




# **SOC 446W**

# **Presentation 2**

Carol Zhou





# Harry Potter and the Order of the Phoenix

## (2003)

Written by J.K. Rowling

### What is the corpus about?

- Harry's fifth year at Hogwarts amid skepticism about Voldemort's return
- Forms "Dumbledore's Army" to teach defensive magic to students because the school controlled by Dolores Umbridge, sent by the Department of Ministry.
- Final battle at the Department of Mysteries, revealing Harry's connection to Voldemort and the loss of a mentor

38 chapters and 17 words per sentence (average)

### Linguistic domain

- Narrative
- Characters and dialogues
- Magical and fantasy theme
- Conflict





# The Scarlet Letter (1850)

Written by Nathaniel Hawthorne

## What is the corpus about?

- Hester Prynne, a woman in Puritan New England who is publicly shamed for having a child out of adultery and forced to wear a scarlet "A" as a symbol of her sin
- Explores themes of guilt, sin, and redemption, particularly focusing on the psychological and moral struggles of Hester and her lover Reverend Dimmesdale

25 chapters and 26 words per sentence (average)

## Linguistic domain

- Narrative
- Characters and dialogues
- Moral and religious theme (sin, redemption, punishment, etc.)
- Archaic language



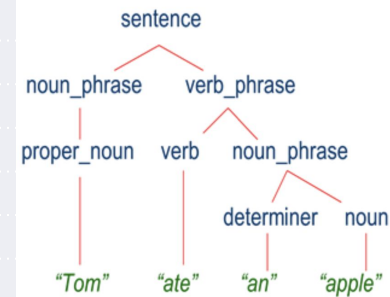


# Main Points from Last Presentation

- Similar themes between two corpus:
  - Character-driven and dialogue-focused
- Corpus-specific themes:
  - Harry Potter: Magic, school, battles, different groups and organizations, etc.
  - The Scarlet Letter: Sin, guilt, repentance, motherhood, public shaming, etc.



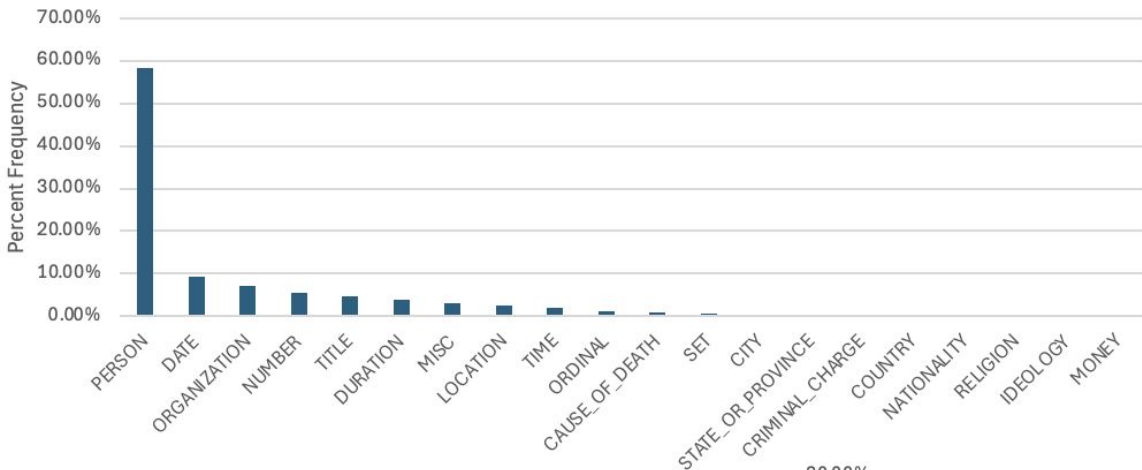
# Parsers



- Analyze and determine the syntactic structure of sentences in a text
- Break down a sentence into its components such as words, phrases, and clauses, and identify the grammatical relationships between those components (Franzosi)
  - Example: Tom ate an apple.
    - “Tom” is the subject
    - “ate” is the verb
    - “an apple” is the object
- The output file from the parser is the CoNLL table.
  - Contains many fields
  - Named Entity Recognition (NER) is an NLP technique used to identify and classify entities in a text into predefined categories such as the names of persons, organizations, locations, dates, times, quantities, etc.

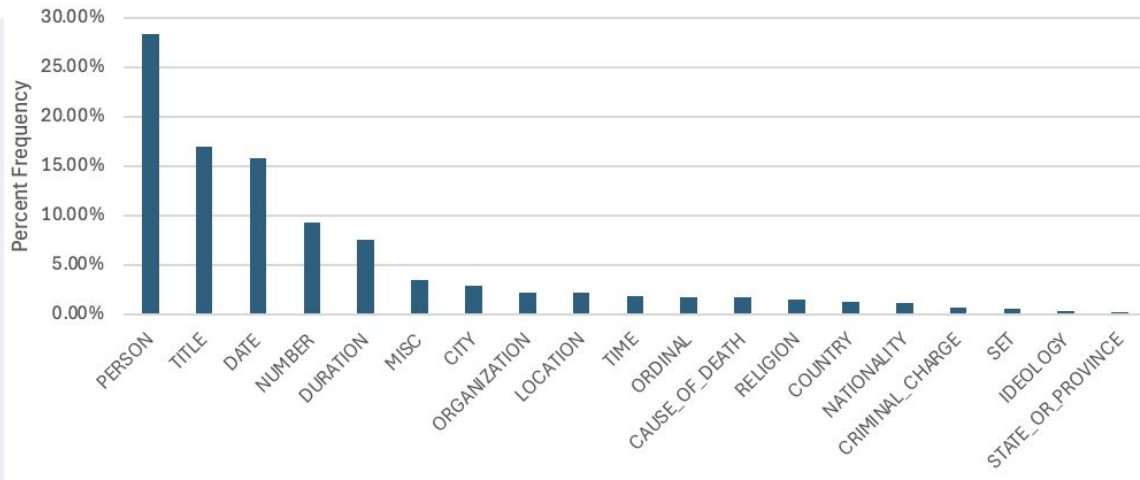


# Named Entity Recognition (NER)



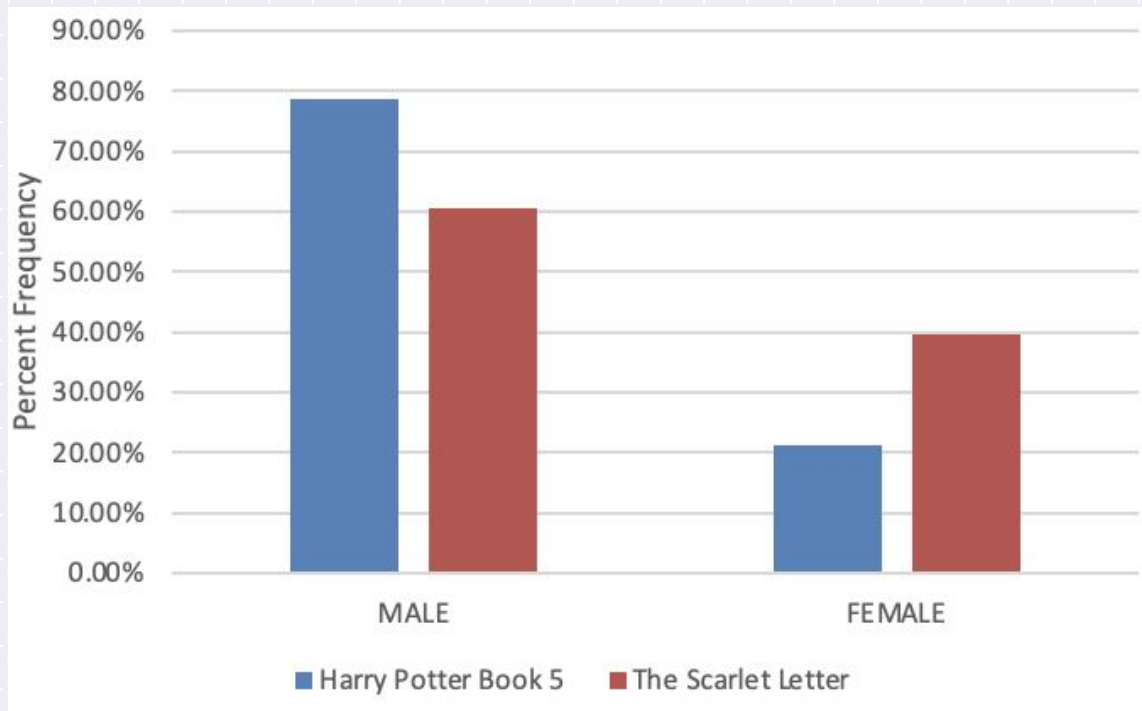
Percent frequency of NER categories (excluding O/undefined values) in the CoNLL Table for Harry Potter Book 5.

Percent frequency of NER categories (excluding O/undefined values) in the CoNLL Table for The Scarlet Letter.





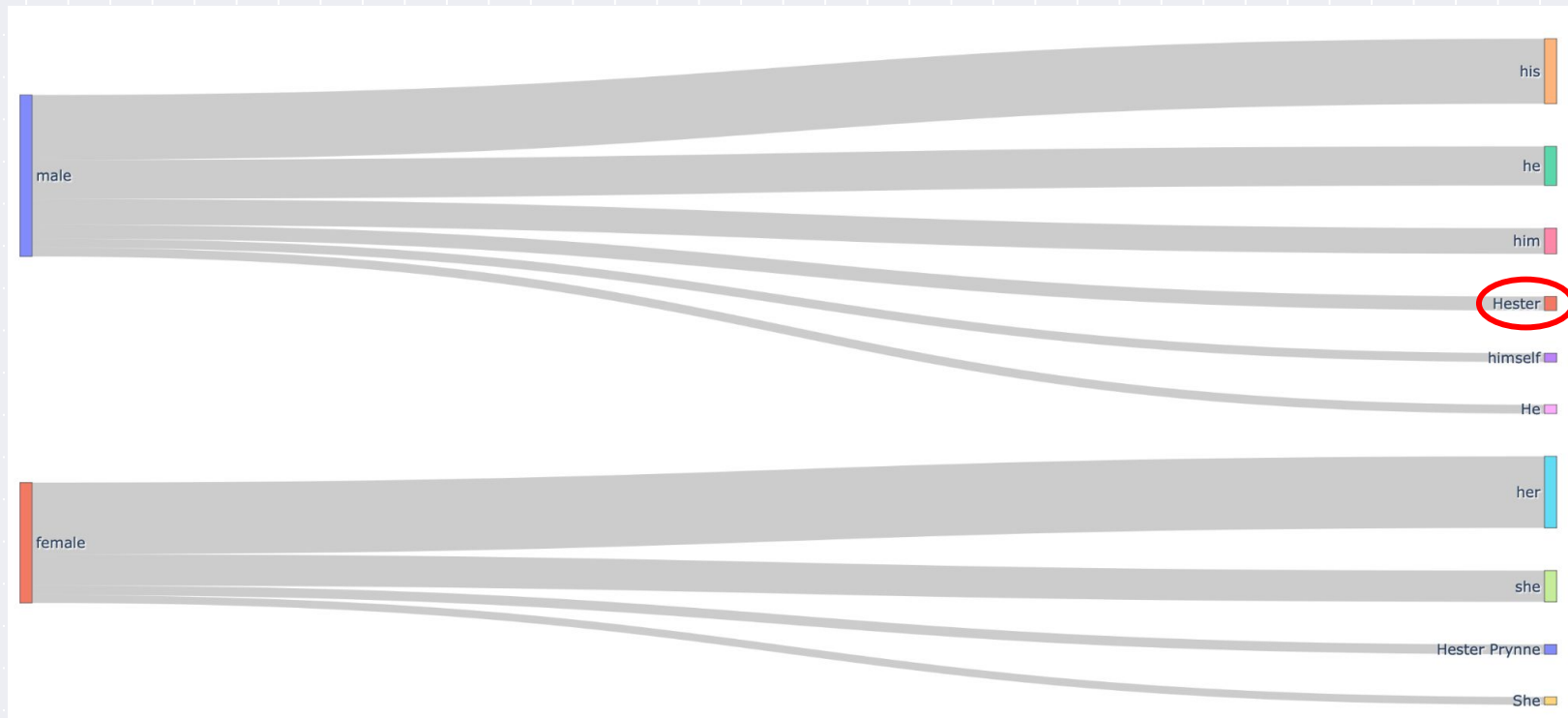
# Gender Annotator



Percent frequency of gender values for identified proper names in Harry Potter Book 5 (blue) and The Scarlet Letter (red) using Stanford CoreNLP's dictionary.



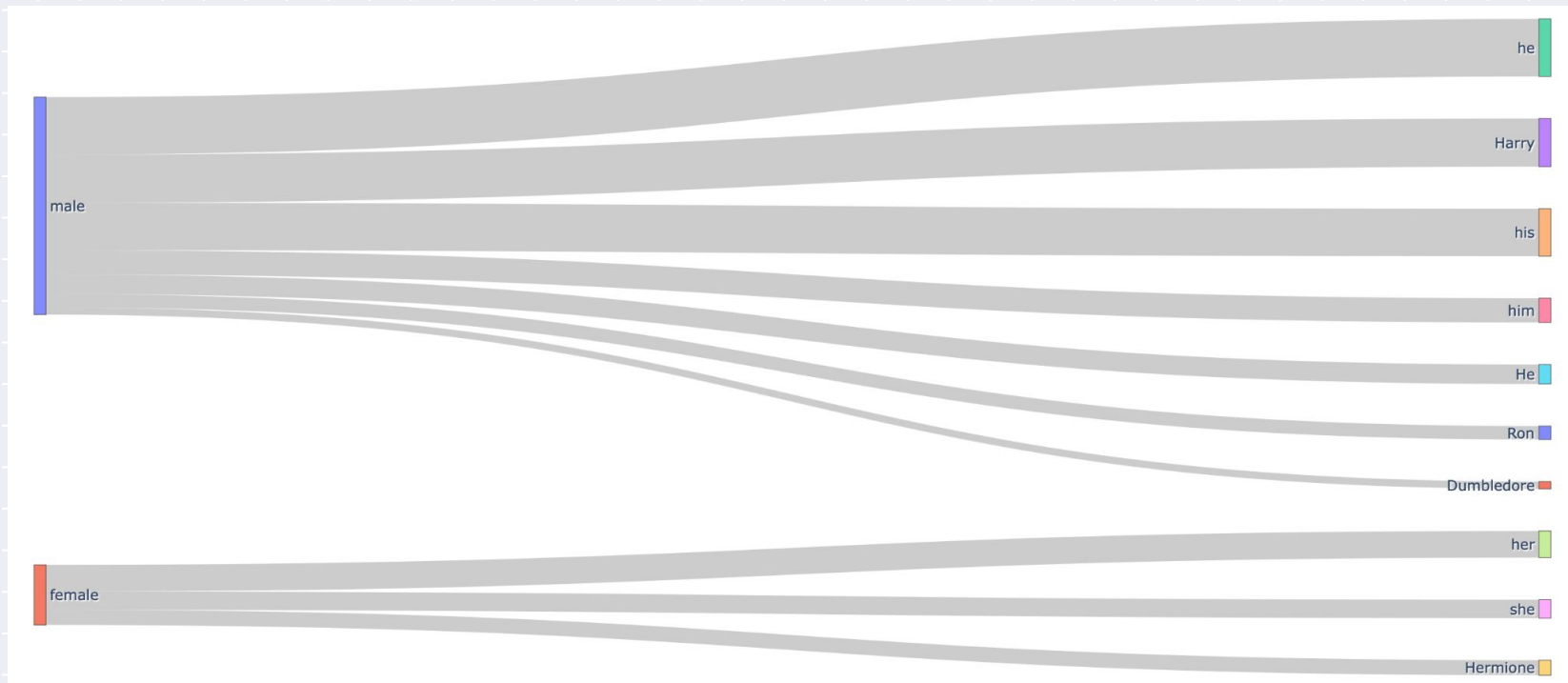
# Gender Annotator – The Scarlet Letter



Sankey chart showing some proper names classified as “MALE” or “FEMALE” in The Scarlet Letter using Stanford CoreNLP’s dictionary.



# Gender Annotator – Harry Potter



Sankey chart showing some proper names classified as “MALE” or “FEMALE” in Harry Potter Book 5 using Stanford CoreNLP’s dictionary.



# CoNLL Table Analyzer

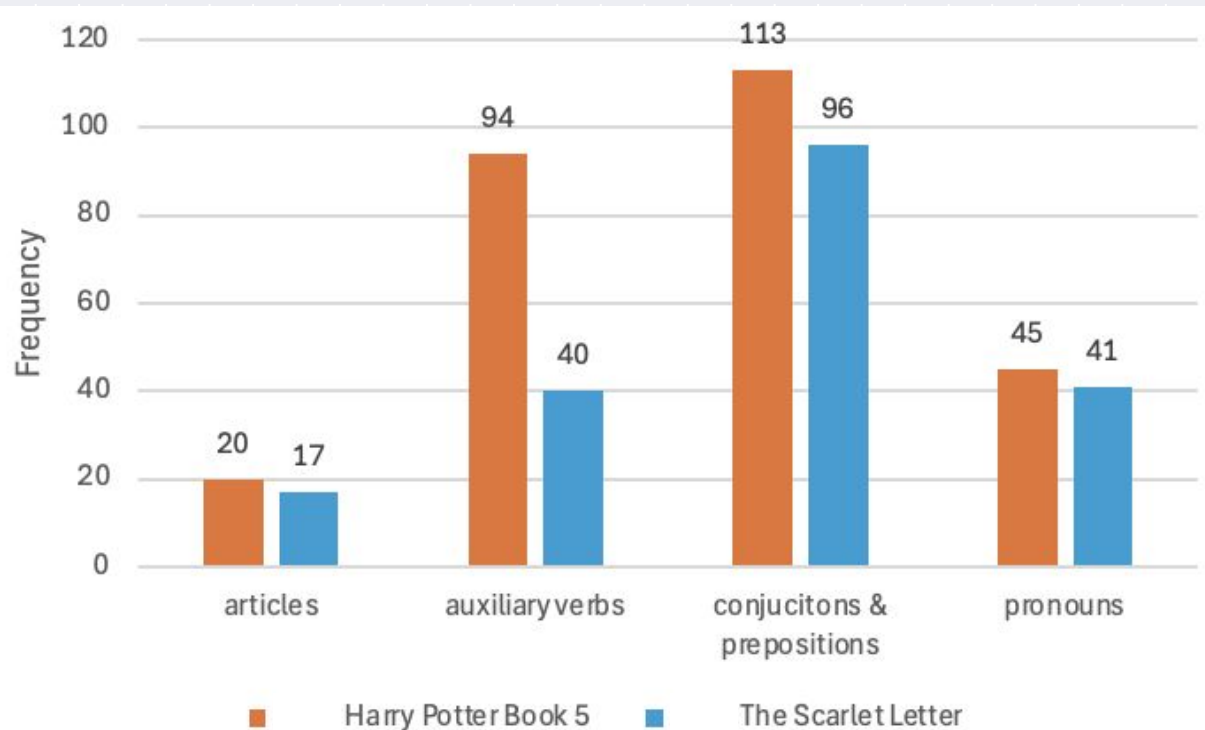
- Perform clause, noun, and verb analyses and provide more insights into our corpora
  - Stopword analysis

Stopwords (or junk words) are words that are filtered out during natural language processing

- Examples: pronouns (e.g., I, you, we, mine), prepositions (e.g., after, in, to, on, with), articles (e.g., a, the), conjunctions (e.g., and, or), and auxiliary verbs (e.g., can, would)
- Though traditionally excluded, stopwords, especially pronouns and auxiliary words, gives insights into the language and style of our texts
  - Example: gender-based language style (Pennebaker)



# Stopword Analysis



## Total stopword

Harry Potter: 272

The Scarlet Letter: 194

Chung and Pennebaker: 300


Frequency distribution of FORM function/stopword categories in Harry Potter Book 5 (orange) and The Scarlet Letter (blue).



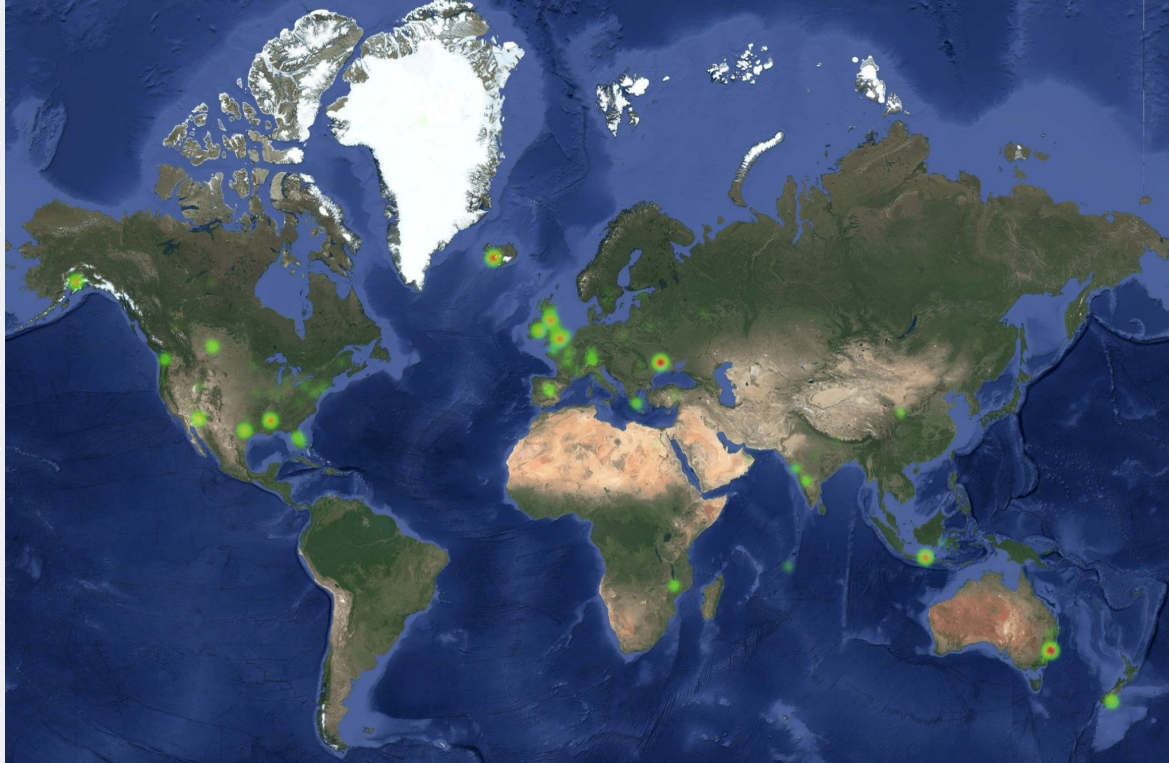
# GIS & Geocoding

- A geographic information system (GIS) puts locations mentioned in a text (“Atlanta” in “Emory is located in Atlanta, GA.”) on a geographic map
- Explore the spatial aspects of literature and the connections between physical locations, characters, and events within a text
- 3 steps in GIS:
  - Extract locations (via Stanford CoreNLP NER)
  - Geocode the locations (via Nominatim or Google)
    - The process of converting addresses into geographic coordinates (like latitude and longitude) and place them on a map
  - Mapping: Google Earth Pro (for pin maps) and Google Maps for heat maps

**Harry Potter Book 5** is primarily set in a fictionalized, magical version of Britain. It is fictional. **The Scarlet Letter** is set in Puritanical 17th-century Massachusetts (specifically Boston). Though it is a work of historical fiction, it draws inspiration from real historical events and punishments of that era.



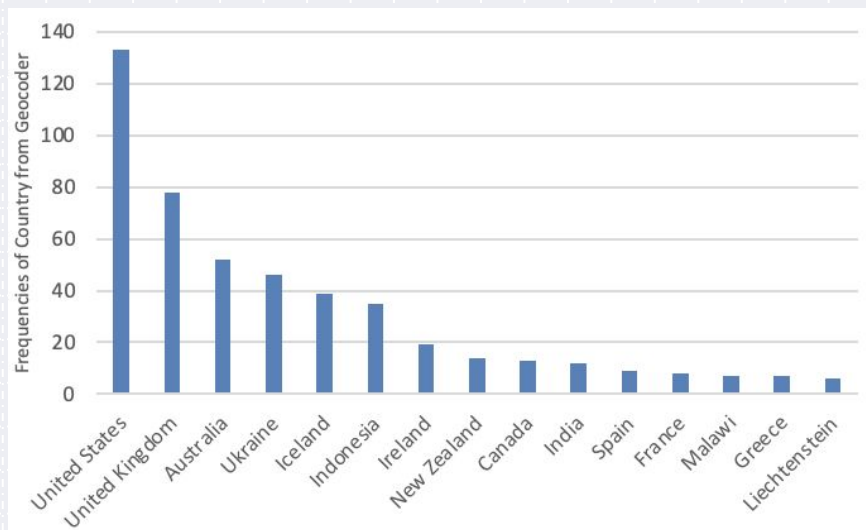
# Geocoding – Harry Potter



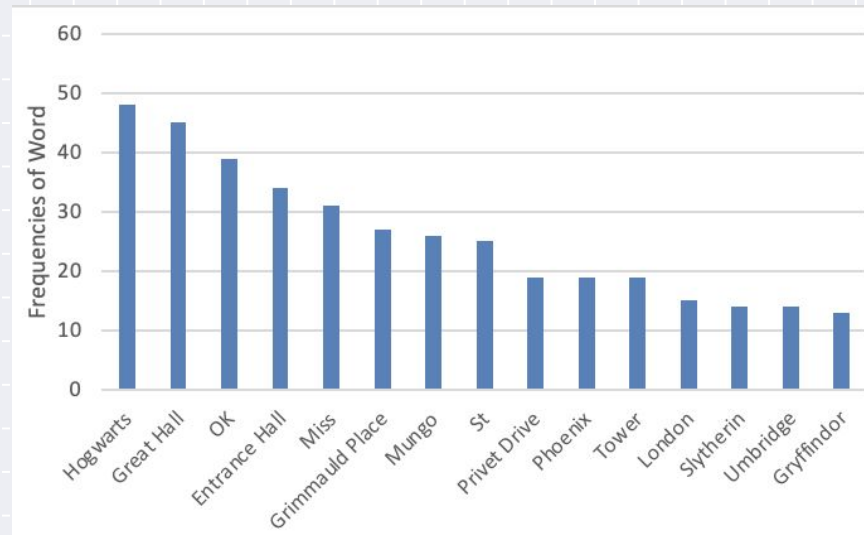
Heatmap (via Google Maps) using Google as the Geocoder.



# Geocoding – Harry Potter



Frequencies of country from the Geocoder (Top 15)



Frequencies of words classified as location by Stanford CoreNLP NER (Top 15)



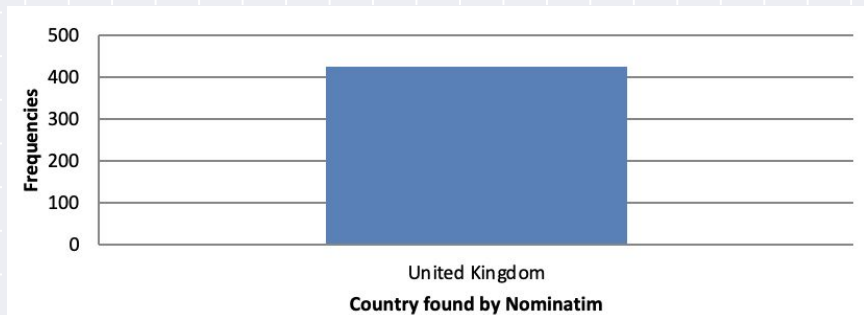
# Geocoding – Harry Potter (improve?)



Heatmap (via Google Maps) using Nominatim as the Geocoder, the UK as the Country Bias, and (58.6350, -8.1482), (49.9590, 1.7620) as the Restricted Area.



# Geocoding – Harry Potter (improve?)



Frequencies of country from the Geocoder



Google Earth screenshot



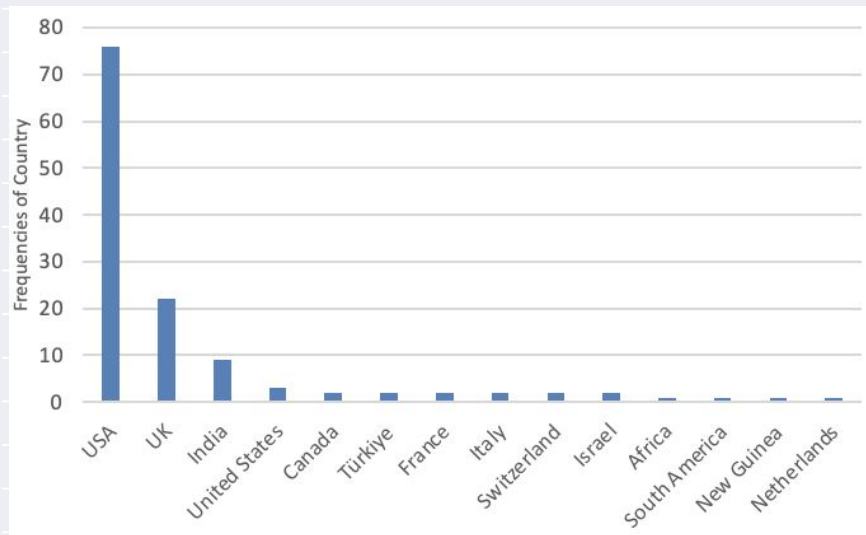
# Geocoding – The Scarlet Letter



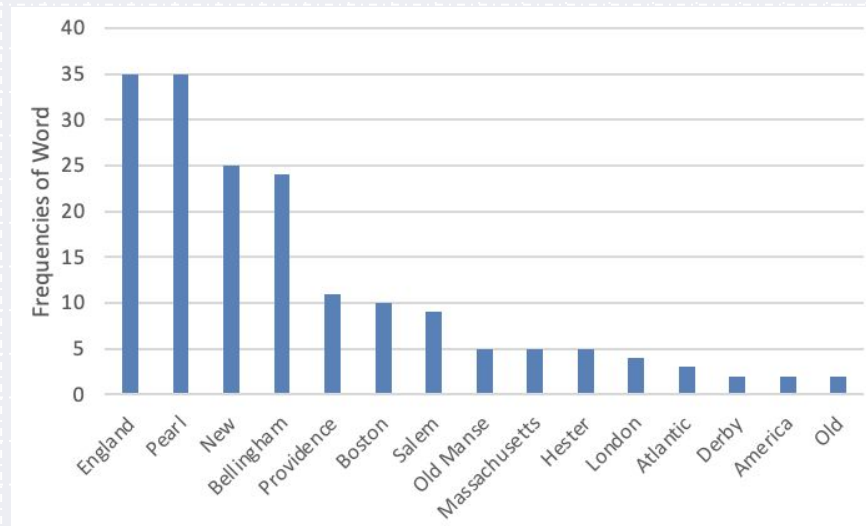
Heatmap (via Google Maps) using Google as the Geocoder.



# Geocoding – The Scarlet Letter



Frequencies of country from the Geocoder (Top 15)



Frequencies of words classified as location by Stanford CoreNLP NER (Top 15)

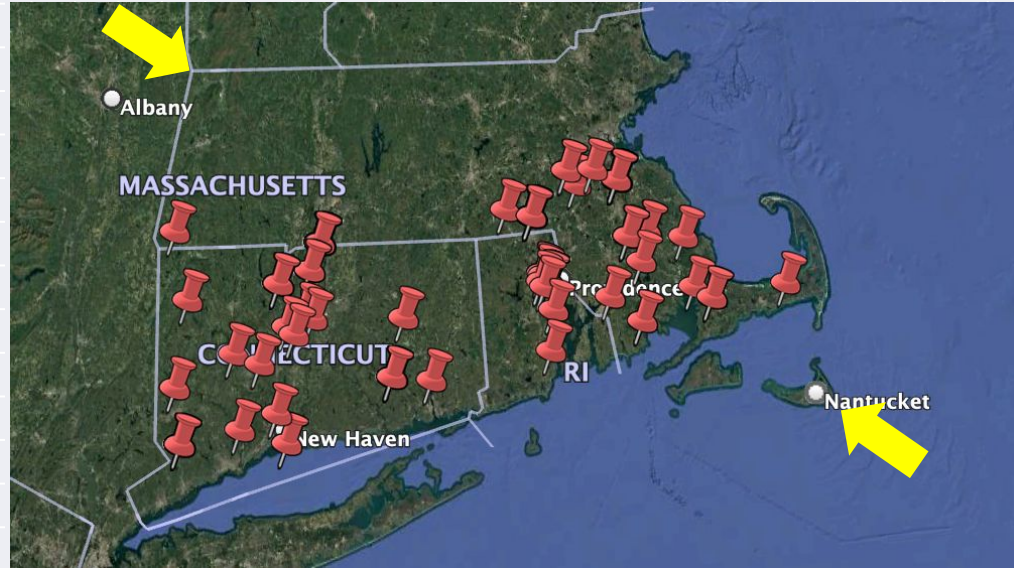


# Geocoding – The Scarlet Letter (improve?)



Heatmap (via Google Maps) using Nominatim as the Geocoder, the US as the Country Bias, and (42.2611, -73.5080), (41.1910, -69.9283) as the Restricted Area.

# Geocoding – The Scarlet Letter (improve?)



Google Earth screenshot





# Works Cited

Franzosi, Roberto. NLP TIPS files.

Chung, Cindy, and James Pennebaker. "The Psychological Functions of Function Words." In K. Fiedler (Ed.), *Social communication*, Psychology Press, 2007 pp. 343-359.

Pennebaker, James W., et al. "Psychological Aspects of Natural Language Use: Our Words, Our Selves." *Annual Review of Psychology*, 2003 vol. 54, pp. 547-77.



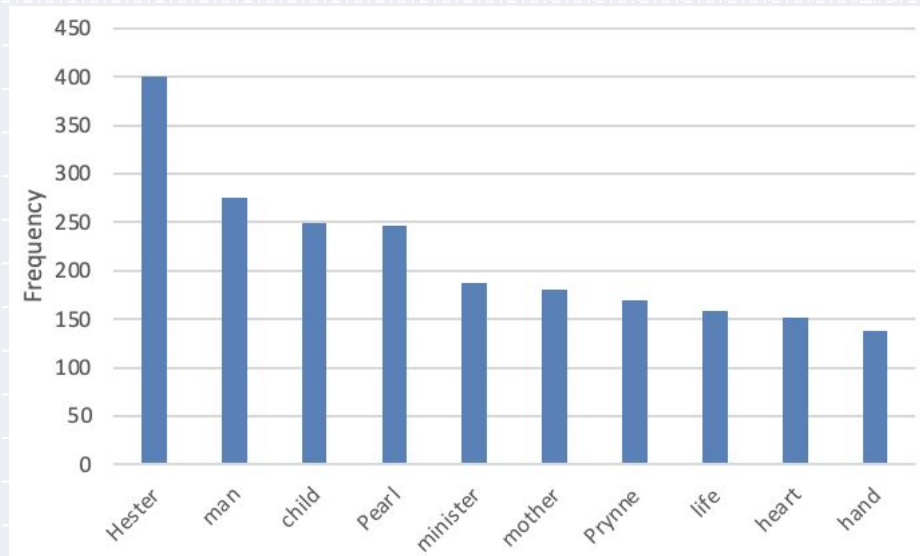
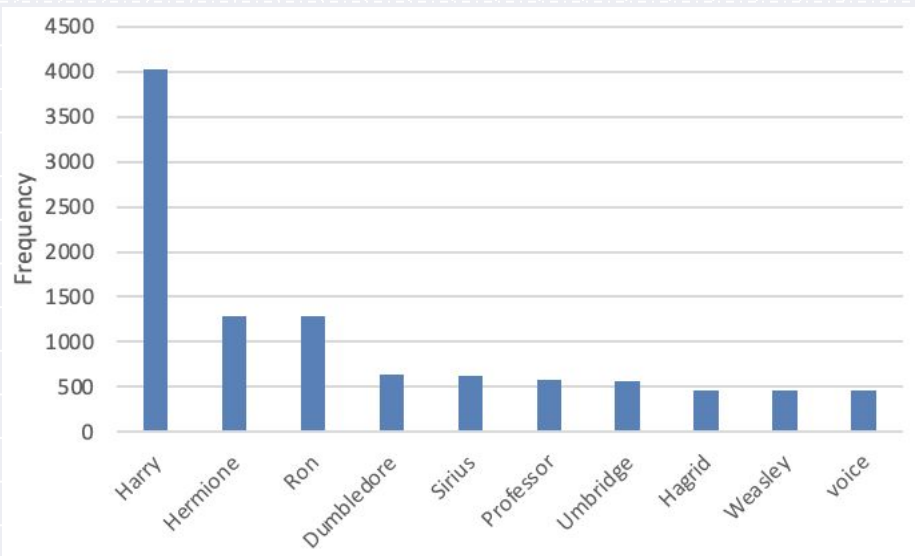
# **Extra Slides**





# Search CoNLL Table Tool

Are there differences in how male and female characters are described?



Frequency distribution of the top 10 most frequent nouns (lemmas) in Harry Potter Book 5 (left) and The Scarlet Letter (right).



■	■	
■		





# Search CoNLL Table Tool – The Scarlet

A word cloud of adjectives co-occurring with the name 'Hester'. The words are arranged in a vertical orientation. The most prominent words are 'helpful' and 'hush' in large, dark blue letters. Other visible words include 'dead' in green, 'alive' in yellow, 'slight' in light green, 'miserable' in teal, 'refuseth' in dark teal, 'own' in light green, 'safe' in yellow-green, 'contrived' in dark teal, 'sainted' in dark teal, 'godly' in dark teal, 'tremulous' in dark teal, 'acute' in dark teal, 'lad' in dark teal, 'embarrassed' in dark teal, 'sensitive' in dark teal, 'wretched' in dark teal, 'conscious' in dark teal, 'startled' in dark teal, 'good' in dark teal, 'pious' in dark teal, 'unhappy' in dark teal, 'unfortunate' in light green, 'venerable' in light green, 'unkind' in light green, and 'sweet' in light green.

A word cloud of adjectives co-occurring with the name 'minister'. The words are arranged in a vertical orientation. The most prominent words are 'poor' and 'old' in large, dark blue letters. Other visible words include 'young' in green, 'same' in teal, 'holy' in dark teal, 'pale' in dark teal, 'wretched' in dark teal, 'conscious' in dark teal, 'startled' in dark teal, 'good' in dark teal, 'pious' in dark teal, 'unhappy' in dark teal, 'unfortunate' in light green, 'venerable' in light green, 'unkind' in light green, 'sweet' in light green, 'acute' in dark teal, 'lad' in dark teal, 'embarrassed' in dark teal, 'sensitive' in dark teal, 'wretched' in dark teal, 'conscious' in dark teal, 'startled' in dark teal, 'good' in dark teal, 'pious' in dark teal, 'unhappy' in dark teal, 'unfortunate' in light green, 'venerable' in light green, 'unkind' in light green, and 'sweet' in light green.

Word clouds of co-occurring adjectives with “Hester” (left) and “minister” (right) in The Scarlet Letter.