PROJETO DE ANÁLISE DE DADOS

Especialização Gestão e Ciência de Dados
Dados, Conhecimento e Inteligência Artificial

Equipe: Sergio Lucio Nunes da Silva

Análise de vagas de emprego anunciadas no Twitter

1. Introdução

O mercado de tecnologia está abalado recentemente, e na pandemia a mídia foi inundada com notícias de muitas vagas nesse mercado, altos salários, falta de profissionais, e digitalização forçada das empresas pareciam trazer para a área uma grande demanda que se manteria ao longo do tempo.

Os dados de vagas de emprego anunciadas no Twitter podem ajudar a responder algumas perguntas sobre esse tópico? Houve de fato um aumento na demanda durante a pandemia? Há uma queda nas vagas anunciadas em 2023 em relação a outros anos? Houve um aumento das vagas remotas? As empresas descentralizaram as vagas dos grandes centros urbanos? Com essas perguntas em mente, o dataset foi explorado.

Ao avaliar o dataset e analisar os dados nele contidos, buscamos: verificar a composição dos dados no pré-processamento; avaliar o balanceamento da amostra; aplicar conceitos de visualização de dados para responder as principais perguntas.

2. Coleta de dados

Os dados foram obtidos através do dataset disponibilizado no Kaggle. O Kaggle utiliza três métricas para avaliar a usabilidade de um dataset: completude, credibilidade e compatibilidade, esse dataset foi escolhido, entre outros fatores, por possuir nota dez, na métrica de usabilidade de um dataset no Kaggle. Ele contém cinquenta mil tweets coletados via API do Twitter entre janeiro de abril de 2019 e março de 2023.

O uso de um dataset com boa usabilidade se deve a alguns fatores, dentre eles: redução de tempo ao buscar dataset mais completos, porém de fontes diferentes e que precisam ser unidos por vínculos não tão triviais; tempo disponível para execução do projeto; complexidade de consumo de dados via conexão direta com API do Twitter; otimização de tempo em sanitização dos dados.

3. Pré-processamento dos dados

O pré-processamento dos dados ocorreu em duas plataformas diferentes, Jupyter Notebook e Power Query. No Jupyter Notebook foram feitas apenas análises triviais usando a biblioteca Pandas para ler o CSV, mostras as primeiras linhas do dataframe, a descrição dos tipos e qualidade das colunas e as dimensões do dataframe que continha 50.000 linhas e 12 colunas. A escolha do Jupyter Notebook se deu apenas para revisão do funcionamento da ferramenta, já conhecida anteriormente.

No Power Query, ferramenta da Microsoft para ingestão e transformação de dados, temos muitas funções de manipulação de tabelas poderosas, tudo isso aliado a uma interface gráfica bonita e com boa experiência ao usuário. Os recursos visuais facilitam a manipulação e operações na tabela, contudo, ficam claras algumas limitações em operações não tão bem definidas e repetitivas, que podem ser muito mais eficientes quando realizadas com Python do que na linguagem M, desenvolvida pela Microsoft para o Power Query.

Os scripts de pré-processamento dos dados feitos em linguagem M, estão disponíveis no repositório do Github, link nas referências. Como os dados já estavam organizados, foram necessárias poucas transformações, entre elas, preenchimento de informações ausentes, extração de textos de colunas e transposição de uma tabela auxiliar que tratava das hashtags. Para executar o script basta abrir a opção “editor avançado” Power Query e alterar o diretório para o diretório local onde os dados do .csv estão.

4. Conclusão

Sugestão de possíveis ações a serem tomadas com base nos resultados

Com base no processo de análise de dados, notamos a importância do balanceamento dos dados, uma vez que a falta de dados mais bem distribuídos ao longo do tempo pode gerar vieses na análise e oculta informações que podem contradizer os próprios argumentos obtidos com base nos dados.

A qualidade dos dados (em relação a uniformidade de tipos, padronização de inputs etc.) é fundamental na análise, uma vez que dados de má qualidade podem gerar conclusões completamente equivocadas ou ainda impossibilidade a realização de uma análise.

Principalmente baseado no campo de ‘texto’ do dataset, é possível desenvolver uma forma de análise textual com aprendizado de máquina que complemente as interpretações do painel e até teste a predição de algumas outras informações do painel ou de outro anúncio de vaga, conforme o padrão estabelecido.

5. Referências

Repositório do Projeto: GitHub - luciosh/analise_job_tweets
Dataset: Job Vacancy Tweets | Kaggle
Link para o relatório: Microsoft Power BI