
Tu	Turn
Po	Position
Pi	Piece
Ps	Piece specifier
Mc	Move compliment
Pa	Piece attribute
Pq	Piece quantity
Re	Region
Ph	Phase
St	Strategy
Ca	Castle
Me	Move eval.
Mn	Move name
Ee	Eval. element
Ev	Evaluation
Ti	Time
Ac	Player action
Ap	Piece action
Ao	Other action
Ot	Other notion

NESTED NER

- Typically NER assumes all categories are in a flat structure.
 - E.g., [Emory University]_{ORG}
- But many named entities have a nested structure
 - E.g., [[Emory]_{PER} University]_{ORG}

NAMED ENTITY RECOGNITION

- Many different ways to identify named entities
 - Max Entropy Models (e.g., Multinomial Logistic Regression)
 - Max Entropy Markov Models
 - Recurrent Neural networks
 - Transformer models

NAMED ENTITY LINKING

- Recognize spans that are named entities
- Also, resolve the named entities to their references
- Usually, by linking them to some external database

Paris is the capital of France

[Paris]_{LOC} is the capital of [France]_{LOC}

Paris, Arkansas

Paris, France

[Paris]_{LOC} is the capital of [France]_{LOC}

<https://en.wikipedia.org/wiki/Paris>



[Paris] is the capital of [France]



<https://en.wikipedia.org/wiki/France>

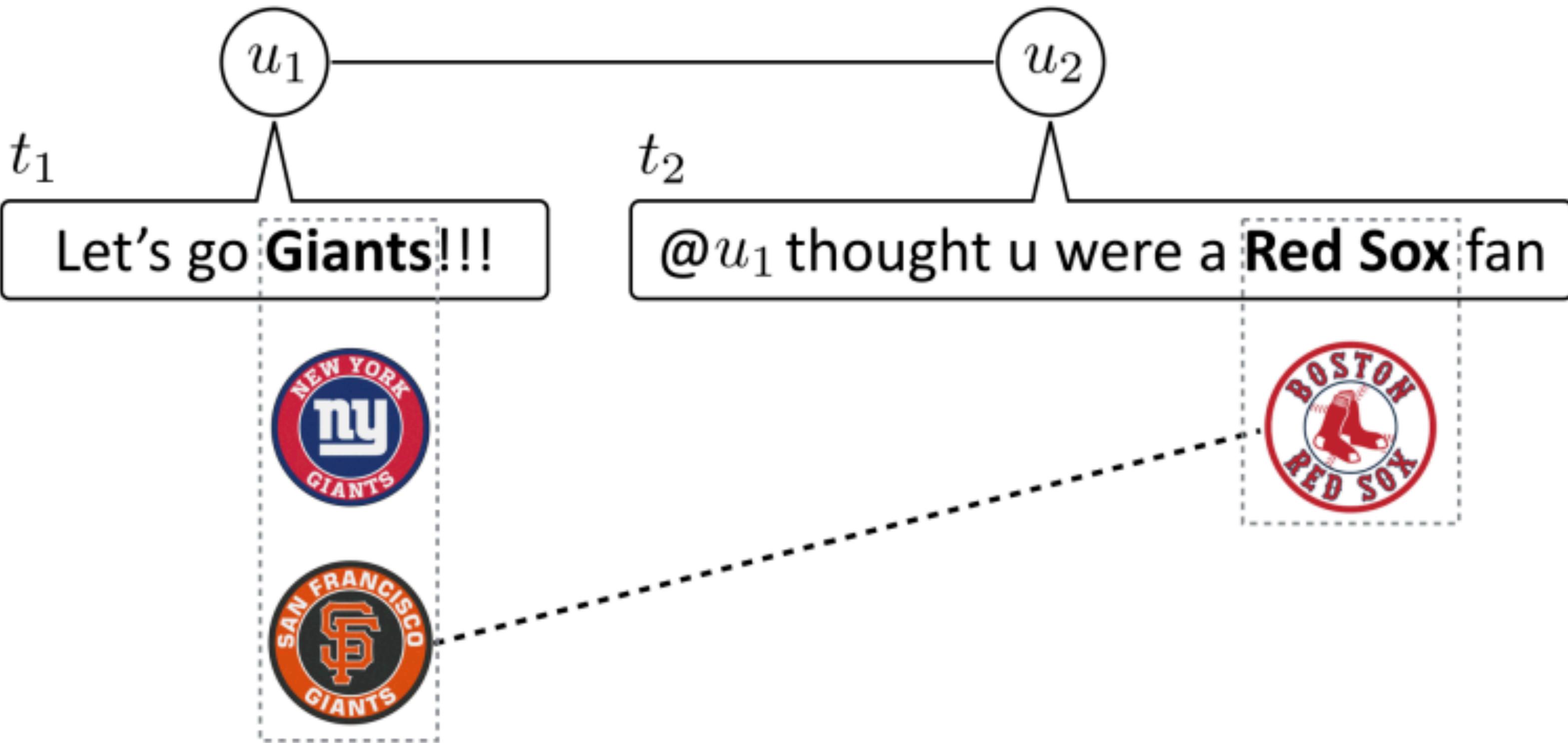


Figure 1: Illustration on leveraging social relations for entity disambiguation. Socially connected users u_1 and u_2 tend to talk about similar entities (baseball in the example).

TOPOONYM RESOLUTION

- Named entity linking but restricted to geographic entities.
- The objective is usually to link a span of text that marks a geographic entity to geographical coordinates on a map

TOPOONYM RESOLUTION

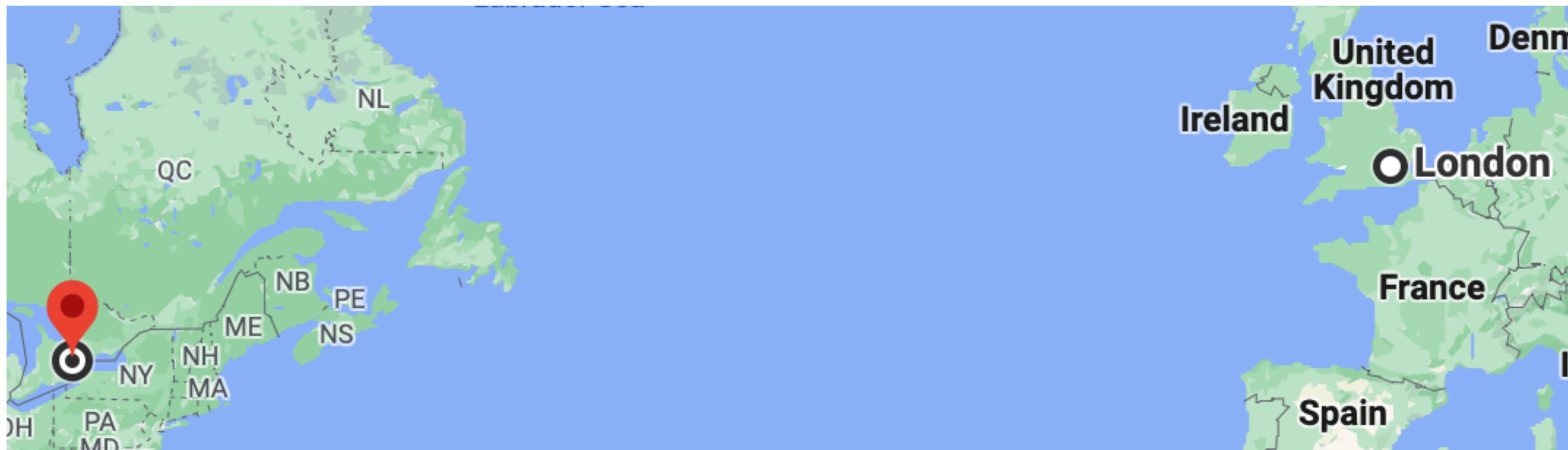
High-speed rail between Toronto and London by 2025, premier says

Wynne: 'Time to take a serious look at an idea that's been around for decades.'

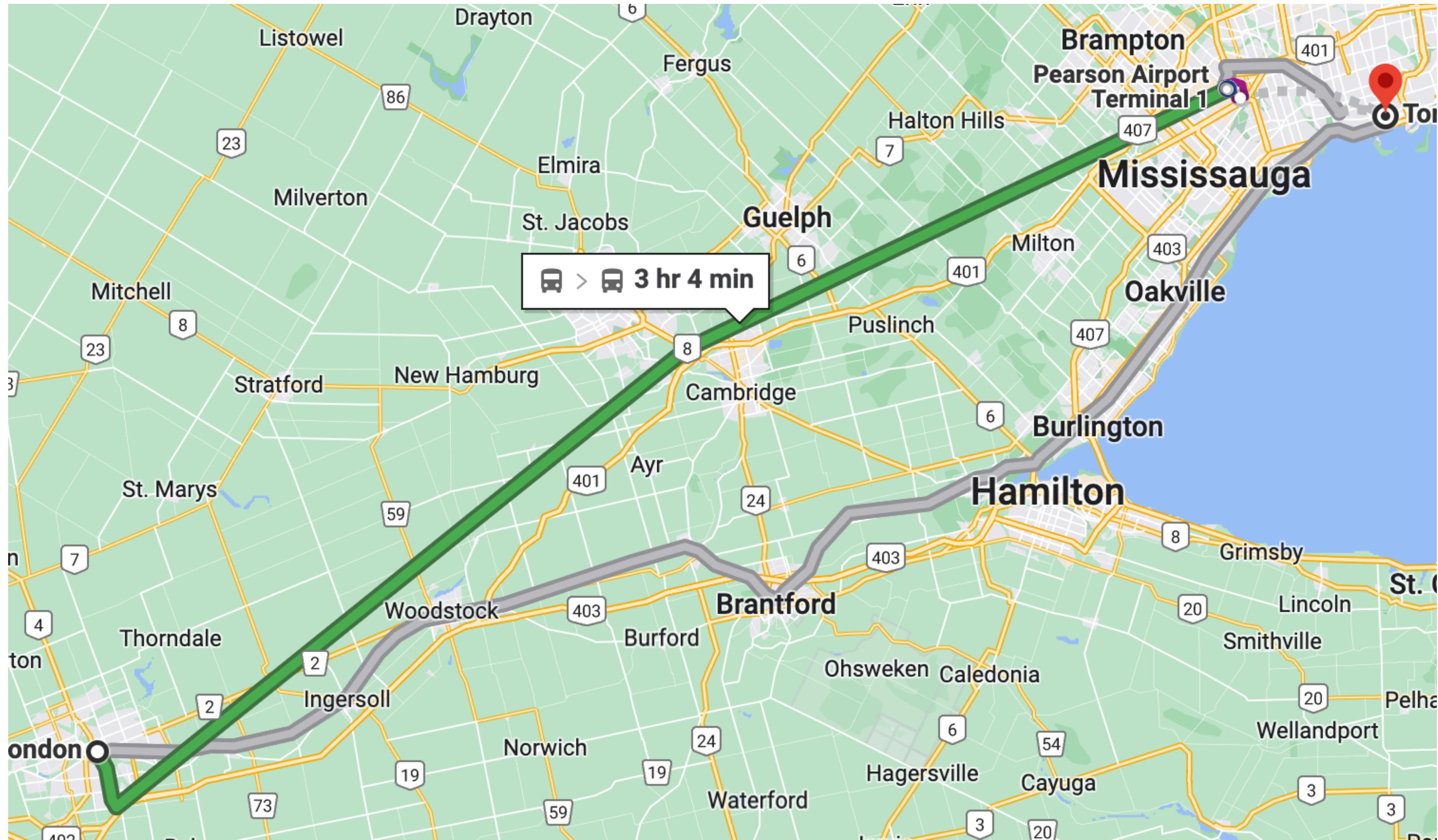
CBC News · Posted: May 19, 2017 9:22 AM EDT | Last Updated: May 19, 2017

<https://www.cbc.ca/news/canada/kitchener-waterloo/wynne-ontario-high-speed-rail-report-1.4123183>

TOPOONYM RESOLUTION

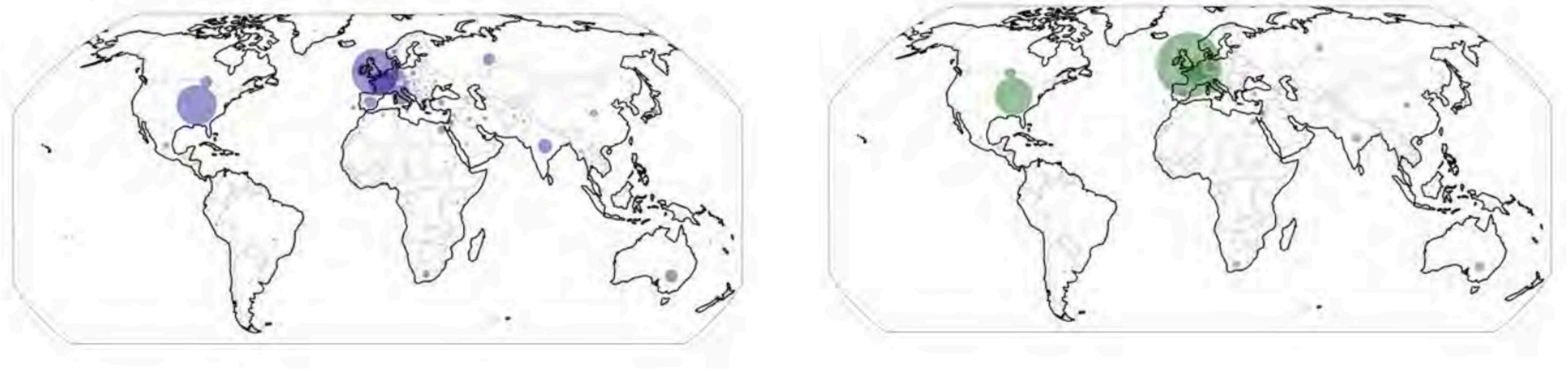


Can there be a high speed rail between London, UK to Toronto, Canada?



In general, we can resolve the place names by linking the mentions to a latitude and longitude

What can geographical named entities tell us about British Fiction?

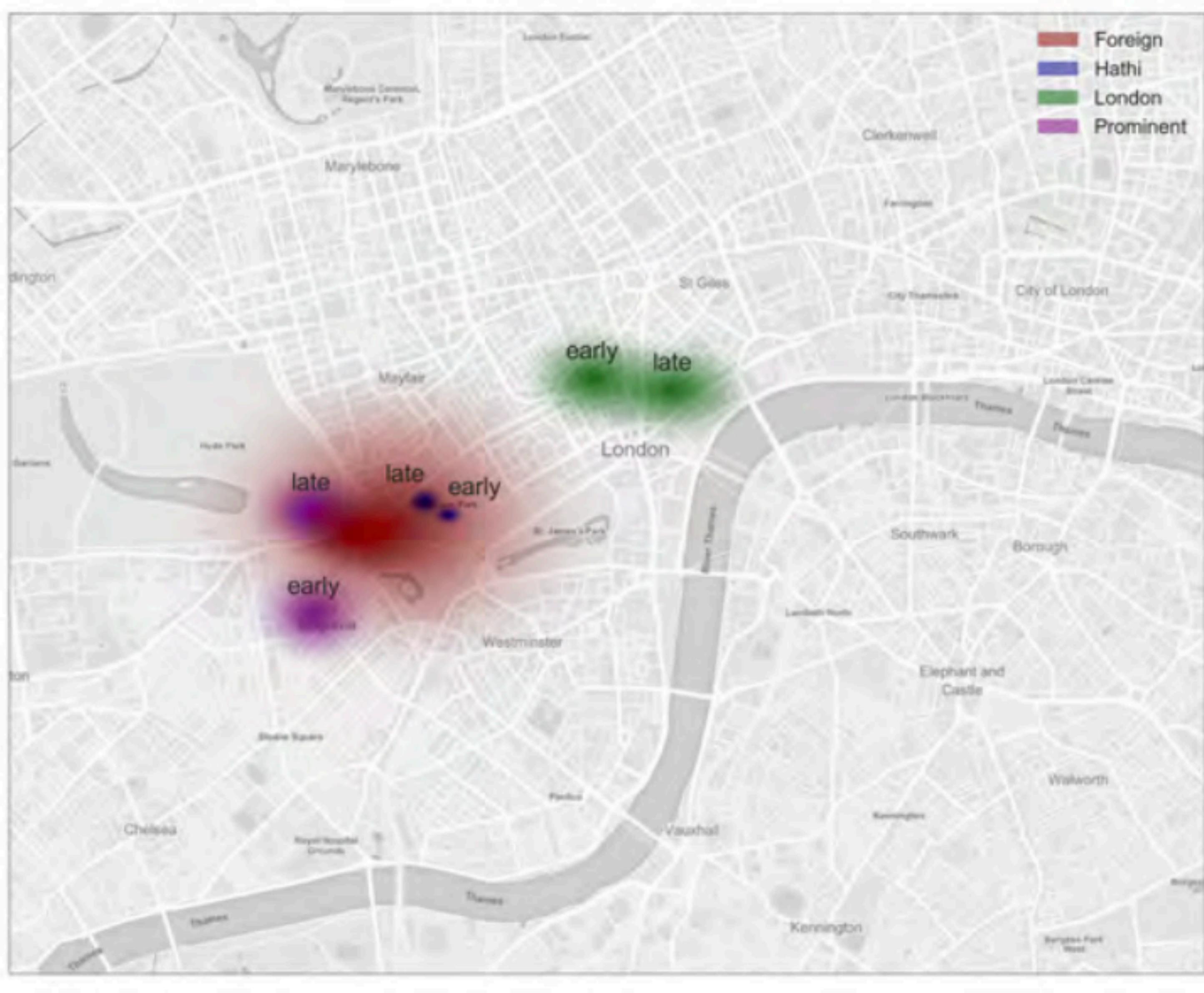


The distribution of global geographical attention in two large book corpora for British fiction

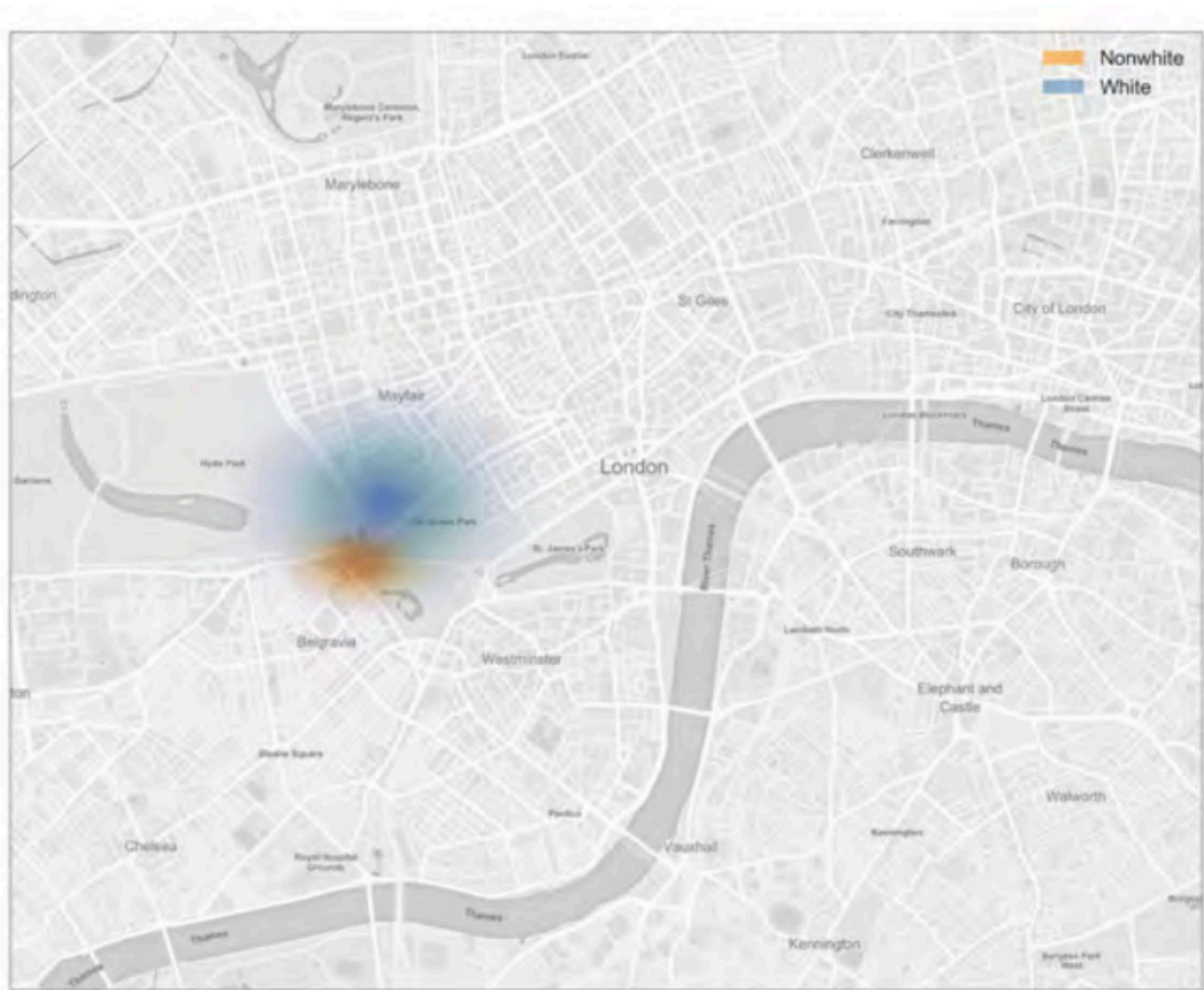
Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).



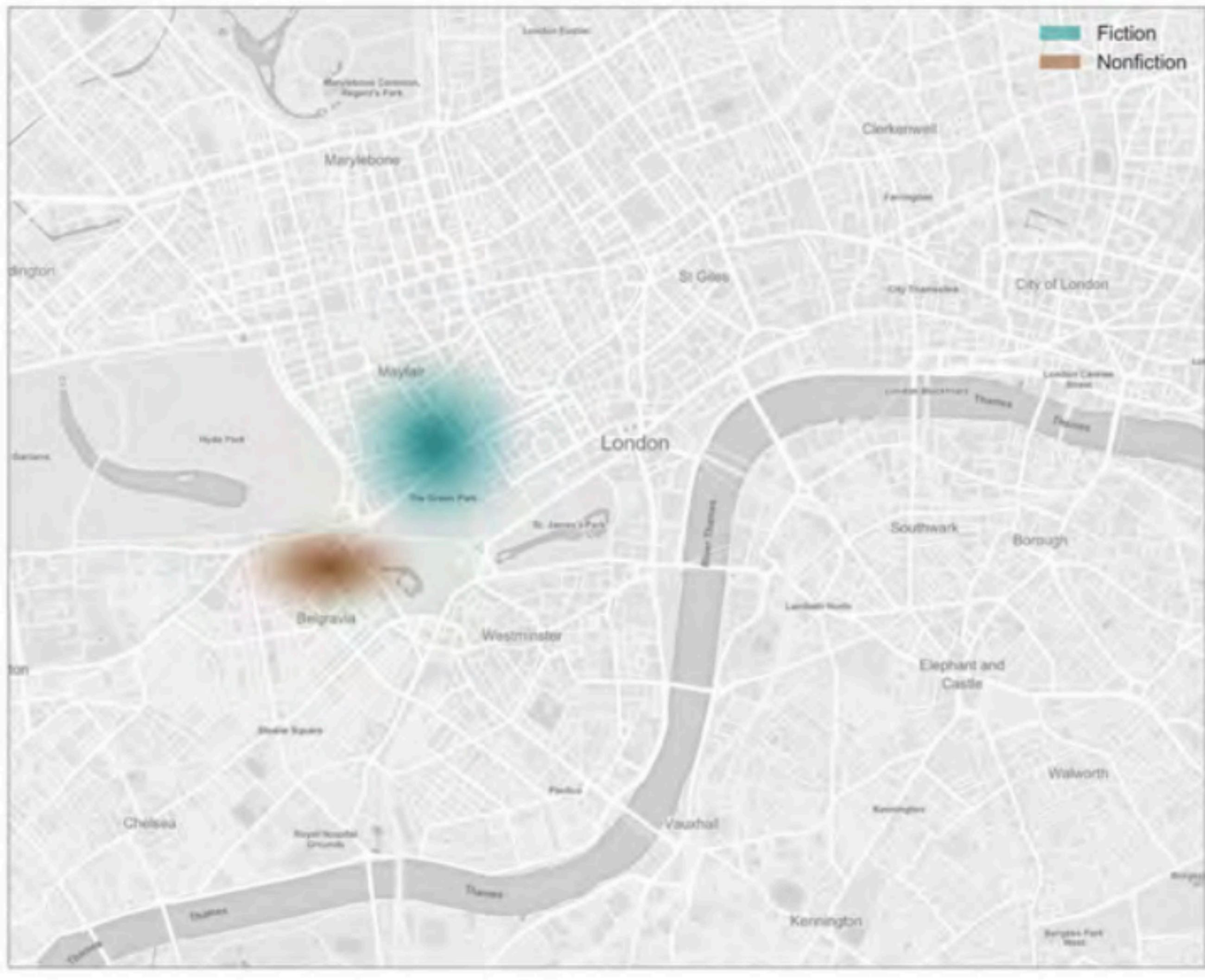
The distribution of attention over places in London in two large book corpora for British fiction



Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).



Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).



Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).

RELATION EXTRACTION

- Named entities mentioned in texts can be used to construct a knowledge base
- Two entities are connected to each other if they have a relationship

My [apartment]_{LOC} has a large [kitchen]_{LOC}

My [apartment]_{LOC} has a large [kitchen]_{LOC}



My [apartment]_{LOC} has a large [kitchen]_{LOC}

COMPONENT_OF (“Apartment”, “Kitchen”)

[Attention head]_{MODEL} is at the heart of [Transformers]_{MODEL}

COMPONENT_OF (“Transformers”, “Attention head”)

[Attention head]_{MODEL} is at the heart of [Transformers]_{MODEL}

COMPONENT_OF (“Transformers”, “Attention head”)

[Barack Obama]_{PER} was born in [Hawaii]_{LOC}

ENTITY_ORIGIN (“Barack Obama”, “Hawaii”)

[Barack Obama]_{PER} was born in [Hawaii]_{GPE}

ENTITY_ORIGIN (PER, GPE)

[Barack Obama]_{PER} was born in [Hawaii]_{GPE}

He was the president of the [United States of America]_{GPE}

He is the husband of [Michelle Obama]_{PER}

ENT_ORIGIN	Barack Obama	Hawaii
PRESIDENT	Barack Obama	United States of America
SPOUSE	Barack Obama	Michell Obama

(1) Stephen Dedalus, displeased and sleepy, **leaned** his arms on the top of the staircase and **looked** coldly at the **shaking gurgling** face that **blessed** him, equine in its length, and at the light untonsured hair, grained and hued like pale oak. (Joyce, *Ulysses*)

(2) My eyes **followed** his trim figure, richly though sombrely clad, then **fell** with a sudden **dissatisfaction** upon my own stained and frayed apparel. (Johnston, *To Have and to Hold*)

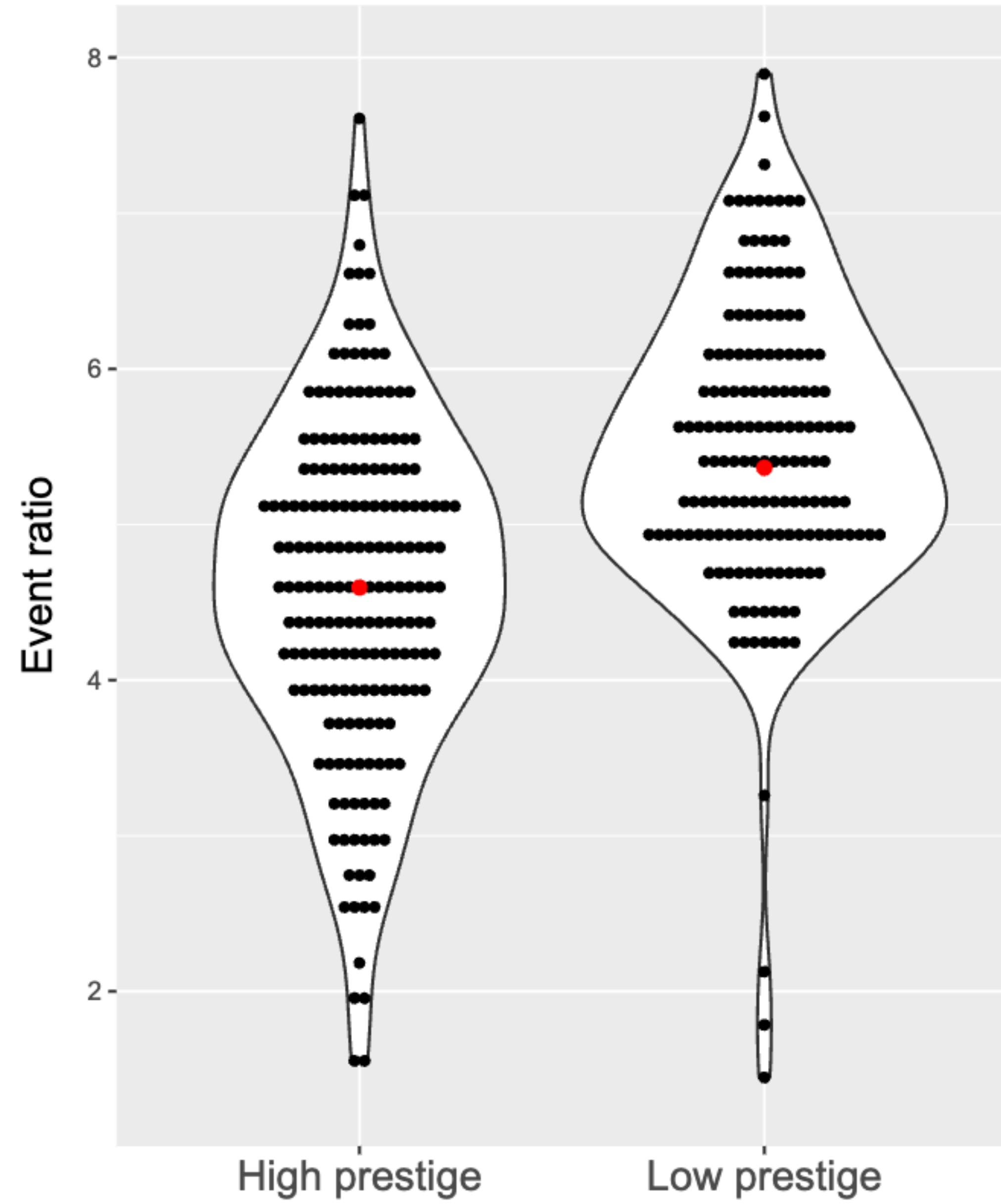
(3) He generally arrived in London (like the influenza) from the Continent, only he arrived unheralded by the Press; and his visitations set in with great severity. (Conrad, *The Secret Agent*)

Table 1: Three annotation examples with tagged event triggers in bold and candidate triggers that would not be tagged underlined.

Source: Sims et. al. Literary Event Detection ACL 2019

- Verbs connect entities and can be used to trace actions or events

- Model learns to detect events
- Model differentiates between grounded events and hypotheticals



Source: Sims et. al. (2019)

IN CLASS

- Sequence labeling demo
- Modify this notebook to find the distribution of named entities in a book from Project Gutenberg