



Processamento de Linguagem Natural com Transformers

Processamento de Linguagem Natural com Transformers

Class-based TF-IDF (Term Frequency-Inverse Document Frequency)

Class-based TF-IDF, ou c-TF-IDF, é uma variação do método TF-IDF (Term Frequency-Inverse Document Frequency) comumente usado na análise de texto. TF-IDF é uma maneira de quantificar a importância de uma palavra em um documento em relação a uma coleção de documentos, levando em consideração tanto a frequência da palavra no documento (TF) quanto a inversão da frequência da palavra em todos os documentos (IDF).

Em sua essência, c-TF-IDF é aplicado da mesma maneira que o TF-IDF regular, exceto que, em vez de calcular as pontuações para cada documento individualmente, c-TF-IDF calcula as pontuações para cada classe ou grupo de documentos. Isso faz com que ele se concentre mais nas palavras que são importantes para uma classe específica, em vez de palavras que são importantes para um único documento.

O c-TF-IDF pode ser útil em cenários onde você está interessado em encontrar palavras que são distintivas para diferentes grupos ou classes de documentos. Por exemplo, se você estiver analisando textos de diferentes gêneros literários, c-TF-IDF pode ajudar a identificar palavras que são particularmente importantes para o gênero de romance, mas não para o gênero de ficção científica.

Aqui está uma descrição passo a passo de como calcular c-TF-IDF:

- Combine todos os documentos dentro de uma mesma classe para formar um "documento de classe".
- Calcule a frequência de termo (TF) de cada palavra no documento de classe, que é a frequência da palavra no documento de classe.
- Calcule a frequência inversa de documento (IDF) de cada palavra, que é o logaritmo do total de classes dividido pelo número de classes em que a palavra aparece.
- Multiplique o TF e IDF de cada palavra para obter a pontuação c-TF-IDF.

As palavras com a maior pontuação c-TF-IDF são as palavras mais distintivas ou importantes para essa classe em particular.

O BERTopic usa o c-TF-IDF para definir os termos mais importantes em cada tópico.