Análise e Sumário dos Resultados

Análise de Similaridade de Vagas de Engenharia de Dados: Comparativo São Paulo vs. Remoto

1. Introdução e Objetivo

Este documento apresenta a análise de dados realizada sobre um conjunto de vagas de emprego, com o objetivo de aplicar conceitos de Álgebra Linear (especificamente, TF-IDF e Similaridade de Cosseno) para identificar e comparar perfis de vagas em diferentes mercados de trabalho.

A análise foi desenhada para responder à seguinte pergunta: "Quão similares são as vagas de Engenheiro de Dados em São Paulo e em regime Remoto, quando comparadas a um perfil ideal (query)?"

2. Metodologia

O processo de análise seguiu quatro etapas principais:

2.1. Fonte de Dados (Dataset)

A análise utilizou o dataset público engenheiro_de_dados_6k.csv, obtido da plataforma Kaggle. Este conjunto de dados contém 1.614 descrições de vagas para a posição de Engenheiro de Dados (Data Engineer), extraídas do LinkedIn no Brasil durante o ano de 2023.

2.2. Definição dos Grupos (Documentos)

Os 1.614 documentos de texto (as descrições das vagas) foram filtrados e segmentados em dois grupos distintos para comparação:

- **Grupo 1 (São Paulo):** Composto por **756 documentos** (vagas) cuja localização continha "São Paulo".
- **Grupo 2 (Remoto):** Composto por **180 documentos** (vagas) marcadas como work_remote_allowed (trabalho remoto permitido).

2.3. Definição do Perfil (Query)

Para simular a busca por um perfil ideal, foi definida a seguinte "query" (consulta) em inglês, para maximizar a correspondência com termos técnicos:

- **Query:** "data engenieer with experience in python and sql"

2.4. Processamento (Álgebra Linear)

1. **TF-IDF (Term Frequency-Inverse Document Frequency):** A Query e todos os documentos de cada grupo foram transformados em vetores numéricos. Este processo atribui "pesos" a cada palavra, dando mais importância a termos que são relevantes em um documento, mas não tão comuns em todos os outros.
2. **Similaridade de Cosseno:** Foi calculada a similaridade (o cosseno do ângulo) entre o vetor da Query e cada um dos vetores de documentos em ambos os grupos (SP e Remoto). Um score de 1.0 (ângulo de 0°) significa identidade total, enquanto um score de 0.0 (ângulo de 90°) significa nenhuma similaridade.

3. Resultados da Análise

A análise ranqueou os documentos em cada grupo com base na proximidade angular (maior similaridade) com a query definida.

3.1. Grupo 1: Top 3 - Vagas em São Paulo

As vagas em São Paulo com maior aderência à query foram:

1. **Rank 1 (ID 571):** Similaridade: 0.3576 (Ângulo: 69.05°)
2. **Rank 2 (ID 193):** Similaridade: 0.3576 (Ângulo: 69.05°)
3. **Rank 3 (ID 132):** Similaridade: 0.2430 (Ângulo: 75.94°)

3.2. Grupo 2: Top 3 - Vagas Remotas

As vagas Remotas com maior aderência à query foram:

1. **Rank 1 (ID 45):** Similaridade: 0.3032 (Ângulo: 72.35°)
2. **Rank 2 (ID 60):** Similaridade: 0.3032 (Ângulo: 72.35°)
3. **Rank 3 (ID 409):** Similaridade: 0.3032 (Ângulo: 72.35°)

4. Discussão e Conclusão

A análise dos resultados revelou descobertas importantes sobre a natureza dos dados e o funcionamento do algoritmo:

- **A Descoberta Principal:** A análise da query revelou um "cluster" (agrupamento) de vagas dominado pela empresa de recrutamento "Turing", que apareceu de forma massiva nos resultados de *ambos* os grupos (SP e Remoto).
- **A Causa (Vagas Duplicadas):** Os resultados mostram scores de similaridade e ângulos *idênticos* para vagas diferentes (ex: Ranks 1 e 2 de SP; Ranks 1, 2 e 3 de Remoto). Isso prova que o algoritmo detectou que

essas vagas são, na verdade, **cópias exatas do mesmo documento de texto**, publicadas múltiplas vezes no dataset.

- **A Implicação (Ruído nos Dados):** A análise expôs um "ruído" significativo nos dados. O uso de "texto boilerplate" (o parágrafo de introdução padrão da Turing) faz com que essas vagas dominem o topo do ranking, mesmo que a vaga real (descrita mais abaixo no texto) possa ser ligeiramente diferente.
- **A Prova de Eficácia:** O algoritmo, no entanto, funcionou como esperado. O Rank 3 de SP (ID 132), referente a uma vaga de "Azure Data Engineer", não é da Turing e foi corretamente ranqueado como menos similar (score 0.2430) do que as vagas da Turing (score 0.3576), provando que o sistema ranqueou os documentos de forma matematicamente correta, mas que as vagas da Turing eram, de fato, as mais similares à query.