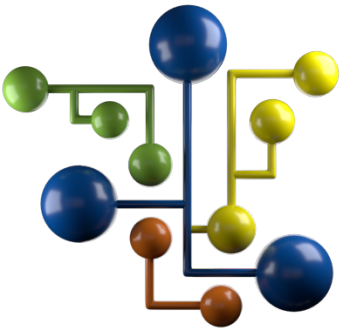


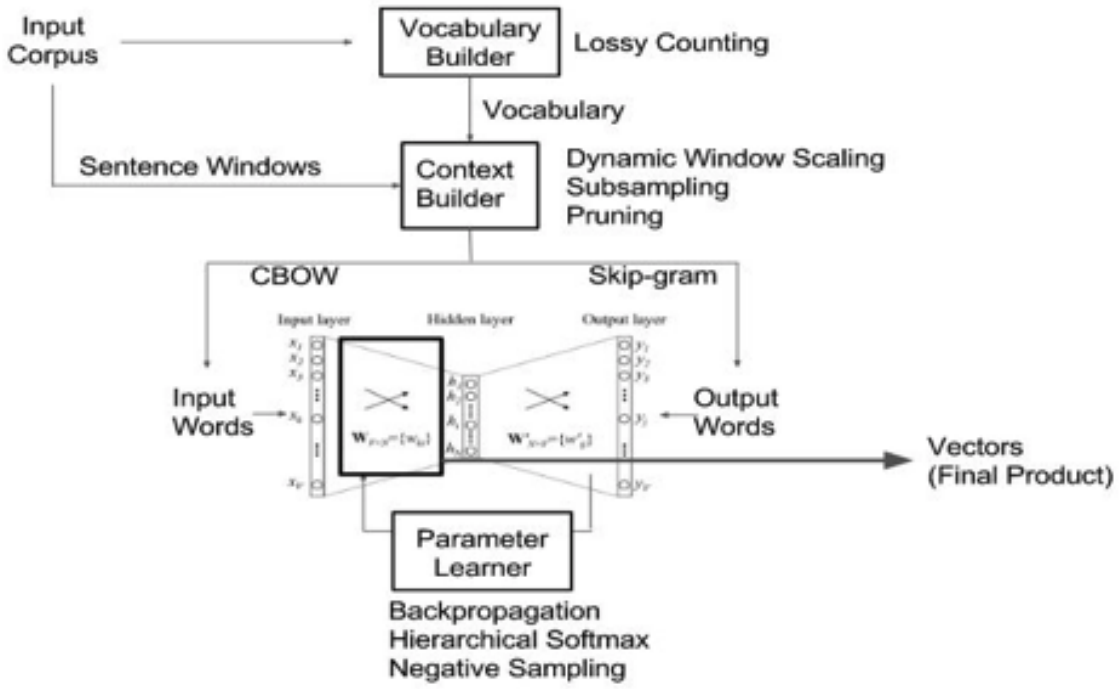
Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Processo de Construção de um Modelo Word2vec

Vamos começar a decompor o modelo word2vec e tentar entender a lógica dele e seu processo de construção. O word2vec é um software e usa vários algoritmos, sendo composto de várias partes, conforme imagem abaixo:

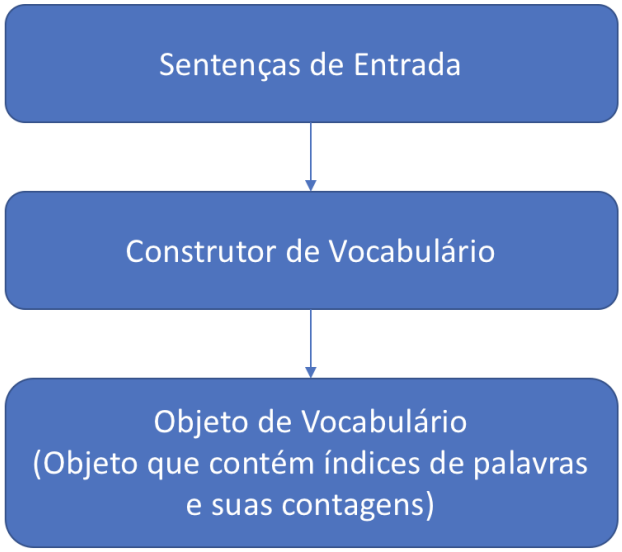


Como você pode ver na figura acima, existem três blocos de construção principais. Vamos examinar cada um deles em detalhes:

- Construtor de vocabulário (Vocabulary Builder)
- Construtor de contexto (Context Builder)
- Rede neural com duas camadas (Neural Net)

Construtor de Vocabulário

O construtor de vocabulário é o primeiro bloco de construção do modelo Word2vec, conforme vimos no capítulo anterior. Ele recebe dados de texto não processados, principalmente na forma de sentenças. O construtor de vocabulário é usado para construir o vocabulário a partir do seu corpus de texto. Ele irá coletar todas as palavras únicas do seu corpus e construir o vocabulário. Cada palavra presente no vocabulário tem uma associação com o objeto de vocabulário, que contém um índice e uma contagem. Essa é a saída do construtor de vocabulário, conforme você pode observar na figura abaixo, que ajuda a entender a entrada e a saída do construtor:



Construtor de Contexto

O construtor de contexto usa a saída do construtor de vocabulário (assim como as palavras que fazem parte da janela de contexto) como entrada e gera a saída. Primeiro de tudo, vamos entender o conceito de uma janela de contexto.

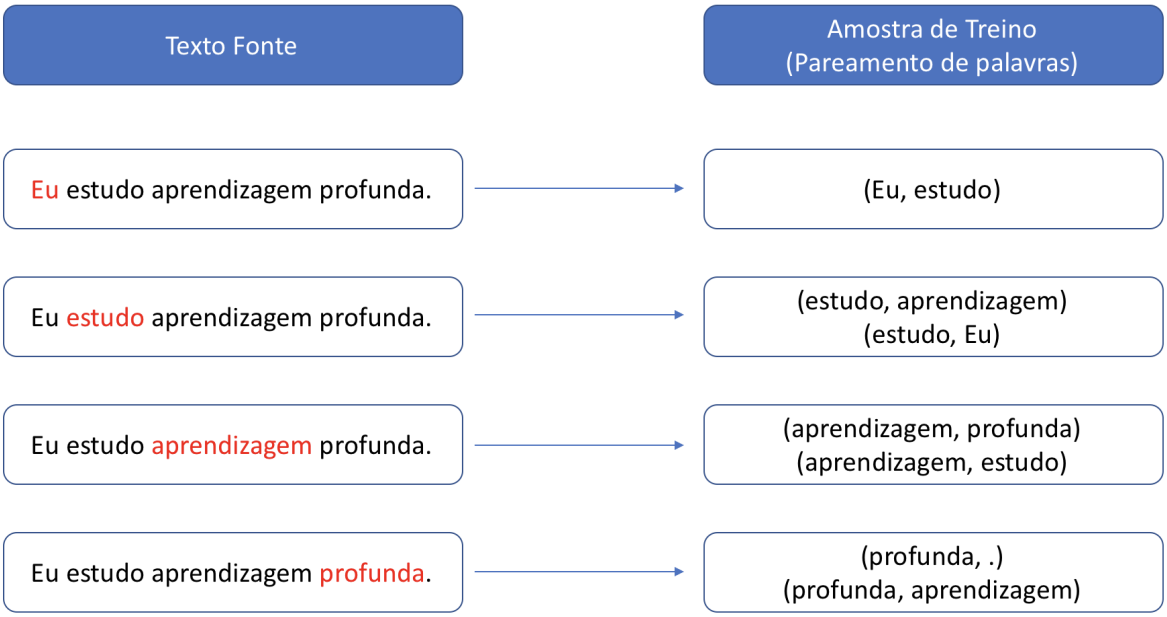
A janela de contexto é uma espécie de janela deslizante. Você pode definir o tamanho da janela conforme o aplicativo de PLN no qual você usará o Word2vec. Geralmente, os aplicativos de PLN usam o tamanho da janela de contexto de cinco a dez palavras. Se você decidir ir com um tamanho de janela de cinco, então precisamos considerar as cinco palavras do lado esquerdo da palavra central e as cinco palavras do lado direito da palavra central. Dessa forma, capturamos as informações sobre o que são todas as palavras ao redor de nossa palavra central. Vejamos um exemplo e, para isso, o tamanho da janela de contexto é igual a um, pois temos um corpus de apenas uma frase:

Eu estudo aprendizagem profunda.

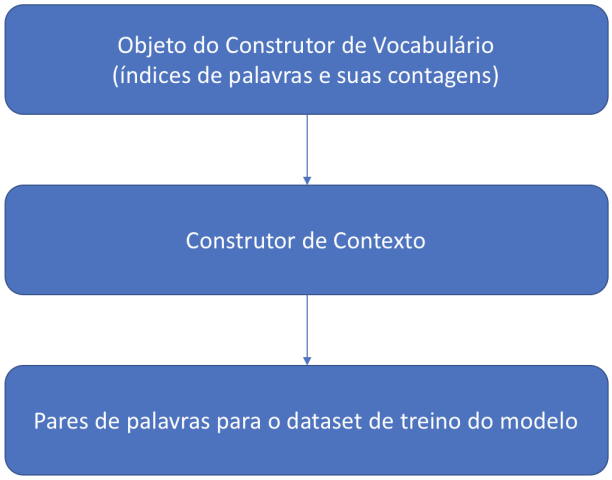
e profunda é a palavra central. Então, você deve considerar as palavras em torno de acordo com o tamanho da nossa janela. Assim, precisamos considerar as palavras estudo e aprendizagem. Na próxima iteração, nossa palavra central estará aprendendo que as palavras que a cercam são profunda e, no final da frase, um ponto (.). Espero que o conceito de janela de contexto esteja claro em sua cabeça. Agora, precisamos vincular esse conceito e ver como o construtor de contexto usa esse conceito e a saída do construtor de vocabulário.

O objeto construtor de vocabulário possui índices de palavras e a frequência da palavra no corpus. Usando o índice da palavra, o construtor de contexto tem uma ideia de qual palavra estamos observando e, de acordo com o tamanho da janela de contexto, ele considera as outras palavras adjacentes.

Essas palavras centrais e as outras palavras adjacentes são inseridas no construtor de contexto. Agora você tem uma ideia clara sobre quais são as entradas para o construtor de contexto. E quais são as saídas do construtor de contexto? Esse construtor de contexto gera o pareamento de palavras. Veja a imagem abaixo para compreender isso em mais detalhes:



Esses pares de palavras serão então enviados à rede neural para treinamento. A rede aprenderá as estatísticas básicas do número de vezes que cada par de palavras aparece. Assim, por exemplo, a rede neural provavelmente terá muito mais exemplos de treinamento (profunda, aprendizagem) do que de (profunda, comunicação). Quando o treinamento estiver terminado, se você der a palavra profunda como entrada, então o modelo produzirá uma probabilidade muito maior para “aprendizagem” do que para “comunicação”. Portanto, este par de palavras é a saída do construtor de contexto e passará para o próximo componente, que é uma rede neural de duas camadas, conforme imagem abaixo:



Rede Neural

O Word2vec usa a rede neural para treinamento e por isso é muito importante entender a estrutura básica da rede neural. Os detalhes estruturais de uma rede neural são dados da seguinte forma:

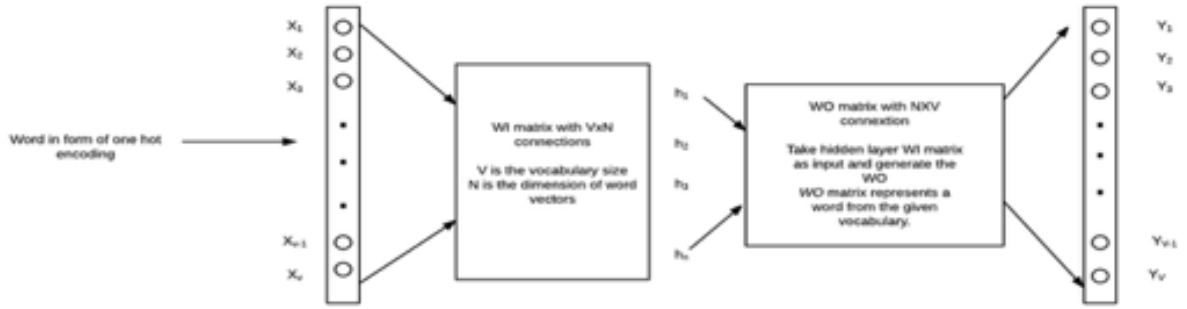
- Há uma camada de entrada
- A segunda camada é a camada oculta
- A terceira e última camada é a camada de saída

Vamos começar a olhar para cada uma das camadas e sua entrada e saída em detalhes. Vamos entender a tarefa de cada camada em resumo:

- **Camada de entrada**: Uma camada de entrada tem tantos neurônios quanto há palavras no vocabulário para treinar.
- **Camada oculta**: O tamanho da camada oculta em termos de neurônios é a dimensionalidade dos vetores de palavras resultantes.
- **Camada de saída**: A camada de saída tem o mesmo número de neurônios que a camada de entrada.

A entrada para a primeira camada de entrada é a palavra com One-hot Encoding. Suponha que o tamanho do nosso vocabulário para aprender vetores de palavras seja V, o que significa que existem números V de diferentes palavras no corpus. Nesse caso, a posição da palavra que representa ela mesma é codificada como 1 e todas as outras posições são codificadas como 0.

Suponha que a dimensão dessas palavras seja N. Assim, a entrada para as conexões da camada oculta pode ser representada por nossa matriz de entrada WI (símbolo da matriz de entrada) do tamanho V * N, com cada linha da matriz WI representando uma palavra do vocabulário. Do mesmo modo, as ligações da camada oculta para a camada de saída significam que a saída da camada oculta pode ser descrita pela matriz de saída da camada oculta WO (símbolo da matriz da camada oculta). A matriz WO é do tamanho N * V. Neste caso, cada coluna da matriz WO representa uma palavra do vocabulário fornecido. A imagem abaixo oferece uma visão cristalina da entrada e saída. Além disso, também veremos um pequeno exemplo para entender o conceito:



Isso foi o que estudamos no capítulo anterior com um exemplo completo no TensorFlow. Na próxima aula, vamos ver um exemplo onde este conceito ficará ainda mais claro, antes de estudarmos itens mais avançados com o Word2vec.