

Historical Background

The **Virginia Company of London (VC)** was first chartered in 1606 as a way for England to increase its standing in international competition through colonization, resulting in the **exploitation** and **genocide** of the Native Nations of America and the **importation of labor**. These are factors that helped develop **capitalism**. The VC, however, was a massive economic failure and was absorbed by the Crown in 1624. Trying to mitigate these early losses, the VC attempted to push the idea that it was not only economically profitable but also **spiritually profitable**, particularly surrounding specific groups in the discourse of:

- Conversion of Native Nations
- Importing of Indentured (Child) Labor
- English colonization of Ireland
- International competition tied with the Protestant Reformation

The VC was a pivotal force in the foundation of capitalism as the definition of profit and the **language** surrounding different groups shifted in these changing times. Since Western capitalism and American society evolved from colonialism, these definitions are essential when thinking about 'ethical consumption'. From stolen land and deep racial hatred to the desperation of the colony to make all types of profit, this confounding history needs to be understood to promote human autonomy over mindless consumption.

Methodology

XML/VEP Processing

- Around 52,000 texts in Early Print (EP) library.
- Around **8,000 texts** in our period: **1590-1639**.

Standardization/Lemmatization (S/L), Close Reading

- The EP texts were partially S/L-ed, but further S/L was needed. Topical texts were close-read and keywords were added to a dictionary.

TF-IDF, Clustering

- Utilized TF-IDF as a clustering method to create sub-corpora w/ keywords.
- Created categories based on important historical players as well as concepts:
 - Spain, Portugal, Irish, Dutch, West Indies, Jesuits, Native Nations, Indentured Child Labor, African Enslavement
- Clustering produced unexpected results, as categories overlapped too much.

Cosine Similarity (CS) Scores (1)

- Tracked the similarity of two words without taking into account frequency.

N-grams (2)

- Tracked the occurrence of two words in conjunction with one another.

Part of Speech (POS) Tagging (3)

- Since N-grams and CS scores showed only the *proximity* and *context* for each word, POS was used to account for *how* each word was used.

To see our code, scan below for our [Github](#).



To see all of our visualizations, scan above for our [R Shiny app](#).



Figure 5: Virginia Top 10 Adjectives Over Time

*While the N-grams corpora were comprised solely of texts found with TF-IDF, the CS corpora also took into account EP metadata tags, making the CS corpora larger than the N-grams corpora.

Acknowledgements

We would like to thank the following for making this project possible:

- Dr. Astrid Giugni
- Dr. Jessica Hines
- Dr. Paul Bendich
- Dr. Gregory Hershlag
- Erin Winters
- Hannah Jorgensen
- Birmingham-Southern College
- Data+
- Ariel Dawn
- Duke University
- Rhodes Information Initiative (RII)
- Early Print (EP) Online Database
- University of Michigan Text Creation Partnership

Results

These CS heat maps (right) help visualize the relationship between words in any category, while the CS line graph (below) views keyword relations over time. The higher the score, the more two words are related to each other.

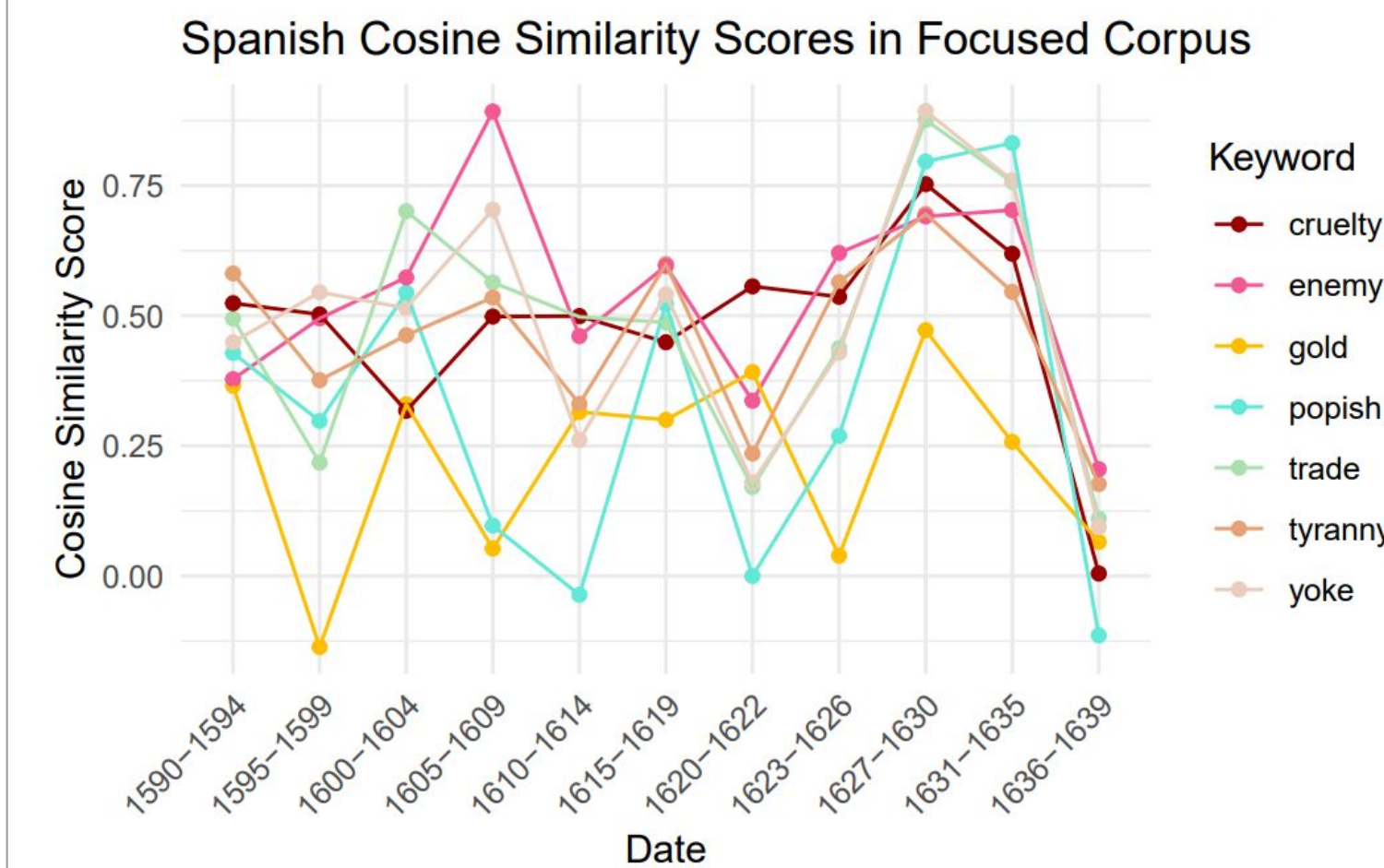


Figure 1: Spanish CS Scores Line Graph in Spanish Corpus*

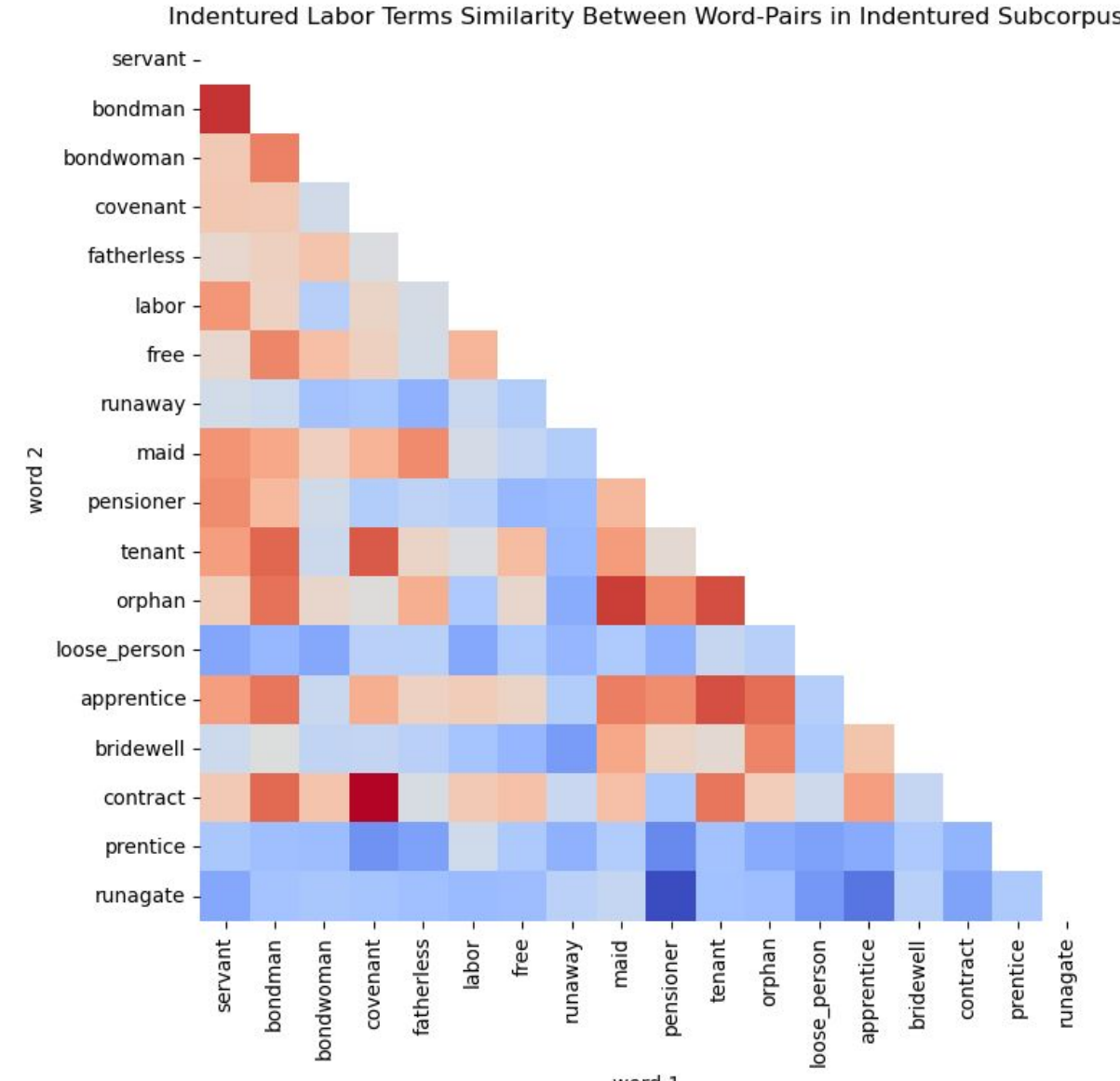


Figure 2: Indentured CS Scores Heatmap in Indentured Corpus*

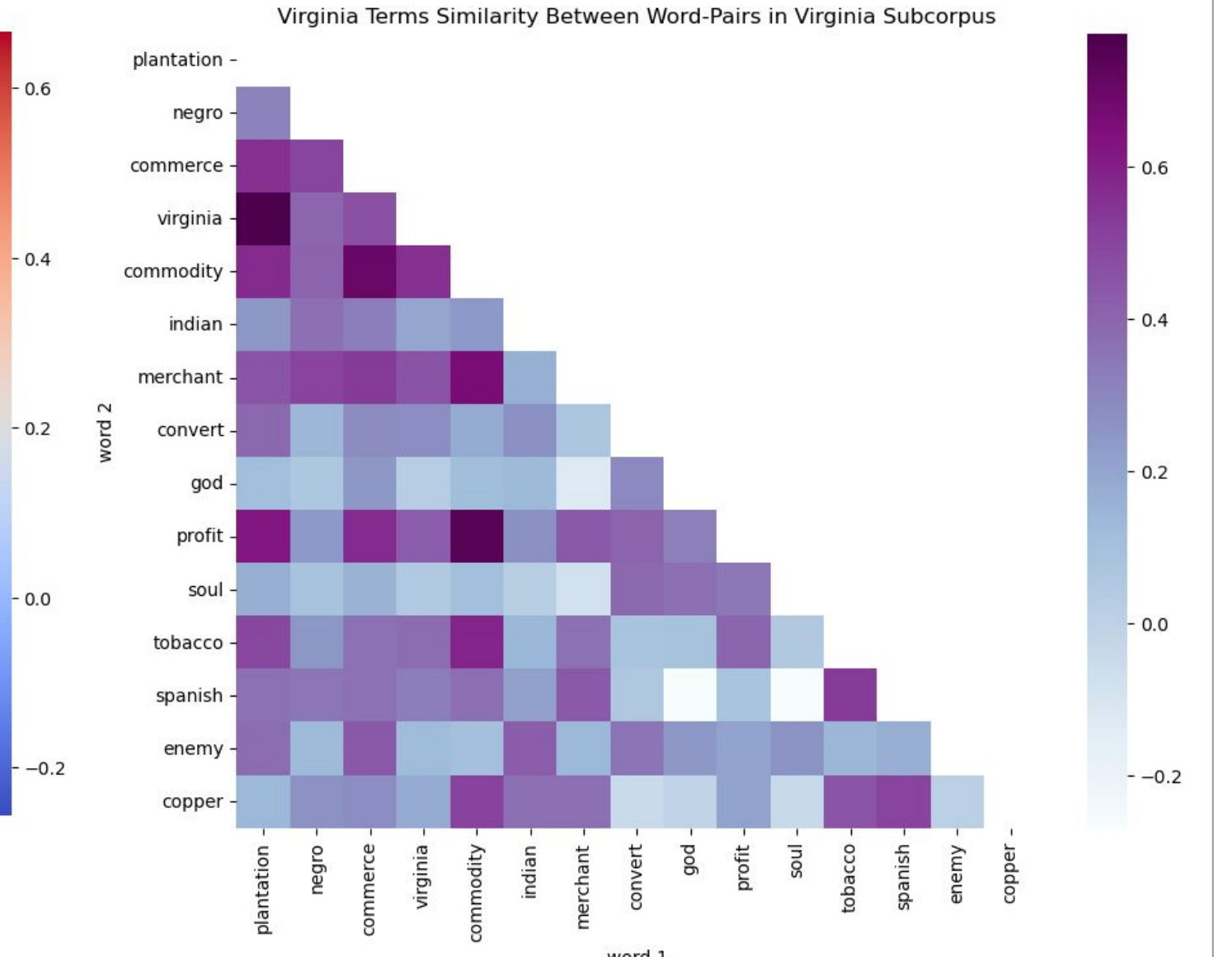


Figure 3: Virginia CS Scores Heatmap in Virginia Corpus*

(1) Cosine Similarity Graphs:

- Calculated using the word embeddings generated by Word2Vec
- Distinct sub-corpora were utilized and different graphs were created to analyze the relationship between keywords over 5- to 10-year periods.

(2) N-gram Graphs:

- N-gram with specific keywords were generated based on significance and general appearance over time
- Limitations: some sub-corpora were very small and did not have a significant amount of texts in every period

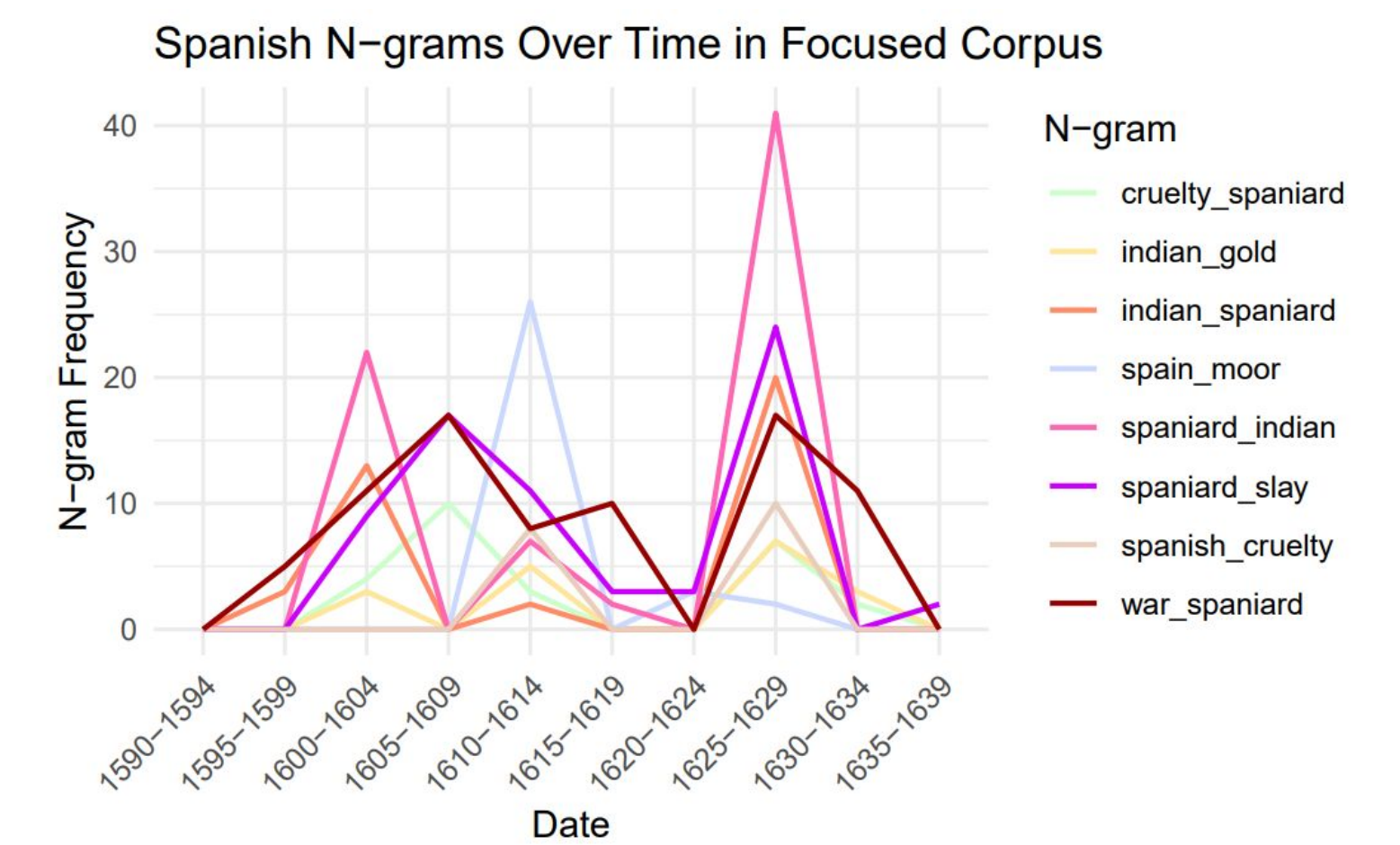


Figure 4: Spanish N-grams Frequency Line Graph in Spanish Corpus

(3) Part of Speech Visualization:

- Top ten verbs, adjectives, and nouns most commonly associated with various keywords were generated over 10-year time periods using original EP XML files that had been cleaned

Conclusion & Future Steps

Based on our examinations and analysis of different sub-corpora and our entire corpus as a whole, we have found the following:

- Key terms that developed surrounding Native Nations in Virginia, especially the Powhatans
 - Emergence of terms such as 'savage' and 'heathen'
- Spain was seen as an enemy in the colonization of the Americas
 - Negative sentiments towards Spain were not fully censored during King James' reign.
- Development of key words and their usage
 - 'Wild' becomes an adjective associated with the term 'Irish' only after the 1600s.
- Intersections between sub-corpora: Similar language was used to describe different groups.
 - 'Native' and 'Irish' tend to 'barbarian' and 'naked' as 'Irish' and 'Indentured' tend to 'beggarly'.

To learn more about what we found as well as our background and references, scan below for our [website](#).



In the future, we hope to do more close analysis of texts in our whole corpus. Identified texts of interest include Walter Raleigh's "The Discovery of Guiana" and "The History of the World" to better analyze Anglo-Spanish competition with American colonization.