



**Data Science
Academy**

www.datascienceacademy.com.br

Processamento de Linguagem Natural

**Técnicas Usadas Para Gerar os Vetores de
Saída e Probabilidades**

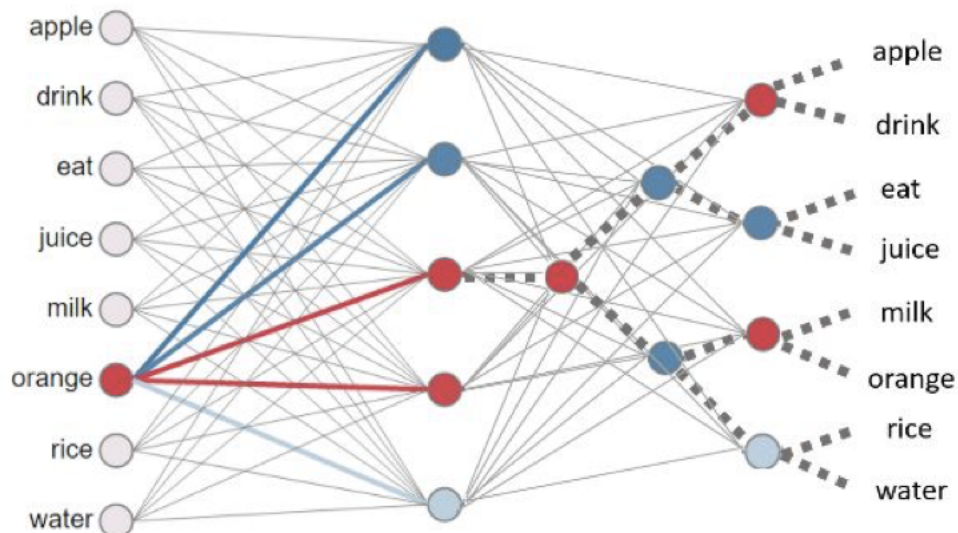
Nesta seção, veremos como vamos gerar o vetor final. Também usaremos algumas heurísticas para gerar saída eficientemente. Então, vamos falar sobre essas heurísticas também. Como já vimos, para gerar o vetor de palavras, precisamos atualizar a entrada e o vetor de saída. Suponha que tenhamos um milhão de palavras em nosso vocabulário, portanto, o processo de atualização dos vetores de entrada e saída levará muito tempo e será ineficiente. Temos que resolver esse desafio. Por isso, usamos algumas das maneiras otimizadas de executar as mesmas operações e essas técnicas são fornecidas da seguinte forma:

- Hierarchical Softmax
- Negative Sampling

Então, vamos começar a entender essas técnicas.

Hierarchical Softmax

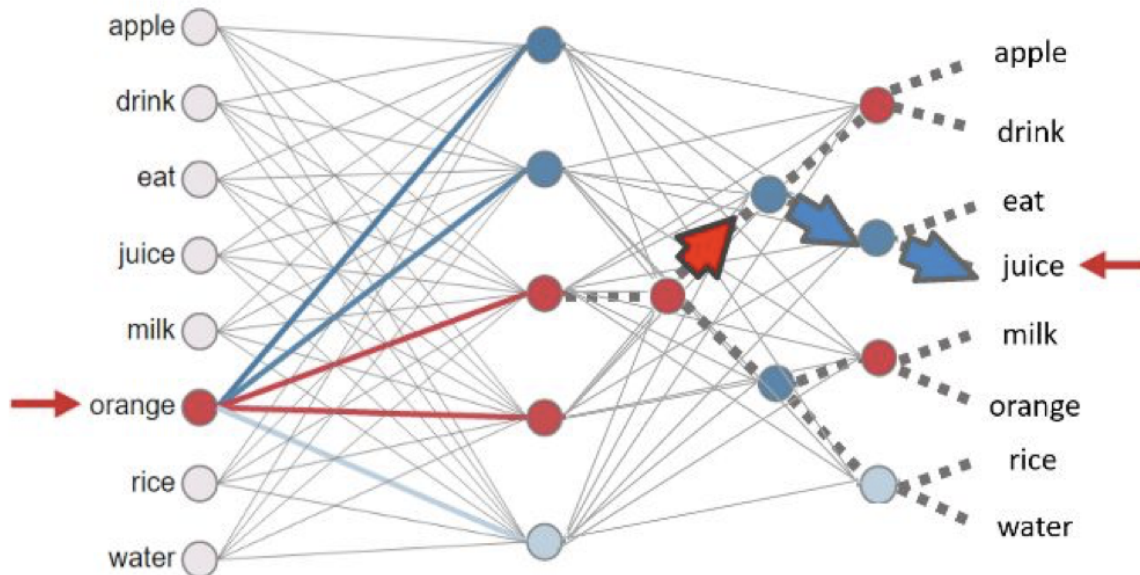
Com o softmax hierárquico, em vez de mapear cada vetor de saída para sua palavra correspondente, consideramos o vetor de saída como uma forma de árvore binária, similar a imagem abaixo:



Aqui, o vetor de saída não está fazendo uma previsão sobre o quão provável a palavra é, mas está fazendo uma previsão sobre qual caminho você quer seguir na árvore binária.

Neste caso, considere o ponto vermelho ativado e o ponto azul ativado indo para baixo). Conforme você pode ver aqui, podemos prever a palavra suco (juice) com alta probabilidade. A vantagem é que, se você quiser retroceder um erro, basta atualizar um vetor de saída e o erro

se propagará para apenas três nós ativados no momento da previsão. Usamos a construção da árvore binária de Huffman para gerar a árvore binária.



Negative Sampling

A amostragem negativa também é um tipo de método de otimização. Neste método, vamos atualizar o vetor da palavra de saída, mas não atualizaremos todos os vetores de outras palavras. Nós apenas pegamos a amostra das palavras além do vetor de saída. Então, estamos selecionando uma amostra do conjunto de palavras negativo da amostra, daí o nome dessa técnica ser uma amostragem negativa.

As duas técnicas serão abordadas durante as próximas aulas.