Project 1 Post-port

Hanzhang Yin

October 19, 2024

Summary of Results

This project aimed to distinguish the tonal styles of Zhu Shuzhen and Du Fu using Markov matrices and various computational methods. The test sequence was analyzed using the log-likelihood method, equilibrium vector comparison via cosine similarity and Euclidean distance, as well as a baseline approach using the infinity norm.

Log-Likelihood Method

The log-likelihood method correctly predicted Zhu Shuzhen as the author of the test tones. With a higher total log-likelihood for Zhu Shuzhen, this method accurately captured the tonal transitions characteristic of her style.

The reason this method worked well is that it directly uses the transition probabilities from the Markov matrix, which are based on observed tonal patterns in each poet's work. Even with a small test case, the log-likelihood method effectively quantifies how likely the transitions are for each poet, resulting in an accurate prediction. This approach is particularly robust when dealing with small datasets, as it inherently accounts for the probabilities of each transition in a cumulative manner.

Cosine Similarity and Euclidean Distance Methods

The equilibrium vector comparison using cosine similarity and Euclidean distance also predicted the correct authorship for both test cases. The cosine similarity values for Zhu Shuzhen's test tones were higher when compared to her own matrix (0.9789) versus Du Fu's (0.9729), and the Euclidean distance was lower for Zhu Shuzhen (0.1049) than Du Fu (0.1187), leading to the correct prediction for Zhu Shuzhen. Similarly, Du Fu's test tones showed higher cosine similarity with his own matrix (0.9756) than Zhu Shuzhen's (0.9598), and a lower Euclidean distance (0.1159) compared to Zhu Shuzhen (0.1483), resulting in the correct prediction for Du Fu.

While these methods performed well, slight discrepancies in the similarity scores and distances suggest that vector-based methods may be less robust when dealing with small datasets.

Here are several potential reasons that might render prediction inconsistency:

- **Small Test Case Size:** The limited size of the test sequence means that the calculated Markov matrix and equilibrium vector do not fully capture the general stylistic information. This can lead to inaccurate comparisons when using vector-based methods like cosine similarity and Euclidean distance.

- **Overfitting to Local Patterns:** The equilibrium vector derived from the small dataset may overemphasize specific transitions that are not representative of Zhu Shuzhen's overall style, leading to closer matches with Du Fu's matrix.
- **Sensitivity of Similarity Measures:** Cosine similarity and Euclidean distance are sensitive to the direction and magnitude of the vectors. Given the sparse data, these measures may incorrectly favor Du Fu's equilibrium vector.

Baseline Infinity Norm Method

The baseline infinity norm method also correctly predicted Zhu Shuzhen and Du Fu as the authors of their respective test tones. The infinity norm differences were smaller for the correct author in both test cases.

The simplicity of this method may contribute to its success, as it captures the general tonal patterns when comparing equilibrium vectors to the test tones.

Conclusion

The log-likelihood method provided highly accurate results, leveraging the probabilistic nature of the transition matrices and proving particularly reliable with limited data. The cosine similarity, Euclidean distance, and infinity norm methods also correctly predicted the authorship in both cases, though minor discrepancies in their values suggest that these methods may be more sensitive to small datasets.

Overall, the results demonstrate that while all methods performed well in this case, the log-likelihood method stands out for its robustness. The vector-based methods, including cosine similarity, Euclidean distance, and the infinity norm, provide correct predictions but may benefit from larger datasets or further refinement to more accurately capture subtle tonal transitions and stylistic nuances in general.