Dimensionality Reduction and Optimization of the GloVe Words Database Using Principal Component Analysis and Birch Clustering

Kacheria, Nishka (School: Interlake High School)

Natural Language Processing systems such as GloVe are necessary for computer interpretations of human language. Previous research has used these systems to predict genomes or stocks. After training NLPs on 300 dimensions, no further compression of the dataset is completed in a variety of languages (English, Mandarin, etc.). This provided motivation for my unique and novel investigation on optimizing the dataset, thus allowing for lower runtime and less memory usage for NLPs. The ramifications of this research are faster and more efficient processing in auto-generating sentences or following words such as in ChatGPT or phone predictive text. Simultaneously, my research creates a system for continuous AI internal refinement. I utilized novel creative methods for determining the relative definition of a word. Further, I comparatively analyze clustering methods as well as dimensionality reduction techniques to reduce the dataset dimensionality while preventing data loss. This yielded merely 143 instead of 300 dimensions necessary with the same amount of data, thus halving memory requirements, and revealed significant runtime improvements of up to 8 minutes for O(n^3) complexities. My work also revealed word associations that reflected racial and gender bias, and outlines how to find these biases in the future. Finally, I have created a framework for determining the necessary dimensionality sizes of specialized corpuses in the future that have less words and dimensions being used for extraneous information. Keywords: Optimization, NLP, Database, PCA, Birch, Dimensionality

Awards Won: