**Conceitos Aprendidos**

**BERTopic** – método usado para encontrar tópicos: faz se clusters e consoante as palavras dos textos em cada um dos clusters é composto pelos tweets de um tópico.

**Embedding** - Cada palavra, frase ou documento é transformado num vetor de números.

Ex: “carro” e “automóvel” → vetores muito próximos.

**c-TF-IDF** – versão adaptada do TF-IDF que, em vez de olhar para cada documento separado, calcula a importância das palavras em grupos de documentos (ou seja, nos clusters).

**MiniLM** – modelo de linguagem “leve” baseado em transformers que gera representações de texto rápidas e eficientes, mantendo uma boa qualidade mesmo sem precisar de muito poder de computação.

**HDBSCAN** - algoritmo de clustering que não só agrupa pontos semelhantes, mas também atribui uma probabilidade de um ponto pertencer a um grupo. Apesar de cada tweet pertencer apenas a um tópico (cluster), ele pode falar de mais que um tema por isso, apesar de o BERTopic falhar nessa hipótese, este algoritmo ajuda a saber se havia algum outro tema no tweet. Ou seja, apesar de ajudar a ver isso, não resolve o problema de o BERTopic não considerar a hipótese e vários tópicos no mesmo tweet.