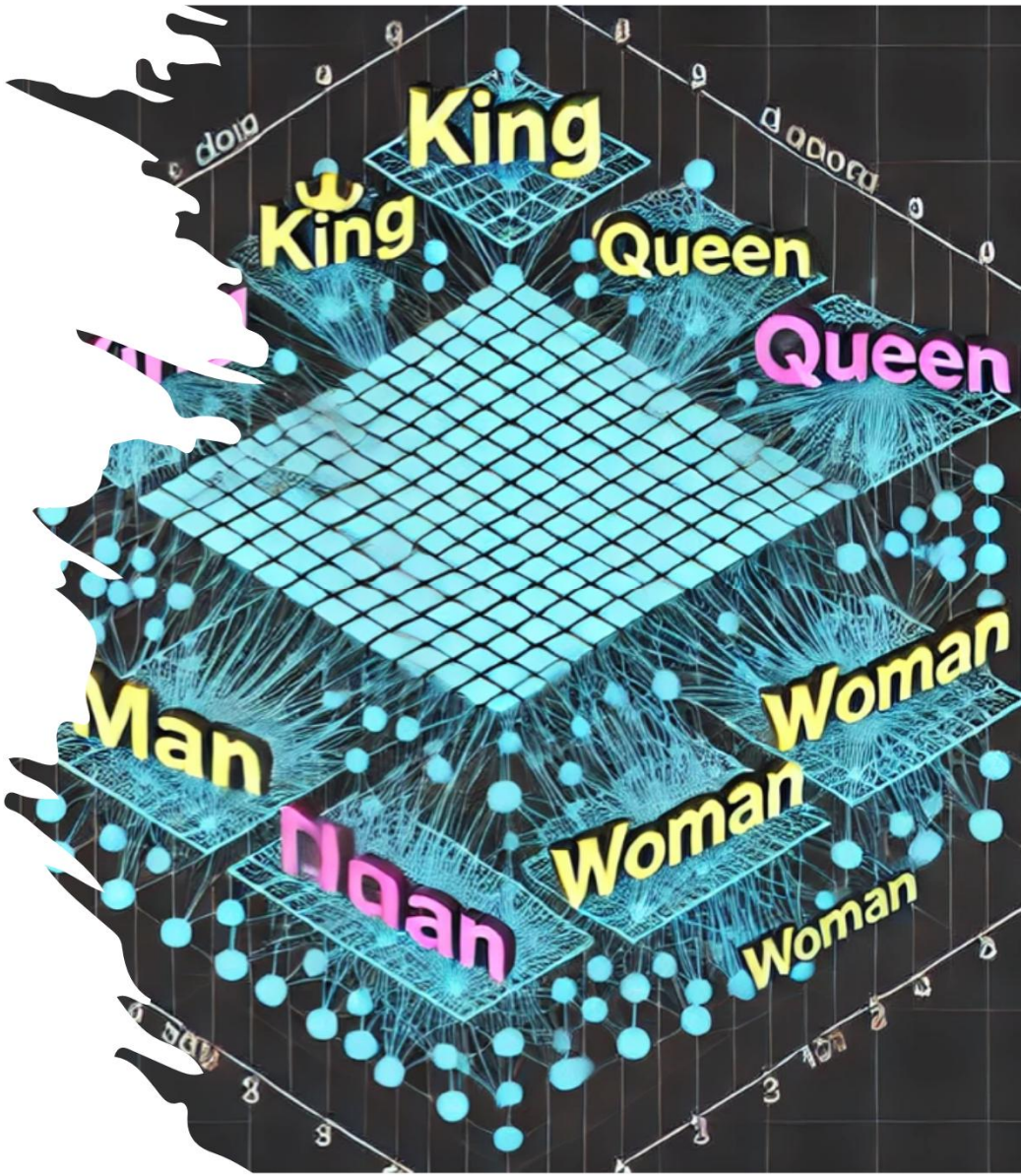
Embeddings

- Representações vetoriais de palavras em um espaço contínuo, onde similaridades semânticas são preservadas.
- Essas representações são usadas em processamento de linguagem natural para capturar relacionamentos e significados contextuais entre palavras



Palavra	Dimensão 1	Dimensão 2	Dimensão 3	Dimensão 4	Dimensão 5
king	0.4	0.3	0.2	0.6	0.1
queen	0.5	0.3	0.1	0.6	0.2
man	0.1	0.7	0.4	0.3	0.4

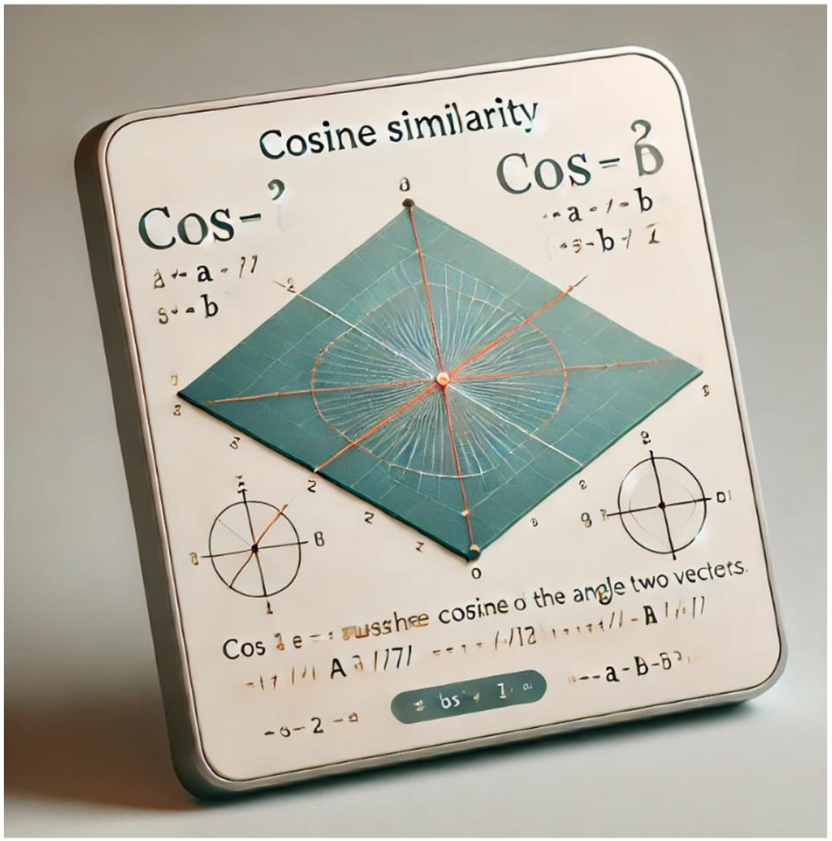
Exemplo

Cálculo de similaridade do cosseno.

King e Queen	0,982
King e Man	0,711

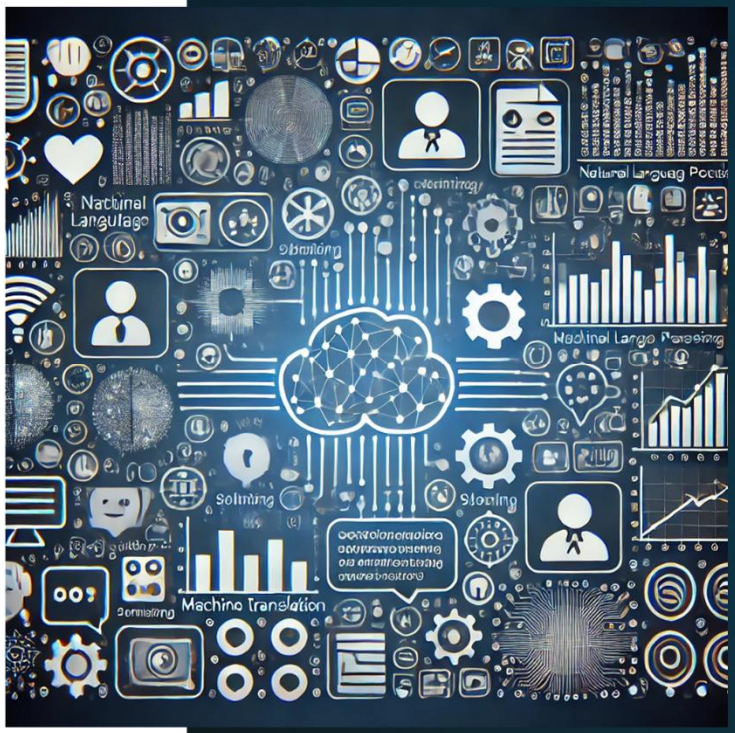
Similaridade do Coseno

- **Valores próximos de 1:** Indicam alta similaridade entre os documentos.
- **Valores próximos de 0:** Indicam pouca ou nenhuma similaridade entre os documentos.
- **Valores negativos:** Indicam dissimilaridade, embora, em muitos casos práticos de embeddings de texto, valores negativos sejam raros e geralmente ocorrem quando os vetores estão em direções opostas.



Outras Métricas de Similaridade

- **Distância Euclidiana**:
- **Distância de Manhattan (ou Distância L1)**:
- **Distância de Minkowski**:
- **Distância de Jaccard**:
- **Distância de Hamming**:

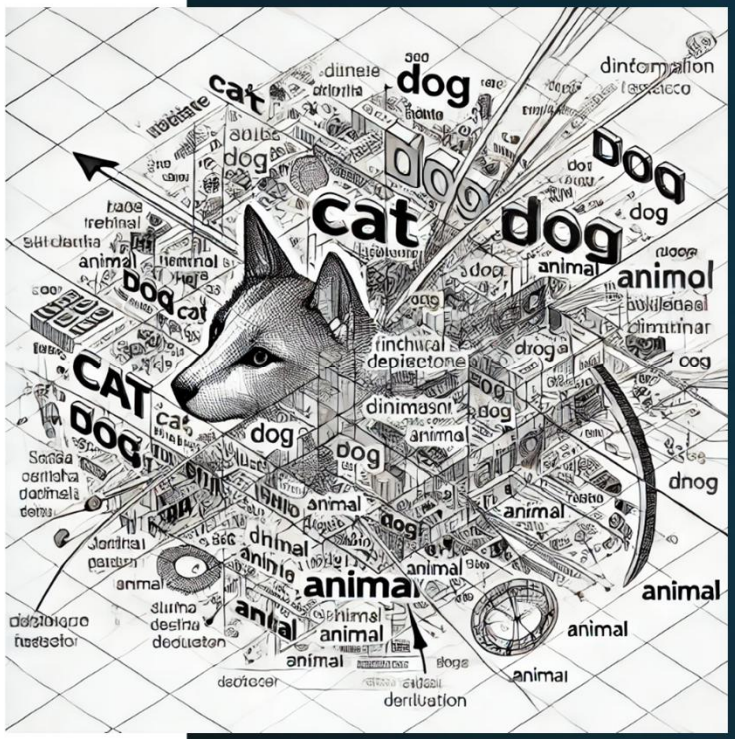


Aplicações

- **Análise de Sentimentos**
- **Classificação de Texto**
- **Machine Translation (Tradução Automática**
- **Reconhecimento de Entidades Nomeadas (NER**
- **Busca Semântica**

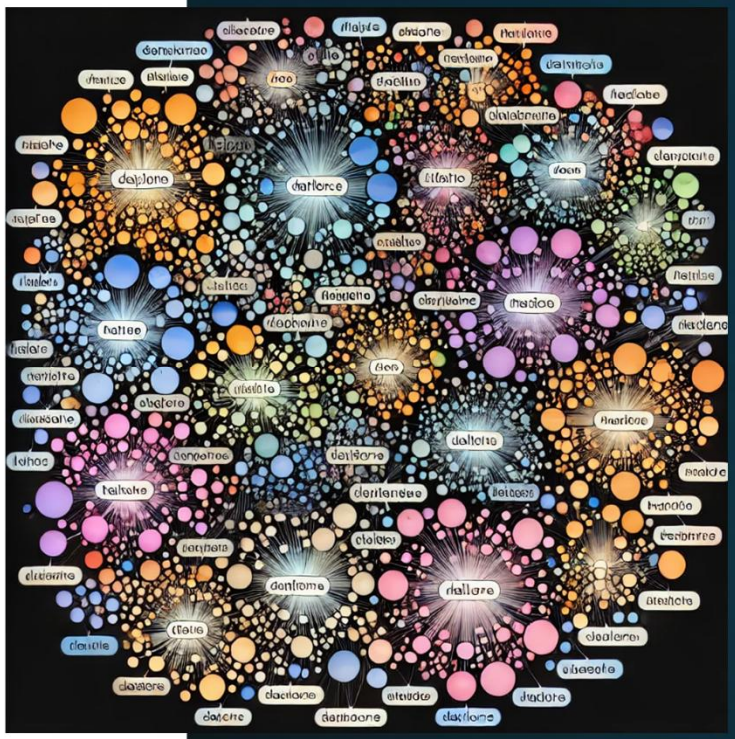
Aplicações

- **Resumo Automático de Texto**
- **Geração de Texto**
- **Sistemas de Recomendação**
- **Detecção de Fraude**
- **Análise de Tópicos**



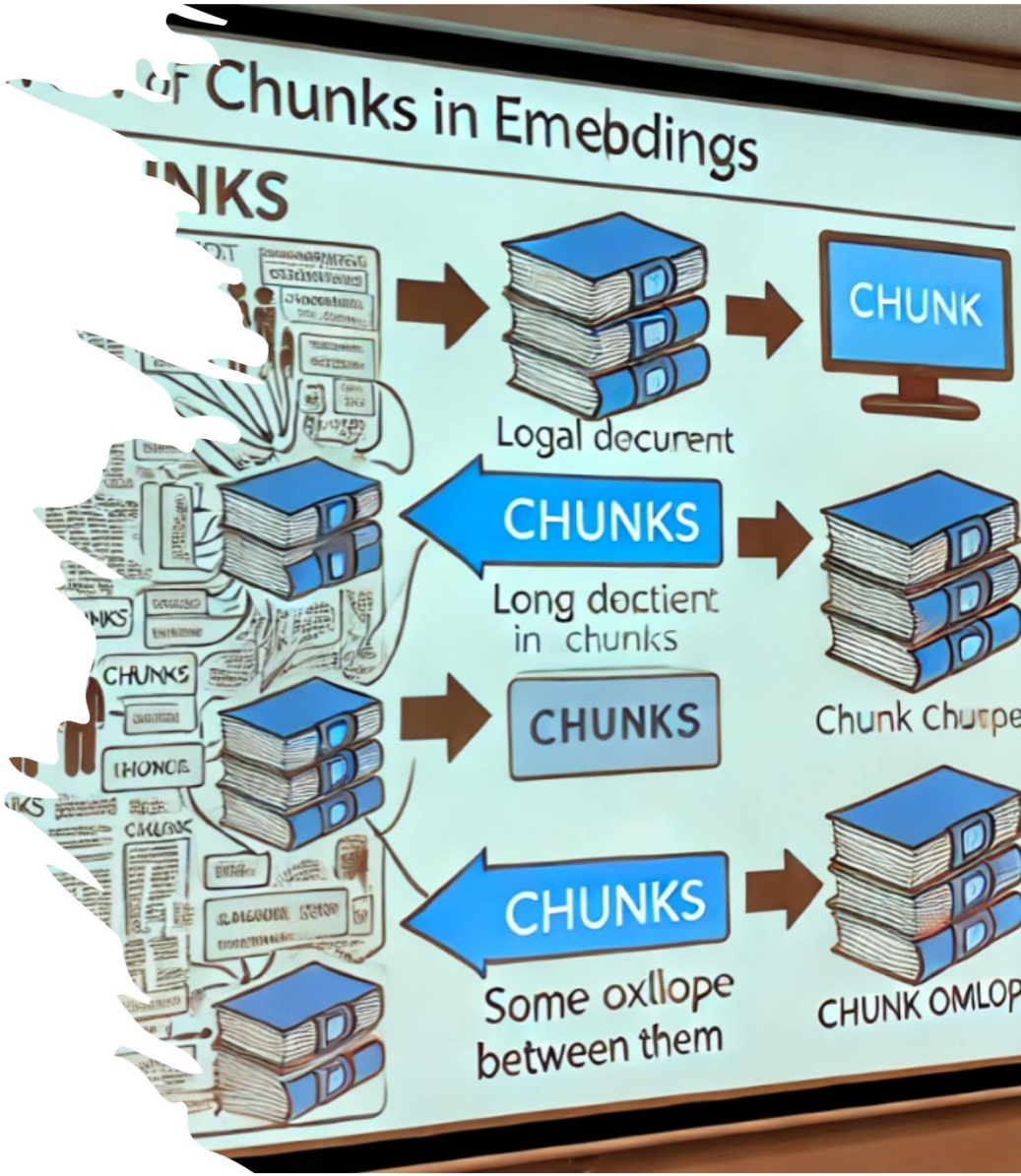
Visualização de Embeddings

- A visualização ajuda a identificar padrões, relações semânticas e clusters de palavras ou frases similares.
- Técnicas:
  - t-SNE (t-Distributed Stochastic Neighbor Embedding)
  - PCA (Principal Component Analysis)
  - UMAP (Uniform Manifold Approximation and Projection)
- Ferramentas
  - TensorFlow
  - Matplotlib
  - Seaborn
  - Plotly



Chunks

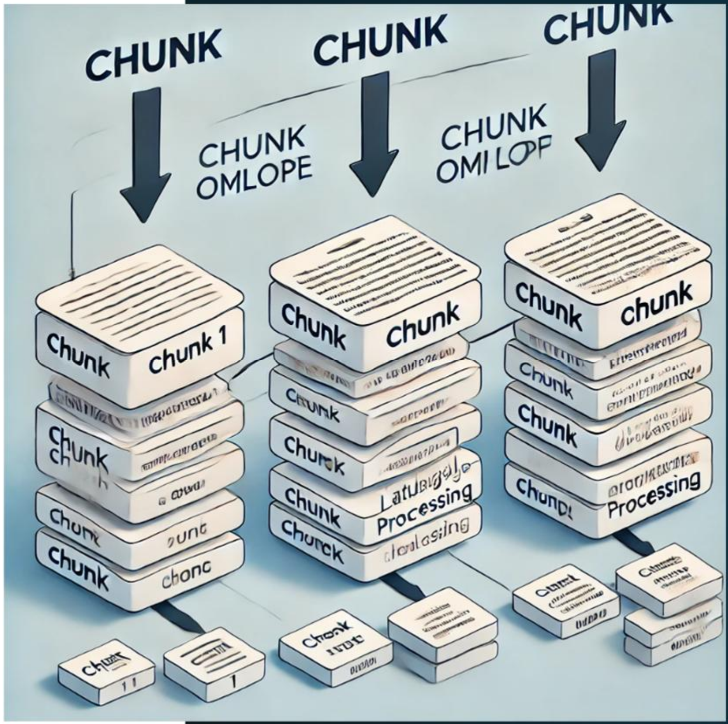
- Divisões menores de um documento
- Chunk overlap é a prática de permitir que os chunks se sobreponham parcialmente
- Overlap ocorre apenas entre chunks do mesmo documento



Chunk overlap

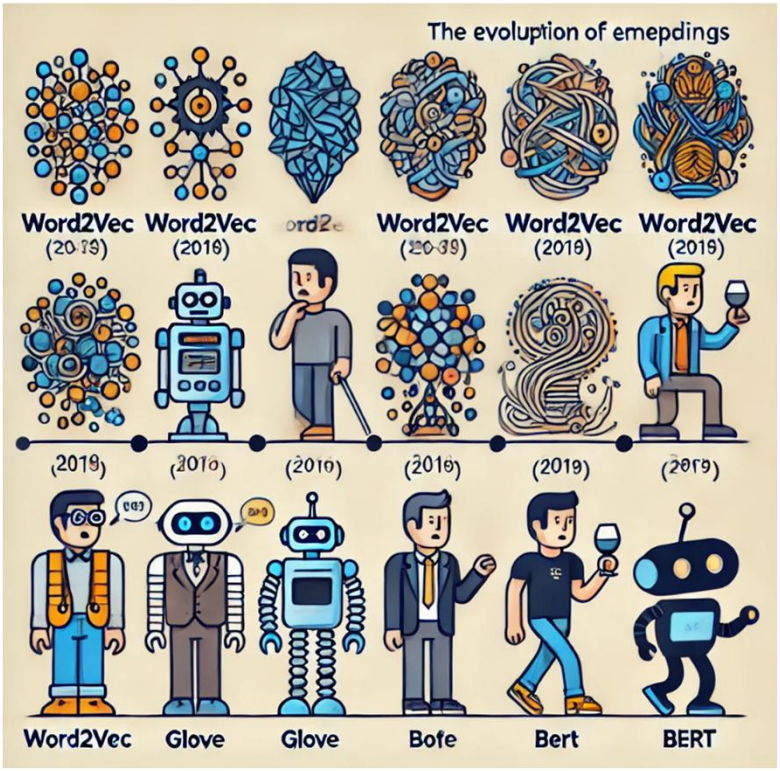
"Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language. In particular, NLP is interested in how to program computers to process and analyze large amounts of natural language data.“

- "Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the **interactions**“
- "**interactions** between computers and human language. In particular, NLP is interested in how to **program computers**“
- "**program computers** to process and analyze large amounts of natural language data."



Evolução dos Embeddings

- Word2Vec (2013)
- GloVe (2014)
- FastText (2016)
- ELMo (2018)
- BERT (2018)
- GPT (2018)



Considerações Finais

- A natureza e a qualidade dos embeddings dependem do modelo específico e do corpus de treinamento utilizado.
- Embedding VS Modelo:
  - Embedding: Representação vetorial de frases
  - Modelo: Estrutura Neural que processa e gera estas representações