Word Embeddings

**Word embedding in NLP** is an important term that is used for representing words for text analysis in the form of real-valued vectors.

In this approach, words and documents are represented in the form of numeric vectors allowing similar words to have similar vector representations. The extracted features are fed into a machine learning model to work with text data and preserve the semantic and syntactic information.

Word2Vec

Word2vec is a technique in natural language processing (NLP) for obtaining vector representations of words. These vectors capture information about the meaning of the word based on the surrounding words.

Word2vec represents a word as a high-dimension vector of numbers which captures relationships between words. In particular, words which appear in similar contexts are mapped to vectors which are nearby as measured by cosine similarity.

Word2vec Two Model Architectures

1.Continuous Bag-Of-Words (CBOW)

This model predicts the target word based on the surrounding context words within a fixed window size.

The input to the model is the one-hot encoded vectors representing the context words.

2. Skip Gram

In the Skip-gram architecture, the model predicts the context words given a target word.

The input to the model is the one-hot encoded vector representing the target word.

GloVe (Global Vectors for Word Representation)

It constructs a co-occurrence matrix from the input text corpus and learns word embeddings by optimizing a global objective function that captures word co-occurrence statistics.

GloVe captures semantic meaning by leveraging the global statistical information of word co-occurrences across the entire dataset.

It learns dense vector representations of words that reflect the distributional properties of words in the corpus, emphasizing meaningful semantic relationships between words.

Context window

The size of the context window determines how many words before and after a given word are included as context words of the given word.

Limitations of Word2Vec and GloVe

Out-of-Vocabulary Words:

Word2Vec and GloVe struggle with out-of-vocabulary words, words that were not present in the training corpus.

Fixed Dimensionality:

This fixed dimensionality may not be sufficient to capture all the semantic information present in complex words or phrases.

Polysemy and Homonymy:

Word2Vec and GloVe may struggle with polysemous words (words with multiple meanings) and homonymous words (words that are spelled the same but have different meanings).

Despite these limitations, Word2Vec and GloVe remain popular and widely used techniques for learning word embeddings due to their simplicity, efficiency, and effectiveness in capturing semantic similarities between words.