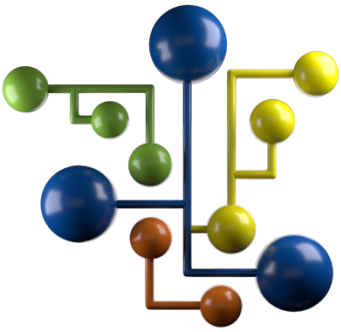

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Encoders e Decoders
One-Hot Encoding

O conceito de codificação em PLN é bastante antigo e útil. Como já mencionamos anteriormente, não é fácil lidar com atributos de dados categóricos presentes em nosso conjunto de dados. Aqui, exploraremos a técnica de codificação denominada One-Hot Encoding, que nos ajuda a converter nossos recursos categóricos em um formato numérico.

Em um aplicativo de PLN, você sempre obtém dados categóricos. Os dados categóricos estão principalmente na forma de palavras e as palavras que formam o vocabulário. As palavras deste vocabulário não podem se transformar em vetores facilmente. Considere que você tem um vocabulário com o tamanho N. A maneira de aproximar o estado do idioma é representando as palavras na forma de uma codificação One-Hot Encoding. Essa técnica é usada para mapear as palavras para os vetores de comprimento n, onde o n-ésimo dígito é um indicador da presença da palavra em particular. Se você estiver convertendo palavras para o formato de codificação One-Hot, você verá vetores como 0000 ... 001, 0000 ... 100, 0000 ... 010 e assim por diante. Cada palavra no vocabulário é representada por uma das combinações de um vetor binário. Aqui, o enésimo bit de cada vetor indica a presença da enésima palavra no vocabulário. Então, como esses vetores individuais são relacionados a sentenças ou outras palavras no corpus? Vejamos um exemplo que ajudará você a entender esse conceito. Por exemplo, você tem uma frase:

"Maria gosta de PLN"

Suponha que depois de aplicar uma codificação One-Hot Encoding, essa sentença se torne 00010 00001 10000. Esse vetor é feito com base no tamanho do vocabulário e no esquema de codificação. Quando tivermos essa representação vetorial, podemos executar a operação numérica nela. Aqui, estamos transformando palavras em vetores e sentenças em matrizes.

Essas técnicas são muito úteis. Vamos ver algumas das aplicações básicas para esta técnica de mapeamento:

- Muitas redes neurais artificiais aceitam dados de entrada no formato de codificação One-Hot Encoding e geram vetores de saída que carregam a representação semântica também.

- O algoritmo word2vec aceita dados de entrada na forma de palavras e as palavras estão na forma de vetores que são gerados por One-Hot Encoding.

Agora é hora de olhar para o conceito de decodificação. Conceitos de decodificação são usados principalmente em aprendizado profundo hoje em dia. Então, aqui, definiremos o decodificador em termos de aprendizado profundo, pois usaremos essa arquitetura de codificação e decodificação mais a frente com Deep Learning para NLU e NLG, para desenvolver um sistema de tradução.

Um codificador mapeia dados de entrada para uma representação de recurso diferente e podemos usar One-Hot Encoding para isso. Um decodificador mapeia a representação do recurso de volta para o espaço de dados de entrada. Na aprendizagem profunda, um decodificador sabe qual vetor representa quais palavras, para que ele possa decodificar palavras de acordo com o esquema de entrada fornecido. Veremos o conceito detalhado do codificador-decodificador quando cobrirmos o modelo de sequence-to-sequence.