



Processamento de Linguagem Natural com Transformers

Processamento de Linguagem Natural com Transformers

Modelo BERTopic

BERTopic é uma técnica de modelagem de tópicos que utiliza transformações de linguagem para criar representações ricas de documentos e, em seguida, emprega um algoritmo de clustering de alta densidade (HDBSCAN) e classificação para extrair os tópicos.

Ao contrário da Latent Dirichlet Allocation (LDA), uma técnica de modelagem de tópicos tradicional que se baseia em contagens de palavras e suposições de distribuições estatísticas, BERTopic usa o BERT para transformar documentos inteiros em representações vetoriais que capturam o significado semântico das palavras e do contexto em que são usadas.

Estas são as etapas gerais que o BERTopic segue:

Extração de recursos: BERTopic começa transformando os documentos em embeddings usando modelos de transformadores de linguagem como BERT. Cada documento é convertido em um vetor de alta dimensão que captura o contexto e o significado semântico do texto.

Redução de dimensões: Em seguida, BERTopic usa a Redução de Dimensionalidade Uniforme de Manifold Aproximado (UMAP) para reduzir esses vetores de alta dimensão a um espaço de dimensões mais baixas que ainda preserva as relações semânticas entre os documentos.

Clustering: BERTopic, em seguida, aplica o algoritmo de clustering HDBSCAN nos vetores de dimensão reduzida para identificar clusters de documentos que são semelhantes em termos de conteúdo.

Extração de tópicos: Por fim, para cada cluster, BERTopic identifica as palavras mais representativas para criar um "tópico". Essas palavras são determinadas calculando a contribuição de uma palavra para um tópico com base na frequência da palavra nesse tópico e na diversidade de tópicos em que a palavra aparece.

BERTopic é uma abordagem poderosa e flexível para a modelagem de tópicos que se beneficia das representações ricas em contexto fornecidas pelos modelos de transformadores de linguagem. Usaremos o BERTopic mais tarde neste capítulo.