Processamento de Linguagem Natural com Transformers

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) é um algoritmo de redução de dimensionalidade utilizado para visualização de dados em ciência de dados e aprendizado de máquina. Foi criado por Leland McInnes, John Healy e James Melville.

UMAP é muito eficaz para lidar com dados de alta dimensão, e em muitos casos supera os métodos tradicionais de redução de dimensionalidade como o t-SNE (t-Distributed Stochastic Neighbor Embedding). Em comparação com o t-SNE, o UMAP tende a ser mais rápido e escalar melhor com grandes conjuntos de dados, enquanto ainda mantém uma estrutura de dados global e local significativa. UMAP tem duas etapas principais:

Construção do gráfico de vizinhança: UMAP começa calculando a distância entre cada ponto em um conjunto de dados e seus vizinhos mais próximos. Essas distâncias são usadas para construir um gráfico ponderado de alta dimensão, onde cada ponto é um nó e as arestas entre nós representam a proximidade relativa entre pontos.

Otimização da projeção de baixa dimensão: UMAP então usa uma técnica chamada "otimização de layout" para encontrar uma representação de baixa dimensão do gráfico de alta dimensão que preserva a estrutura global e local do gráfico original o máximo possível.

A ideia principal do UMAP é preservar tanto as relações de distância globais (o que é um ponto distante de outro) quanto as relações de distância local (o que é um ponto perto de outro) em um espaço de menor dimensão.

UMAP é um método poderoso e flexível que pode ser usado tanto para visualização quanto para transformação de dados em um pipeline de aprendizado de máquina. É também uma ferramenta eficaz para explorar a estrutura de conjuntos de dados complexos. O UMAP é uma das técnicas usadas no BERTopic.

O link abaixo tem detalhes sobre o funcionamento do algoritmo:

https://umap-learn.readthedocs.io/en/latest