**Conceitos Aprendidos**

**RESUMO 1**

**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.

**RESUMO 2**

**CFDTM** - modelo que acompanha a evolução dos tópicos ao longo do tempo, evitando que eles fiquem repetitivos ou que apareçam palavras que não têm nada a ver com aquele momento.

**RESUMO 3**

**HDP (Hierarchical Dirichlet Process)** - modelo que permite descobrir automaticamente quantos tópicos existem num conjunto de documentos, sem precisar definir esse número antes. Cada grupo (ex: cada documento) tem o seu próprio DP local, mas todos partilham uma distribuição global de tópicos, pois um documento pode falar de vários tópicos, não apenas de um. Assim sendo, cada documento pode estar em vários tópicos ao mesmo tempo, com pesos diferentes. Ex: um discurso político pode ser 40% economia, 30% saúde, 30% educação.

**Gibbs Sampling** -método de inferência estatística que atualiza uma variável de cada vez; funciona, mas pode ser lento e preso a más soluções.

**MCMC** – Markov chain Monte Carlo - família de métodos que usam cadeias de Markov para explorar distribuições complexas.

**Metropolis-Hastings**: regra para decidir se uma mudança (como um split ou merge) deve ser aceite ou não.

**Chinese Restaurant Franchise** - metáfora usada para representar como documentos compartilham tópicos: cada documento é um restaurante, mesas são subgrupos de palavras e os pratos são os tópicos compartilhados.

**Burn-in** - fase inicial da cadeia MCMC em que o modelo ainda está a “aquecer” antes de estabilizar.

**RESUMO 4**

**Bootstrap resampling** - uma técnica estatística que “simula” várias versões dos dados originais (networks), para testar se os padrões que aparecem são mesmo sólidos ou só ruído

**Significance clustering** - em vez de confiar cegamente no algoritmo de clustering, eles verificam se as comunidades (clusters) se mantêm estáveis quando os dados são reamostrados

**Alluvial diagrams**: gráficos que mostram como grupos se dividem (splits) ou se juntam (merges) ao longo do tempo