2 Natural Language Processing

2.1 Word Embeddings

The most basic form word embedding is the one-hot representation, which uses a vector the same size as the vocabulary, and a given word has one index of value 1 and the rest of value 0. The problem with this one-hot representation, however, is that the vectors are very high-dimensional and sparse, and all words are the same distance from one another. A better idea represents each word with a distributed representation that is not sparse and less high-dimensional. These word embeddings will have the property that words with similar meaning will be closer in the embedding space.

Training these word embeddings typically uses some clever formulation of context-target pairs for self-supervised learning. Here are some of the popular methods for learning word embeddings:

1. **Skip-gram** (Mikolov+ '13). Sample one context and one target word within some window, and then use the context word to predict the target word. (You are trying to predict some word "skipping" a few words to the left or right.) This works fine, but computing the softmax over 10,000 outputs (your vocabulary size) is expensive.
2. **Negative sampling** (Mikolov+ '13). Reformulate skip-gram as a binary classification task. Given two context words, predict whether those two context words were indeed from the same window. You can negatively sample somewhere between the observed frequency and uniform distribution.