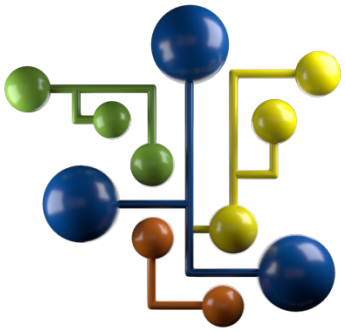

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Técnicas Usadas Para Construção de Contexto

Ao gerar pares de contexto de palavras, o construtor de contexto usa as seguintes técnicas:

- Dynamic Window Scaling ou Dynamic Context Window
- Subsampling
- Pruning

Dynamic Window Scaling ou Dynamic Context Window

O dimensionamento dinâmico de janelas (Dynamic Window Scaling) faz parte do construtor de contexto. Vamos ver como isso pode ser útil e que tipo de impacto ele gera quando o usamos. O dimensionamento dinâmico de janelas também é conhecido como janela de contexto dinâmico (Dynamic Context Window).

Na implementação Word2vec, a janela de contexto dinâmico é uma técnica opcional que pode ser aplicada para gerar uma saída mais precisa. Você também pode considerar essas técnicas como hiperparâmetros. As técnicas de janela de contexto dinâmico usam o esquema de peso para palavras de contexto em relação às palavras de destino. Portanto, a intuição aqui é que as palavras que estão próximas da palavra alvo são mais importantes do que outras palavras que estão distantes da palavra alvo. Vamos ver como isso será útil quando estamos construindo pares de palavras.

A janela de contexto dinâmico considera que as palavras de contexto próximas têm mais importância para prever a palavra de destino. Aqui, estamos aplicando o esquema de ponderação usando amostragem uniforme no tamanho real da janela entre 1 e L. Por exemplo, suponha que o tamanho da janela de contexto seja 5 e agora o peso das palavras de contexto seja distribuído de maneira uniforme, portanto o peso de a maioria das palavras próximas é 5/5, o próximo peso da palavra de contexto é 4/5, e assim por diante. Assim, o peso final para palavras de contexto será 5/5, 4/5, 3/5, 2/5, 1/5. Assim, fornecendo peso, você pode ajustar o resultado final.

Subsampling

A subamostragem também é uma das técnicas que usamos quando estamos criando pares de palavras e, como sabemos, esses pares de palavras são exemplos de dados de treinamento. A subamostragem é o método que remove as palavras mais frequentes. Essa técnica é muito útil para remover palavras de parada (stop words). Essa técnica também remove palavras aleatoriamente, e essas palavras escolhidas aleatoriamente ocorrem no corpus com mais frequência. Assim, as palavras que são removidas são mais frequentes do que um limiar t com uma probabilidade de p, onde f marca a frequência do corpus das palavras e usamos t = 10 - 5 em nossos experimentos, conforme fórmula abaixo:

$$p = 1 - \sqrt{\frac{t}{f}}$$

Isso também funciona como um dos hiperparâmetros úteis, e é muito útil porque estamos removendo as palavras mais frequentes e desnecessárias do corpus, bem como da janela de contexto, e dessa forma, estamos melhorando a qualidade de nossa amostra de treinamento.

Pruning

Pruning também é usado quando estamos construindo nossos pares de palavras para fins de treinamento usando o construtor de contexto. Quando você tem uma grande quantidade de vocabulário para lidar, se você incluiu palavras menos frequentes, então você precisa removê-las. O Pruning é usado para podar o tamanho da amostra de treinamento, bem como melhorar a qualidade dela. Se você não remover as palavras raramente encontradas do conjunto de dados, a precisão do modelo poderá ser reduzida. Este é um tipo de truque para melhorar a precisão.

Referências:

https://www.semanticscholar.org/