Shannon Rumsey

San Jose, CA - shannonalmyrumsey@gmail.com - shannonrumsey.github.io

EDUCATION:

University of California, Santa Cruz September 2024-Present
M.S. Natural Language Processing

University of California, Santa Barbara September 2021-December 2023
B.S. Statistics and Data Science
Recipient of National Science Foundation's Harnessing the Data Revolution Data Science Corps Award

Santa Barbara City College August 2019-May 2021
A.A. Liberal Arts-Science: Science & Math

EXPERIENCE:

Postbaccalaureate Research Assistant, CPLS Lab, Remote June 2023-April 2024

- Validated and expanded upon 15 prior linguistics studies on the Functional Load Hypothesis, identifying potential predictors using a mixed-effects logistic regression model
- Quantified the likelihood of minimal pair confusion in context through Word2Vec embeddings trained on curated, preprocessed corpora
- Examined the impact of varying window sizes and part-of-speech tags to explore the relative importance of syntactic versus semantic context in determining confusability

Capstone Researcher, Evidation Health, Santa Barbara, CA January 2023-June 2023

- Directed a team of 5 to analyze wearables data, including resting heart rate and caloric expenditure, for predicting outcomes of Respiratory Viral Infection lab tests
- Explored data imputation techniques like MICE and applied dimensionality reduction methods, including UMAP and t-SNE
- Presented findings at the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining

Undergraduate Data Science Fellow, Central Coast Data Science Fellowship, Santa Barbara, CA September 2022-June 2023

- Led outreach initiatives to inspire students and promote departmental courses to a diverse student body
- Organized and presented at the Department of Probability and Statistics Project Showcase

RESEARCH PROJECTS:

Penn-Treebank Model Generation November 2024-December 2024

- Built a decoder-only Transformer model to generate sentences similar to the Penn Treebank dataset
- Implemented dynamic embeddings and hyperparameter tuning to improve performance
- Achieved a perplexity of 269, compared to GPT-2's perplexity of 65

Entity Aware Machine Translation September 2024-December 2024

- Designed a Seq2Seq and Transformer model for translating English sentences with named entities into 3 target languages
- Leveraged a knowledge graph for named entity translation and integrated it into end-to-end machine learning pipelines
- The Transformer model outperforms mBart with a COMET score of 0.68
- Findings will be submitted to the 2025 SemEval workshop

Medical Transcription Simplification Using BERT April 2023-June 2023

- Enhanced the readability of medical transcriptions by lowering the Flesch-Kincaid metric, improving accessibility to those without domain expertise
- Identified synonym substitutions through cosine similarity analysis between BioBert embeddings of Harvard Health medical definitions and context-specific terminology found in the transcriptions
- Leveraged Bert2Bert to summarize and further refine the documents

LANGUAGES, TOOLS, & SKILLS:

Python, R, SQL, PyTorch, Git, Machine Learning, NLP, Computer Science, Statistics