Name: Michael Tamburello
NetID: mjt11
Date: November 19, 2023

Project Progress Report
Visualized Media Bias and Polarization Detection

Team Members

Name (NetID) - Role
Michael Tamburello (mjt11) - Captain

Feature Progress Breakdown

- Incorporate spaCy's Named Entity Recognition to pinpoint key subjects and actors in news articles. **[COMPLETE]**
- Leverage spaCy's Part-of-Speech tagging and Dependency Parsing for intricate sentence analysis, illuminating potential bias. **[COMPLETE]**
- Utilize spaCy's advanced text classification features to train a machine learning model that quantifies media bias. **[COMPLETE]**
- Employ spaCy's similarity metrics to juxtapose articles from disparate outlets, thereby revealing journalistic dissonance. **[TODO]**
- Design an embedding projection visualization to spatially represent word biases in a multi-dimensional vector space which highlights stark differences found between sources on identical events. **[TODO]**

Workload Progress Breakdown

- Data Procurement and Cleansing: 5 hours **[COMPLETE]**
- Text Preprocessing Pipeline with spaCy: 5 hours **[COMPLETE]**
- Model Building with PyTorch: 5 hours **[TODO]**
- Embedding Visualization Design and Development: 5 hours **[TODO]**
- Frontend and User Experience Design: 5 hours **[TODO]**
- Documentation and Evaluation: 5 hours **[TODO]**

Overall Time Spent: 12 Hours / 30 Hour Estimated Commitment

Challenges

The data I am using has varied in format and quality. Also, there are unique parts of speech inherent to news material such as quotations, external references, jargon, etc. that must be handled. I had to implement a series of data normalization and cleansing scripts that could adapt to various data formats and inconsistencies. This caused the data cleansing and preprocessing pipeline steps in my workflow to take longer than I had anticipated.