Thesis Defense

Computer Science Master's Program

"Extracting California's Land Grant History: A Comparative Study of Rule-Based and LLM Methods for Historical Document Processing"

By **Anthony Colin Herrera**

Abstract:

This thesis explores methods to extract structured data from historical documents, making them more accessible for analysis and visualization. The goal is to provide historians and Digital Humanities scholars with tools that save time and reduce the manual effort required for this kind of research. We work with census records and historical texts from the Early California Population Project and Spanish Colonial census, covering California's 18th-19th century transition from Spanish and Mexican control to U.S. statehood. The system has three components: person matching across records, family tree generation, and land grant extraction. For person matching, we use a modified Levenshtein distance algorithm to link individuals across census and mission records, accounting for Spanish colonial name variations. Matched records feed into family tree generation, connecting ancestors to descendants across generations. The main focus of this thesis is land grant extraction from California Ranchos, a book documenting private land grants across California counties. We develop a rule-based NLP pipeline using spaCy to extract grant names, transaction histories, and geographic coordinates from semi-structured historical text. We then compare this approach against Large Language Models (GPT-4o and Grok-3) to evaluate whether modern LLMs can outperform traditional NLP methods on domain-specific historical documents. To ensure fair comparison, we built a human-annotated golden set of 100 land grants and developed a unified scoring algorithm that weights fields by importance: identification fields (3 points), transaction details (2 points), and supplementary information (1 point). Results show that LLM extraction (Grok-3: 84.7\%, GPT-4o: 83.7\%) outperforms baseline rule-based extraction (80.3\%) by 3-4 percentage points. However, targeted post-processing enhancements addressing OCR artifacts, coordinate formatting, and name normalization improved the rule-based approach to 83.0, narrowing the gap to just 1.7 percentage points. The largest remaining differences appear in supplementary fields requiring contextual interpretation. Our hope is that this work demonstrates the viability of automated extraction for historical records and provides useful methodology for the Digital Humanities community.

Date: Thursday, May 28th
Time: 4:00 PM – 6:00 PM
Location: 14-238b
Committee: Dr. Foaad Khosmood, Dr. Cameron Jones, Professor Kirk Duran