Word2Vec với thuật toán Skip-gram

1.Dataset

Dataset tập dataset 5000 câu tiếng Việt

2. Pre-processing

Đầu tiên, xác định các giá trị của các hyperparameter: context_size, vocab_size:
Với vocab_size là toàn bộ các từ đơn xuất hiện trong bộ dataset. Sau đó, với mỗi câu Tiếng Việt, chọn ra một từ tại vị trí i gọi là target word (0 <= i <= độ dài câu -1), với mỗi target word xác định các context word xung quanh, context words là các từ cần phải thoả mãn nằm ở vị trí (0 <= i <= độ dài câu - 1).
Sau khi xác định các cặp đó xong, với từ target ta ghép gặp với mỗi từ trong context words.

3. Mô hình

Chúng tôi sử dụng lớp Embedding và Linear trong thư viện Pytorch để triển khai mô hình

4. Huấn luyện

Với mỗi cặp từ (target_word, context_word), ta có P(context_word | target_word) mục đích của ta là đi tối ưu hoá giá trị xác suất. Ở đây chúng tôi sử dụng CrossEntropyLoss để tối đa hoá giá trị xác suất.

5. Trực quan hoá