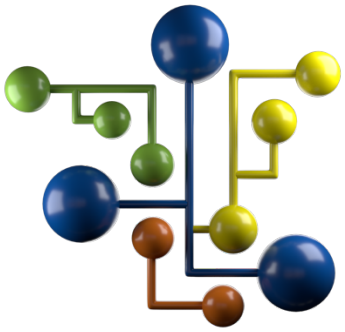

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Técnicas Usadas Para Construção de Vocabulário

Ao gerar o vocabulário do conjunto de dados, você pode usar uma técnica de otimização, e a contagem com perdas (Lossy Counting) é a mais usada para o modelo Word2vec.

Contagem com Perdas (Lossy Counting)

O algoritmo de contagem com perdas é usado para identificar elementos em um conjunto de dados cuja contagem de frequência excede um limite determinado pelo usuário. Esse algoritmo usa fluxos de dados como uma entrada em vez do conjunto finito de um conjunto de dados. Com a contagem com perdas, você remove periodicamente os elementos de contagem muito baixa da tabela de frequência.

As palavras acessadas com mais frequência quase nunca teriam contagens baixas, e, se o fizessem, provavelmente não ficariam ali por muito tempo. Aqui, o limite de frequência é geralmente definido pelo usuário. Quando definimos um parâmetro de **min_count = 4**, removemos as palavras que aparecem no conjunto de dados menos de quatro vezes e não as consideraremos. Isso é definido durante a construção do modelo.

Contagem com perdas é muito útil, especialmente quando você tem um corpus grande e não quer considerar as palavras que aparecem muito raramente. Neste momento, a contagem com perdas é muito útil porque o usuário pode definir a contagem de frequência mínima de palavras como um limite, portanto, palavras que ocorram abaixo da contagem de frequência não serão incluídas em nosso vocabulário. Se você tem um corpus grande e deseja otimizar a velocidade do treinamento, podemos usar esse algoritmo. Em outras palavras, você pode dizer que usando esse algoritmo você reduz o tamanho do seu vocabulário, assim, você pode acelerar o processo de treinamento.

Além do word2vec, o algoritmo **Lossy Counting** é usado em medições de tráfego de rede e análise de logs de servidor web.

Referências:

https://www.semanticscholar.org/