Ethical Consumption Before Capitalism (2025)

Project Leads: Dr. Astrid Giugni, Dr. Jessica Hines; **Project Manager:** Nitin Luthra.
Team: Raghd Aljohani, Yewon Chang, Anya Deng, Isabella Xu.

RHODES INFORMATION INITIATIVE AT DUKE UNIVERSITY

Data

Project Summary

This project uses ***The Records of the Virginia Company of London* (1606–1626)** to visualize the changing strategic importance of locations tied to the English Colony in Virginia.

Objectives

- **Automate cleaning of VCR** pages OCR'd by the 2024-25 Bass Connections Team.
- **Analyze** the **VCR** and supplementary texts for English **travel routes, settlements, and territorial disputes** with Indigenous peoples.
- **Create interactive visualizations** of changes of the location and boundaries of colonial territories and their significance to the English.

Methodology

- **Fine-tune BART** and **GPT-4o** on Early Modern English texts (including the VCR).
- **Correct OCR errors** across **2,721 VCR pages** with **GPT-4o**.
- Extract **GPE-** and **LOC-labeled entities** from the VCR with **spaCy's NER model**.
- **Manually validate** and **add** relevant (Virginian and other English colonial) entities.
- **Generate 43 RegEx patterns** matching entities to their historical name variations.
- Re-extract entities from the VCR with **RegEx patterns**; store data by year, entity, and mention frequency in CSV format.
- Develop interactive **RShiny time series map** to visualize distribution and mention density of key locations throughout the VCR.

Team's GitHub

Acknowledgements

We would like to express our gratitude to Dr. Astrid Giugni, Dr. Jessica Hines, Nitin Luthra, Dr. Gregory Herschlag, and Ariel Dawn. Thank you all!

Results

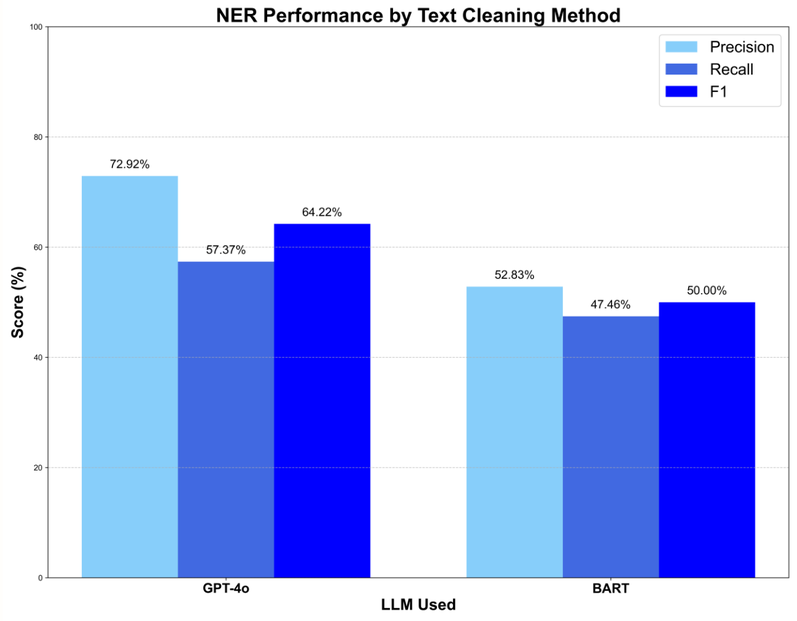

Figure 1: Evaluations of NER given sample cleaned VCR page.

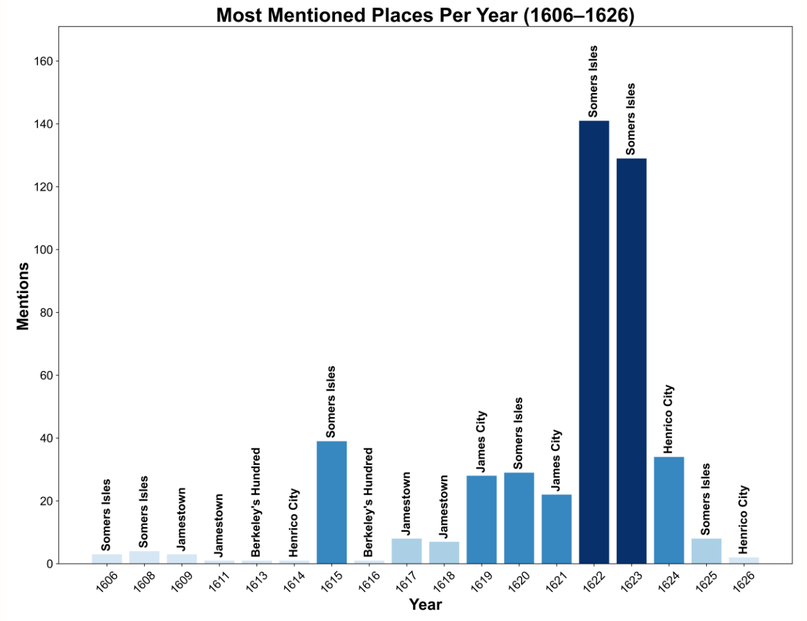

Figure 2: Location with highest frequency of mentions by year.

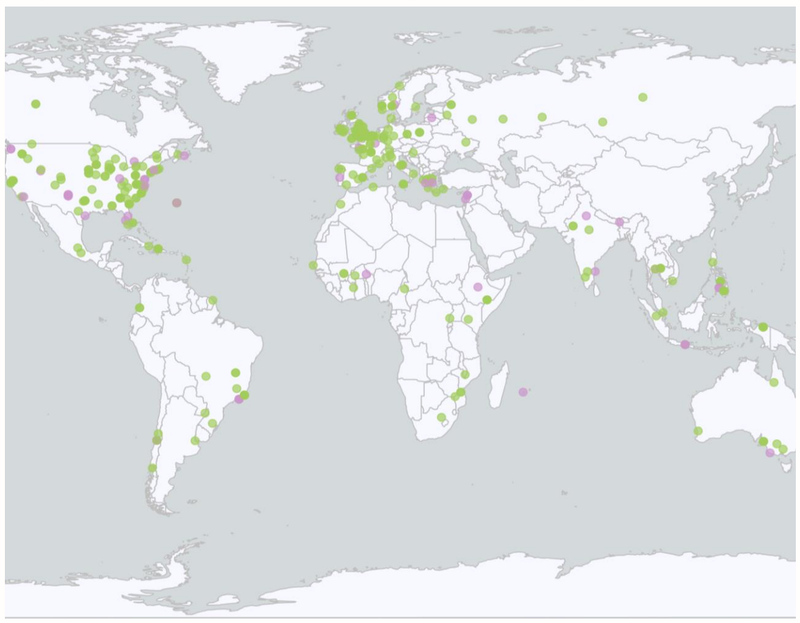

Figure 3: 1622 world map of VCR locations extracted with only NER.

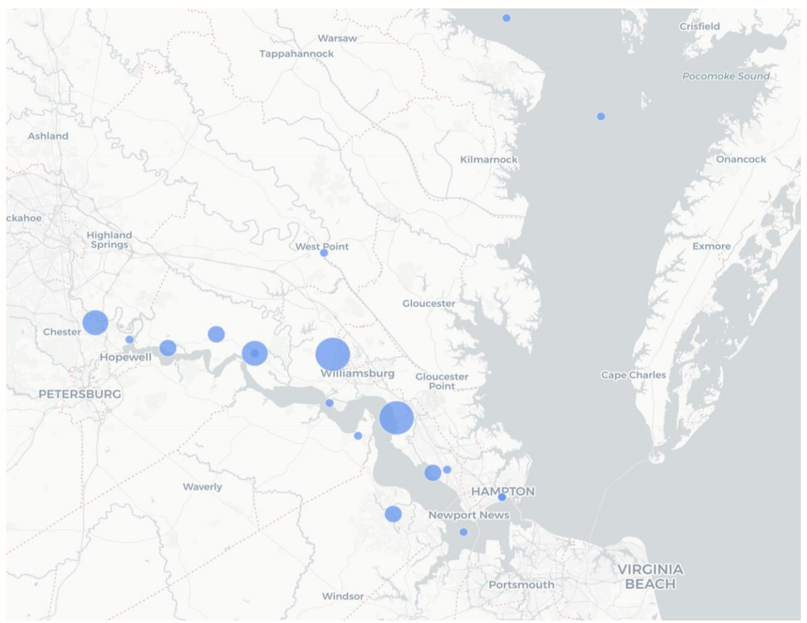

Figure 4: 1622 Virginia-centric map of VCR location mention density extracted with NER and RegEx.

Conclusion

- **Text Cleaning:** The performance scores of both GPT- and BART-cleaned texts reveal that our NER model struggles with the VCR regardless of text cleaning. Notably, the model struggles with identifying LOC entities and the "ambiguity" of Indigenous names (ex. "Powhatan" could be a GPE, PERSON, or NORP). To minimize the need to manually search for locations and their name variations in the VCR, we recommend pre-training an NER model on the VCR.
- **Mapping:** The evolution of key locations as visualized by our map is mostly consistent with the colony's history. For example, settlements most affected by the Massacre of 1622 (ex. Martin's Hundred, Warrascoyack, Henrico) increased the most in mention density post-1622. However, the VCR's introductions, annotations, and titles add noise to our data. For example, Somers Isles (the most common location in titles) is the most mentioned location in 8 out of 18 recorded years and experiences an outsized spike in frequency from 1622-1623. Therefore, we additionally recommend an automated filtering process for non-primary source texts in the VCR.