Ricky Reyes
Dr. Clovis Gladstone
Digital Texts
14 March 2024

Digital Texts Final Write Up (Part II)

Introduction

In this analysis, I explore three unique text analysis methods to extract insights using distant reading from my corpus: document-term matrix exploration (TF-IDF and co-occurrences), sentiment analysis, and topic modeling (LDA). The goal was to identify key terms, evaluate tone and sentiment, and uncover the underlying currents especially as it related to how insiders (those from Guam and CHamoru people) talk about the island and culture compared to insights from outsiders (those who are not CHamoru or from Guam). To prepare for analysis, I utilized various preprocessing steps including stop word removal, lemmatization, part-of-speech filtering, and n-gram extraction to produce a series of high-quality outputs while analyzing my corpus. This write-up details the methodologies, preprocessing details, and finding reflections from each approach along with an outlook for future additions and analysis to this corpus.

Method 1: TF-IDF & Co-occurrences

Selection: I used TF-IDF to analyze meaning across my corpus at the word level. TF-IDF is unique in that instead of using raw word counts, TF-IDF relies on key filtering and distribution to calculate a series of impactful words within a given corpus. In keeping with word-level analysis, I also incorporated co-occurrence analysis and wrote a script to test specific words for their co-occurrences across texts. The main questions I was answering with this technique included:

- What are the most significant terms within the corpus, and how does their importance vary across documents?
- Word associations, meaning, and patterns underlying a given corpus

Preprocessing Details:

1. Stopword Removal: Removal of common words with the purpose of clearing up issues of noise within the corpus analysis process.
2. Lemmatization: Reduced words to their base forms to standardize variations.
3. Part-of-Speech Filtering: Specifically focused on nouns and adjectives, focusing on words most likely to carry thematic weight.

These steps helped to reduce noise and, during the process, were adjusted as useless results were produced by early code.

Results Analysis:

After completing the TF-IDF analysis, interesting observations emerged that served as a starting point for future closer reading. Examples of interesting findings include *A Marianas Mosaic* which chose to center concepts like "*generation*, *culture*, and *people"* pointing toward a focus on intergenerational identity and cultural preservation (according to the analysis results). In contrast, texts like *Guam: Two Invasions and Three Military Occupations* centered terms like "*Japanese*, *naval*, and *gun"*, focusing the text within a more militarized global framework. These patterns give direction to future investigative questions about whether insider-authored texts emphasize cultural knowledge, community, and people, compared to outsider-authored texts, which may tend to view Guam and CHamoru people through lenses of conflict, occupation, or colonial encounters.

This method also reveals where key-terms overlap - especially considering importance within a given text- including terms like "*island"*, which appeared across both insider and outsider texts—even though its contexts and co-occurrences differed.

The TF-IDF results provide a useful lens for comparing what kinds of knowledge, history, and identity are privileged in each text, and they begin to show how linguistic emphasis may reflect deeper undercurrents or positional differences in how Guam and CHamorus are represented in literature and other forms of writing.

Method 2: Sentiment Analysis

Sentiment analysis was chosen to evaluate the emotional tone and framework of the texts. This method was particularly useful in:

- Comparing sentiment across works written by insiders and comparing results to sentiment across works written by outsiders.
- Looking at the overall sentiment distribution across the corpus.
- Identifying noticeable sentiment trends based on different sections of the texts (insider/outsider, non-fiction/fiction, cross-genre analysis, document type, etc.)

I utilized VADER (lexicon-based analysis) to classify sentiment as positive, negative, or neutral.

Preprocessing Details:

1. Lowercasing: Text standardization for consistency in the analysis process.
2. Stopword Removal: Removal of common words with the purpose of clearing up issues of noise within the corpus analysis process.
3. Punctuation Removal: Cleans the inputs before analyzing sentiments.

Results Analysis:

The sentiment analysis revealed relatively low and neutral overall polarity scores across the corpus, but there are nuanced distinctions that are important to be observed between insider and outsider-authored texts. Insider-authored texts like *Legacy of a Political Union* and *CHamoru Legends* scored the highest in sentiment, reflecting more affirming and/or hopeful language tied to themes like community, storytelling, and sovereignty (see TF-IDF findings above). On the other hand, outsider-authored texts such as *Guam: Two Invasions and Three Military Occupations* and *Destiny's Landfall* had the lowest sentiment scores, likely due to the nature of their authoring as historical documents and non-fiction texts and their focus on war, colonization, and trauma—narratives that view the history of CHamoru people through a more neutral, historical, or negative lens.

While sentiment analysis cannot capture nuance or cultural context on its own—especially when a large portion of CHamoru-language text is missing due to translation limitations—it does support the impression that insider-authored texts may employ a more positive or resilient narrative tone. This subtle difference helps better understand how these texts reflect different emotional and political relationships to place and people—including the very authors of the texts within the analyzed corpus.

Method 3: Topic Modeling (LDA)

I selected LDA for its ability to create topic models with the goal of uncovering themes and topics across my corpus. This method helped answer key questions including:

- What underlying topics are present in the corpus dataset?
- How do topics vary across different texts?
- What words are associated with be closer/further apart based on distant reading analysis?

Preprocessing Details: LDA required detailed preprocessing to ensure meaningful topic extraction. Steps to preprocess included:

1. Stopword Removal & Lemmatization: To improve model coherence.
2. Bigram & Trigram Extraction: To identify multi-word expressions that convey richer meaning.
3. TF-IDF Filtering: Removing words that are common and removing words that are rare in an effort to reduce distortion.

Results Analysis:

Topic modeling offers another form of analysis, focusing on various clusters of themes that correspond to different high-level ideas within the corpus. One topic, dominated by terms such as "*guamanians,*" "*naval,*" "*invasion,*" and "*army,*" illustrates themes of war and military histories, potentially presented in outsider-authored texts (based on the previous analysis we've completed). Another topic, which clustered terms like "*ancient,*" "*padre,*" and "*missionaries,*" connects more closely to colonial narratives, aligning with outsider frameworks of Guam's past. In contrast, topics containing terms like "*CHamoru*," "*generation*," and "*art*" appear to represent contemporary cultural expressions and identity-making—more commonly emphasized in insider-authored texts, I would hypothesize.

This method is useful in that it depicts how insider and outsider texts cluster around different conceptual concerns. While some overlap exists, the topic distributions reinforce the idea that positionality informs not only what is written, but also how collective memory, cultural knowledge, and people are written about. Topic modeling, therefore, is a useful tool for identifying not just what stories are told, but which patterns of storytelling dominate in each perspective within our corpus.

Reflection

This project demonstrated how different text analysis methods uncover unique aspects within a specified corpus. In reflecting on my process for corpus analysis, each method provided a distinct, yet complementary set of insights. Using these Python libraries, I was able to uncover meaningful patterns and gained insights into the interesting undercurrents of my corpus and its authors pointing to areas for potential future close reading. Future work for this project could include looking at hybrid models for richer and more context-aware text analysis, integration of CHamoru language tools to analyze the large portion of CHamoru language texts, and improvement of the preprocessing process to better train my language model to enhance the accuracy and depth of corpus analysis.