



Ask the Data no Dados.IPB

Adília Alvesa, Clarisse Paisb

^a Instituto Politécnico de Bragança, aialves@ipb.pt

^b Instituto Politécnico de Bragança, clarisse@ipb.pt

Resumo

O Instituto Politécnico de Bragança (IPB) mantém os seus dados de investigação na plataforma dados.IPB (<https://dados.ipb.pt>). Os dados armazenados nesta plataforma abrangem diversas áreas de conhecimento e estão guardados em distintos formatos (ex: .xlsx, .docx, .pdf, .jpg, .tab). O Dataverse possui metadados que classificam os *datasets* e que permitem a recuperação dos mesmos e respetivos dados após a introdução, pelos utilizadores, de keywords em campos de pesquisa. Os resultados são obtidos com a execução de *queries* SQL sobre a base de dados Dataverse PostgreSQL. Mais recentemente é possível pesquisar na plataforma Dataverse utilizando, não só os metadados, mas também os próprios dados. A Inteligência Artificial é um componente em crescimento no ecossistema Dataverse, também na funcionalidade de pesquisa de informação nos dados armazenados nos datasets. A extensão *Ask the Data* (<https://github.com/IQSS/askdataverse/tree/main/askthedata>) é uma ferramenta externa ao Dataverse que permite fazer perguntas em linguagem natural sobre os dados contidos nas tabelas do Dataverse (dados tabulares, com extensão .tab). Um ficheiro .tab é um ficheiro de texto delimitado por tabulações, ou seja, é um formato de texto simples onde os valores são separados pelo carater de tabulação (\t). Estes ficheiros são frequentemente usados para armazenar e trocar dados estruturados, como folhas de cálculo, tabelas ou bases de dados. Ficheiros .tab são: leves, legíveis por humanos, compatíveis com várias plataformas, menos ambíguos do que .csv porque o texto pode ter vírgulas e não tem o carater '\t'. São, por isso, amplamente utilizados. O IPB instalou a extensão *Ask the Data* no Dataverse. Este *Large Language Model* analisa apenas o conteúdo dos dados tabulares e não a descrição do conjunto de dados ou quaisquer metadados. Depois de instalar esta extensão o IPB começou a desenvolver outra ferramenta externa ao Dataverse, *Ask the Data2* utiliza o *YOLO*, que irá detetar objetos previamente definidos em imagens guardadas no Dataverse e responder a perguntas em linguagem natural.

Palavras-chave: Dataverse, Ask the Data, YOLO, Dados.IPB

Designação do projeto/infraestrutura/iniciativa

Público-alvo

A audiência privilegiada da apresentação são os coordenadores de projetos, investigadores, alunos de mestrado e doutoramento.

Ligações web úteis

AI Guide — Dataverse.org. <https://guides.dataverse.org/en/latest/ai/index.html>

YOLOv12 <https://github.com/sunsmarterjie/yolov12>

Building external tools <https://guides.dataverse.org/en/latest/api/external-tools.html>