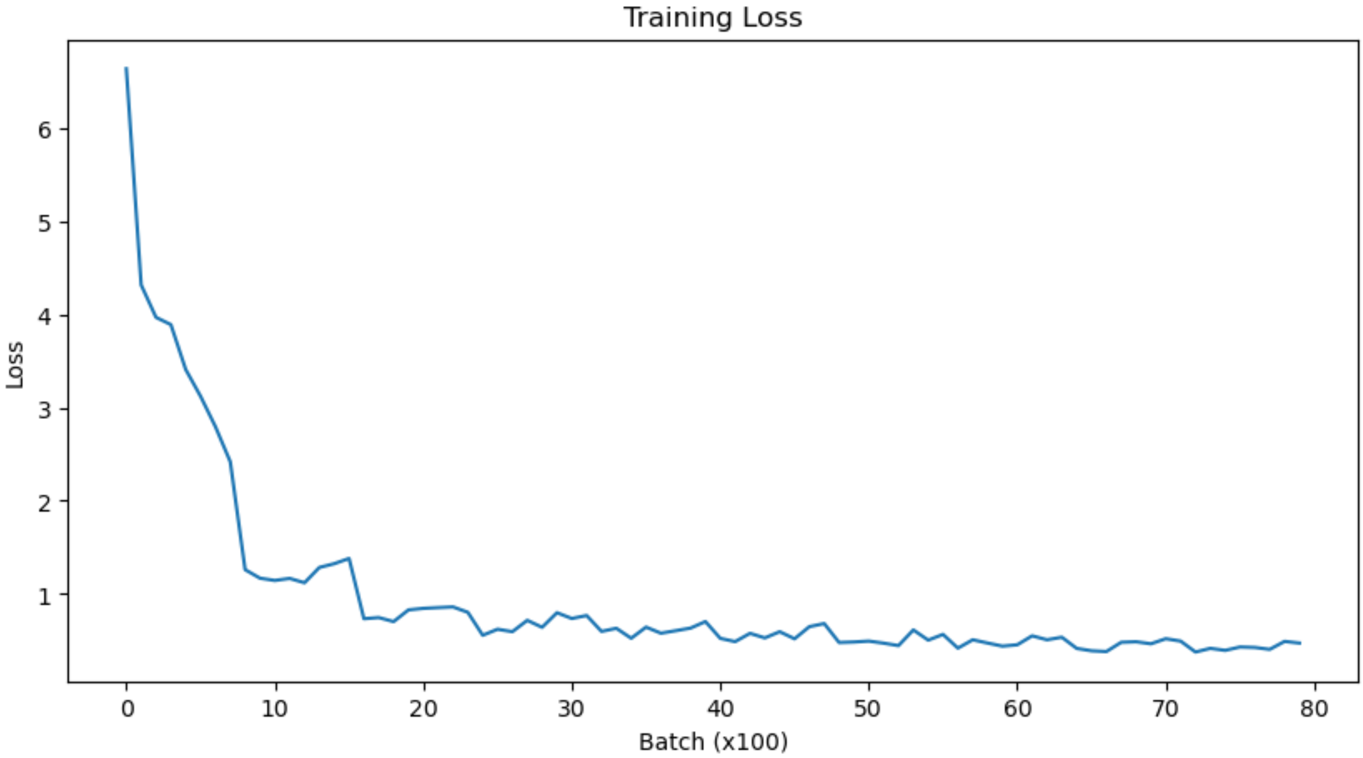
CS310 Natural Language Processing - Assignment 2: Word2vec Implementation

Task: Train a word2vec model using the skip-gram architecture and negative sampling.

- The corpus data being trained on is the full text of 《论语》.
- Use the code from **Lab 4** to help you.

3. Training Process Analysis

3.a Loss Variation



3.b Determination of Training Epochs

By observing the loss curve, it was found that after the 10th–20th epoch, the loss tended to stabilize. Therefore, we chose **15 epochs** as the final training duration, ensuring that the model is sufficiently trained without overfitting.

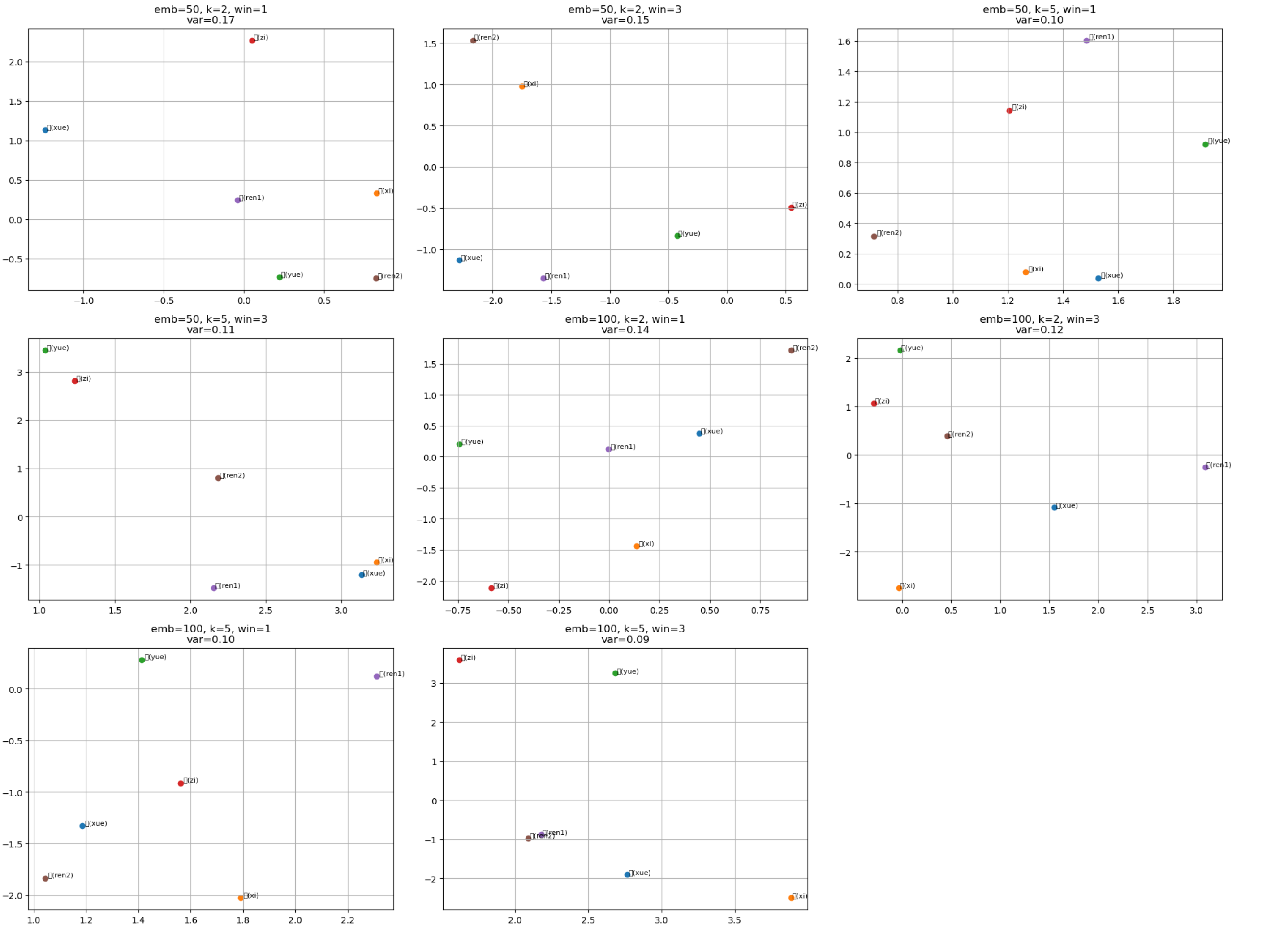
4. Hyperparameter Experiment Results

4.a Experiment Setup

- **Embedding Dimension (emb_size):** 50, 100
- **Negative Sampling Count (k):** 2, 5

- **Window Size (window_size):** 1, 3

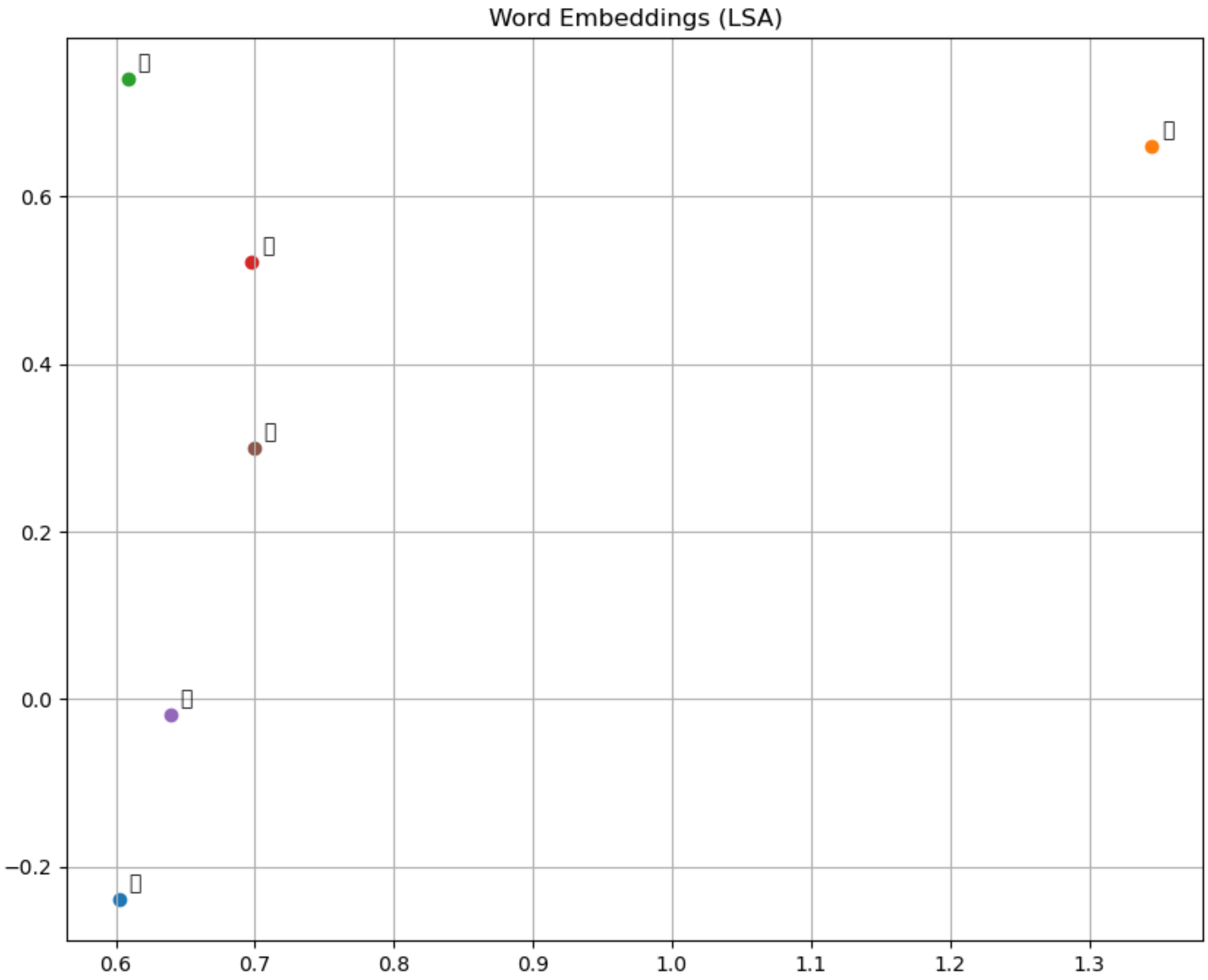
5. Embedding Vector Visualization Analysis



5.b Comparison with LSA Method

We compare the results of **emb_size=100, k=5, window_size=3** with the LSA results from Lab 4:

解释方差比: 0.0895 词对相似度分析： 学-习: 0.2292 子-曰: 0.4987 人-仁: 0.1127



- Similarities:
  - Semantically related words (e.g., "学" and "习") exhibit close proximity in both methods.
- Differences:
  - Word2Vec captures contextual relationships better.
  - LSA focuses more on global co-occurrence statistics.
  - Word2Vec performs better in identifying synonyms.