Project 1 Post-port

Hanzhang Yin

October 17, 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 two approaches: the log-likelihood method and a comparison of equilibrium vectors using cosine similarity and Euclidean distance.

Log-Likelihood Method

The log-likelihood method correctly predicted Zhu Shuzhen as the author of the test tones. The total log-likelihoods were computed as follows:

$$L_{Zhu} = 3\log\left(\frac{35}{103}\right) + 5\log\left(\frac{53}{156}\right) + 4\log\left(\frac{61}{199}\right) + \ldots = \text{(summed result)}$$

$$L_{Du} = 3\log\left(\frac{27}{80}\right) + \log\left(\frac{1}{3}\right) + 5\log\left(\frac{88}{279}\right) + \ldots = \text{(summed result)}$$

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

An alternative method was employed, comparing the equilibrium vectors of the Markov matrices using cosine similarity and Euclidean distance. However, both methods incorrectly predicted Du Fu as the author. The detailed results were:

- **Cosine Similarity:**
  - With Zhu Shuzhen: 0.5729
  - With Du Fu: 0.6121
- **Euclidean Distance:**

  - To Zhu Shuzhen: 0.8218
  - To Du Fu: 0.7970

Both methods favored Du Fu based on these metrics, which may be due to the following reasons:

- **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.

Conclusion

The log-likelihood method provided accurate results due to its probabilistic nature, directly utilizing the transition probabilities from the Markov matrices, which makes it more reliable even with small datasets. In contrast, the cosine similarity and Euclidean distance methods yielded incorrect predictions due to the limited ability of the small test case to capture comprehensive stylistic patterns. The cosine similarity values of 0.5729 with Zhu Shuzhen and 0.6121 with Du Fu, along with Euclidean distances of 0.8218 to Zhu Shuzhen and 0.7970 to Du Fu, highlight this discrepancy.
Overall, the results demonstrate the effectiveness of the log-likelihood method in authorship attribution tasks, especially when data is limited.