Tian Xiaoyang 26001904581

Tokenization: in natural language processing, tokenization is the process of dividing texts into smaller units like sentences or words.

One-hot encoding:

TFIDF: term frequency inverse document frequency. DF, document frequency, is the number of occurrences of a specific term in a document set, meaning number of documents that contain the term.

Pretraining: the process of training an NLP model on a large amount of general-domain language data so it can be trained later using smaller quantity of domain-specific data. It is a form of transfer learning.

CNN: CNN is convolutional neural network; it contains spatially local connections and a pattern of weights replicated across units in each layer. Kernel is the pattern of weights that is replicated across multiple local regions

RNN: recurrent neural networks allow cycles in their computation graphs. In RNN each cycle as a delay, units in the layers may take input as a value calculated by themselves from an earlier cycle. In this case, the RNN can be said to have a memory, prior inputs affect current outputs.