**Analysis of Vector Representations and Similarity Measures in Shakespeare's Works**

**Introduction**

This study explores various vector representations and similarity measures to uncover patterns and relationships within the text of Shakespeare's plays. By examining the textual data through different computational lenses, we aim to reveal insights into character similarities, genre clustering, and potential distinctions between the language use of male and female characters.

**Methodology**

**Vector Representations**

* We utilized four primary vector representations to encode the textual data:

**Term-Document Matrix (TDM):** Captures the frequency of terms across documents, providing a foundational representation of text data.

**TF-IDF Matrix**: Applies term importance weighting to the TDM, highlighting unique terms in each document.

**Term-Context Matrix (TCM):** Focuses on the context of term occurrences, aiming to capture semantic relationships based on co-occurrences.

**Positive Pointwise Mutual Information (PPMI) Matrix**: Enhances the TCM by emphasizing significant co-occurrences, potentially revealing deeper semantic connections.

**Similarity Measures**

* To compare vector representations, we employed three similarity measures:

**Cosine Similarity**: Measures the orientation similarity between vectors, ideal for high-dimensional data.

**Jaccard Similarity**: Evaluates shared members between sets, adapted for weighted vectors by considering non-zero elements.

**Dice Similarity**: Similar to Jaccard but gives more weight to intersection, suitable for analyzing highly similar data.

**Character Similarity**

Analysis of character speech revealed distinct patterns of similarity within and across plays. Certain vector representation and similarity measure combinations were particularly effective at grouping characters by thematic roles or affiliations within plays. For instance, cosine similarity applied to TF-IDF vectors successfully grouped antagonists and protagonists across different works, suggesting that thematic elements of character roles are captured through their spoken words.

**Similarity and dissimilarity**

characters are compared using vector representations like TF-IDF and similarity measures such as cosine similarity. This approach can reveal both similar and dissimilar (or opposite) characters to a given reference character, such as Juliet. Based on the script and similarity analysis of the vector the following was the output.

A screen shot of a computer

Description automatically generated

**Play Clustering**

When clustering plays into comedies, histories, and tragedies, the TF-IDF representation combined with cosine similarity measures proved effective in distinguishing between genres. This combination highlighted each genre's unique terms and stylistic elements, aligning well with traditional classifications. Cluster analysis, particularly hierarchical clustering, visually demonstrated the genres' separation, with comedies forming one cluster, tragedies another, and histories a distinct third.

A screenshot of a computer program

Description automatically generated

**Gender Differences**

Exploration of gender-specific language through aggregated vectors for male and female characters uncovered notable language use and theme differences. PPMI matrices combined with cosine similarity highlighted these distinctions most clearly, suggesting that semantic relationships encoded in the PPMI matrix are key to capturing nuanced differences in gendered language use. For example, female characters' speech was enriched with terms related to family and relationships, while male characters' language was more aligned with power and conflict.

**Conclusion**

This study demonstrates the power of computational text analysis to uncover patterns in literary works. We gained insights into character similarities, genre distinctions, and gender-specific language use by applying different vector representations and similarity measures to Shakespeare's plays. The findings suggest that certain combinations of representations and measures are more suited to specific types of analysis, offering a valuable toolkit for future literary studies and beyond.