



Processamento de Linguagem Natural com Transformers

Processamento de Linguagem Natural com Transformers

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

HDBSCAN é um algoritmo de agrupamento (clusterização) baseado em densidade que é uma extensão do DBSCAN (Density-Based Spatial Clustering of Applications with Noise). O "HD" no HDBSCAN significa Hierarchical DBSCAN, e, como o nome indica, este algoritmo utiliza uma abordagem hierárquica para encontrar clusters de alta densidade em um conjunto de dados.

Aqui estão as principais etapas do algoritmo HDBSCAN:

Transformação de Espaço de Características: Inicialmente, o algoritmo transforma o espaço de características originais em um espaço que reflete a densidade de dados. Isto é feito criando uma árvore de abrangência mínima ponderada em que as arestas têm pesos iguais à distância entre os pontos que eles conectam.

Criação da Hierarquia: Em seguida, o algoritmo constrói uma hierarquia de clusters, começando com cada ponto como seu próprio cluster e então fundindo iterativamente os clusters mais próximos. A proximidade é determinada pelas distâncias entre os pontos mais próximos nos clusters (ao contrário do DBSCAN tradicional, que considera apenas um raio fixo).

Formação de Clusters Estáveis: O HDBSCAN identifica clusters "estáveis" como regiões da hierarquia de cluster que têm uma alta densidade relativa de pontos. A estabilidade de um cluster é definida pela persistência de sua existência ao longo da hierarquia.

Extração e Seleção de Clusters: Por último, o algoritmo extrai os clusters estáveis da árvore de cluster, descarta os clusters menos estáveis e trata os pontos restantes como ruído.

O HDBSCAN tem várias vantagens em relação a outros algoritmos de agrupamento. Ele pode encontrar clusters de várias formas e tamanhos, diferentemente de algoritmos baseados em partição (como k-means) que tendem a encontrar clusters esféricos. Além disso, o HDBSCAN não requer que o usuário especifique o número de clusters antecipadamente e é capaz de identificar pontos de ruído que não pertencem a nenhum cluster.

O HDBSCAN é a outra técnica usada no BERTopic.