Desafio 1 - Engenheiro de Dados

O desafio é que você consiga desenvolver um processo de **web crawling** para estruturar dados de portais concorrentes de notícias, agregar métricas sobre esses dados e disponibilizar informações.

Deadline de Submissão: 7 dias;

Propósitos

- Demonstração de habilidade em codificar uma solução de dados;
- Demonstração de capacidade de organização de dados;
- Avaliação de decisões técnicas;

Quesitos

1. Organização, Reprodução e Documentação;

Escreva um README.md que explique como fazer *build* (se aplicável) e como executar seu código em Linux ou MacOS X.

Documente também, em um log de trabalho, o que você fez e o processo de tomada de decisões.

Ao final, apresente algumas métricas de performance do seu código e como seria possível escalar a solução.

Você pode elaborar mais e escrever sobre estratégia de *deploy* dessa solução em produção.

2. Aquisição e Estruturação dos dados

Escreva um *web crawler* e *scrapper*, utilizando *frameworks* de sua preferência para adquirir dados de portais concorrentes em uma área específica (ex.: games, tech, carros) e estruturar as principais informações (título, corpo do texto, autor, data, url, tags).

*A estratégia de crawling é livre.

*Persista dados de maneira a otimizar consultas analíticas.

*Implemente ou considere aspectos de monitoramento dos *crawlers/scrappers* . *Considere a importância de salvar metadados da aquisição - para manutenção da base.

3. Enriquecimento dos dados

Implemente métricas para cada artigo estruturado.

Ex.:

1. Quantidade;
2. Métricas derivadas do texto;
3. Alguma métrica de engajamento.

As métricas podem vir da aquisição inicial (quantidade), serem extraídas a partir de características do próprio dado (ex.: parágrafos) ou virem de *lookup* externo (ex.: engajamentos).

Disponibilização da informação

Seu processo deve, ao final, escrever novos dados estruturados de artigos em uma base para consultas analíticas. Deve-se utilizar estratégia incremental de carga, ou seja, apenas novos artigos devem ser escritos na base final, com dados em granularidade ideal para análises e métricas normalizadas para a granularidade escolhida.

*Dê exemplos de análises que poderiam ser feitas.

Disponibilização

Submeta seu projeto compartilhando um repositório (github, bitbucket, gitlab);

Envie o link com permissão de acesso para balaum@gmail.com, brunors@me.com, raniel90@gmail.com, felipe.bormann@gmail.com ou mantenha o repositório público.